Rolf Steyer and Werner Nagel

Probabilistic Foundations of the Empirical Sciences

Probability and Conditional Expectation

October 22, 2015

University of Jena

Preface

Why another book on probability?

This book has two titles. The first, 'Probabilistic Foundations of the Empirical Sciences' reflects the intensions and the motivation of the first author for writing this book. He received his academic training in psychology, but considers himself a methodologist. His scientific interest is in explicating fundamental concepts of empirical research (such as causal effects and latent variables) in terms of a language that is *precise* and at the same time is *compatible with the statistical models* used in the analysis of empirical data. Applying statistical models we aim at estimating and testing hypotheses about parameters such as expectations, variances, covariances, etc. or of functions of these parameters such as differences between expectations, ratios of variances, or regression coefficients, etc., all of which are terms of probability theory. Precision is necessary for securing logical consistency of theories, whereas compatibility of substantive theories with statistical models is crucial for probing the empirical validity of theoretical propositions via statistical inference.

Much empirical research uses some kind of regression in order to investigate how the expectation of one random variable depends on the values of one or more other random variables. This applies for analysis of variance, regression analysis, the general linear model, the generalized linear model, factor analysis, structural equation models, hierarchical linear models, and analysis of qualitative data. Using these methods we aim at learning about specific regressions. A regression is a synonym for what, in probability theory, is called a factorization of a conditional expectation, and this explains the second, more concrete title of this book: 'Probability and Conditional Expectation'.

What is it about?

Since the seminal book of Kolmogoroff (1933/1977) the fundamental concepts of probability theory are considered to be special concepts of measure theory. A probability measure is a special finite measure, random variables are special measurable mappings, and expectations of random variables are integrals of measurable mappings with respect to a probability measure. This motivates Part I of this book with three chapters on the measure-theoretical foundations of probability theory. Although at first sight this part seems to be far-off from practical applications, the contrary is true. This part is indispensable for probability

Figure 1

theory and for its applications in substantive sciences and empirical research. This does not only apply to the concepts of a measure and an integral but in particular to the concept of a measurable mapping, although we concede that the full relevance of this concept will become apparent only in the chapters on conditional expectations. The relevance of measurable mappings is also the reason why this chapter is more extended than the corresponding chapters in other books on measure theory.

The Part II of the book is fairly conventional. The material covered, probability, random variable, expectation, variance, covariance, and some distributions, is found in many books on probability and statistics.

Part III is not only the longest, it is also the core of the book that distinguishes it from other books on probability or on probability and statistics. Only few of these other books contain detailed chapters on conditional expectations. Exceptions are Billingsley (1995), Fristedt and Gray (1997), and Hoffmann-Jørgensen (1994). Our book does not cover any statistical model. However, we treat in much detail *what* we are estimating and *about what* we test or evaluate hypotheses using statistical models. *How* we are estimating is important but *what* we are estimating is of most interest from a substantive point of view and this point is typically neglected in books on statistics and in books on probability theory such as Bauer (1996) or Klenke (2008). A simple example in case is the meaning of the coefficient β_2 in the equation $E(Y|X,Z) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX$. Oftentimes, this coefficient is misinterpreted as the 'main effect' of X. However, sometimes β_2 has no autonomous meaning at all, e. g. if $P(Z=0) = 0$. In general, this coefficient is just a component of the function $g_1(Z) = \beta_2 + \beta_3 Z$ that can be used to compute the conditional effects of X on Y for various values z of Z (see chapter 15 for more details). The crucial point is that such concepts can be treated most clearly within probability theory, without referring to a statistical model, sample, estimation or testing. This also includes exemplifying the limitations of conditional

expectations. Simple examples show that conditional expectations do not necessarily serve the substantive purpose of evaluating the effects of an intervention on an outcome variable. But even in these situations conditional expectations are indispensable for the definition of the parameters of substantive interest (see, e. g., chapter 14).

There is much overlap of Parts II and III with Steyer (2003). However, that book is written in German and the mathematics is considerably less rigorous. This explains the cooperation with the second author writing this book.

In the first chapter of Part III we gently introduce conditional expectation values and discrete conditional expectations. In the next chapter of this part we then present the general theory of conditional expectations that has been introduced by Kolmogoroff (1933/1977) and is since that time treated in many books on probability theory — although much too briefly in order to be intelligible for researchers in empirical sciences. Our chapter on conditional expectations contains many more details and is supplemented by a number of other chapters on important special aspects and special cases. Such a special aspect is the concept of a residual with respect to a conditional expectation. Residuals have many interesting properties and they are used in order to introduce the concepts of conditional variance and covariance, as well as the notion of a partial correlation. We then turn to specific parametrizations of a conditional expectation, including the concepts of a linear regression and a linear logistic regression. Note that these concepts are introduced as probabilistic concepts. As mentioned above, they are what we aim at estimating in applying the corresponding statistical models. The next two chapters provide the probabilistic foundations of the analysis of conditional and average effects of treatments, interventions, or expositions to potentially harmful or beneficial environments. To our knowledge this material is not found in any other text book. Note, however, that although these two chapters provide important concepts, they do not cover the theory of causal effects, which is another book project of the first author.

Part IV uses conditional expectations in order to introduce conditional independence and conditional distributions. Although these two chapters are more extensive than comparable chapters or sections in other books, the material is found in other books on probability theory as well.

For Whom is it?

This book has been written for two kinds of readers. The first are applied statisticians and substantive researchers who want to understand in a proper language, i. e., in terms of probability theory, what they estimate and test in their empirical studies. The second kind of readers are mathematicians who want to understand in terms of probability theory what applied statisticians and substantive researchers estimate and test in their empirical research. Both kinds of readers are potential contributors to the methodology of empirical sciences.

Many exercises and their solutions provide extensive material for assignments in courses, but they also facilitate independent learning. At the same time, these exercises and their solutions help streamlining the main text.

Note that we do not provide all proofs, in particular in the chapters on measure, integral, and distributions. In these cases we refer to other textbooks, instead. We decided to include only those proofs that may help to increase understanding of the background and to learn important mathematical procedures. Of course, we provide proofs of all propositions for which we did not find an appropriate reference.

Prerequisites

We assume that the reader is familiar with the elementary concepts of logic, sets, functions, sequences, and matrices as presented, e. g., in chapters 1 and 2 of Rosen (2012). We try to stick to his notation as close as possible. One of the exceptions is the symbol for the implication for which we use $\Rightarrow$ instead of $\rightarrow$. Another exception is the symbol for the equivalence for which we use $\Leftrightarrow$ instead of $\leftrightarrow$. Box 0.1 summarizes the most important notation to start with. The concepts referred to by these symbols are defined, e. g., in Rosen (2012) or in Ellis and Gulick (2006).

Acknowledgements

This book could not have been written without the help of many. First of all we thank Ivailo Partchev who prepared the LaTex framework, many of the figures, tables, and boxes. Some of the figures have been produced by Désirée Thielemann, who also cared for references, read some of the chapters, and hinted at errors. For supporting us with respect to LaTeX, finding errors or suggesting other improvements we also thank Karoline Bading, Marcel Bauer, Sonja Hahn, Gregor Kappler, Andreas Neudecker, Axel Mayer, Erik Sengewald, Jan Plötner, and Tom Landes. Thanks are also due to Ernesto San Martin for suggesting section 1.7 and proposition (iv) of Theorem 16.37. The proof of Lemma 12.37 is due to Peter Vogel. Finally, we would like to thank our students who kept us thinking on how to improve the text.

Box 0.1 A First List of Symbols

$\neg$	not
$\wedge$	and
$\vee$	or
$\Rightarrow$	implies
$\Leftrightarrow$	equivalent to
$\exists$	there is (there exists)
$\forall$	for all
$a \in A$	a is an element of the set A
$\emptyset$	empty set
I	finite, countable, or uncountable index set
$(A_i, i \in I)$	family of sets A_i, $i \in I$
$A \cup B$	union of the sets A and B
$\bigcup_{i \in I} A_i$	union of the sets A_i, $i \in I$
$A \cap B$	intersection of the sets A and B
$\bigcap_{i \in I} A_i$	intersection of the sets A_i, $i \in I$
$A \setminus B$	set difference of the set A and the set B
$A^c := \Omega \setminus A$	complement of a set A with respect to a set Ω
$A \subset B$	A is a subset of the set B; $A \subset B$ includes $A = B$
$A \times B$	Cartesian product (product set) of A and B
$\bigtimes_{i=1}^{n} A_i$	Cartesian product of the sets A_i, $i = 1, \ldots, n$
$f\colon A \to B$	mapping f assigning to each $a \in A$ (the domain) one and only one element $b \in B$ (the codomain)
$\sum_{i=1}^{n} a_i$	sum of the real numbers $a_1, \ldots, a_n$
$\prod_{i=1}^{n} a_i$	product of the real numbers $a_1, \ldots, a_n$
$\lim_{n \to \infty} a_n$	limit of a sequence $a_1, a_2, \ldots$ of real numbers
$\sum_{i=1}^{\infty} a_i$	$\lim_{n \to \infty} \sum_{i=1}^{n} a_i$, where $a_1, a_2, \ldots$ and therefore $\sum_{i=1}^{1} a_i, \sum_{i=1}^{2} a_i, \ldots$ are sequences of real numbers

Contents

Part I Measure-Theoretical Foundations of Probability Theory

1 Measure 3
- 1.1 Introductory Examples 3
- 1.2 σ-Algebra and Measurable Space 4
  - 1.2.1 σ-Algebra Generated by a Set System 9
  - 1.2.2 σ-Algebra of Borel Sets on $\mathbb{R}^n$ 12
  - 1.2.3 σ-Algebra on a Cartesian Product 13
  - 1.2.4 ∩-Stable Set Systems That Generate a σ-Algebra 15
- 1.3 Measure and Measure Space 16
  - 1.3.1 σ-Additivity and Related Properties 17
  - 1.3.2 Other Properties 18
- 1.4 Specific Measures 20
  - 1.4.1 Dirac Measure and Counting Measure 21
  - 1.4.2 Lebesgue Measure 22
  - 1.4.3 Other Examples of a Measure 22
  - 1.4.4 Finite and σ-Finite Measures 23
  - 1.4.5 Product Measure 24
- 1.5 Continuity of a Measure 24
- 1.6 Specifying a Measure via a Generating System 26
- 1.7 σ-Algebra That is Trivial With Respect to a Measure 27
- 1.8 Proofs 28
- 1.9 Exercises 30

2 Measurable Mapping 41
- 2.1 Image and Inverse Image 41
- 2.2 Introductory Examples 42
  - 2.2.1 Example 1: Rectangles 42
  - 2.2.2 Example 2: Flipping two Coins 44
- 2.3 Measurable Mapping 46
  - 2.3.1 Measurable Mapping 46
  - 2.3.2 σ-Algebra Generated by a Mapping 51
  - 2.3.3 Final σ-Algebra 53
  - 2.3.4 Multivariate Mapping 54
  - 2.3.5 Projection Mapping 56
  - 2.3.6 Measurability With Respect to a Mapping 56

2.4 Theorems on Measurable Mappings 58
2.4.1 Measurability of a Composition 59
2.4.2 Theorems on Measurable Functions 61
2.5 Equivalence of Two Mappings With Respect to a Measure 65
2.6 Image Measure 68
2.7 Proofs 70
2.8 Exercises 75

3 Integral 83
3.1 Definition 83
3.1.1 Integral of a Nonnegative Step Function 83
3.1.2 Integral of a Nonnegative Measurable Function 88
3.1.3 Integral of a Measurable Function 93
3.2 Properties 96
3.2.1 Integral of μ-Equivalent Functions 98
3.2.2 Integral With Respect to a Weighted Sum of Measures 100
3.2.3 Integral With Respect to an Image Measure 102
3.2.4 Convergence Theorems 103
3.3 Lebesgue and Riemann Integral 104
3.4 Density 106
3.5 Absolute Continuity and the Radon-Nikodym Theorem 108
3.6 Integral With Respect to a Product Measure 110
3.7 Proofs 111
3.8 Exercises 120

Part II Probability, Random Variable and its Distribution

4 Probability Measure 127
4.1 Probability Measure and Probability Space 127
4.1.1 Definition 127
4.1.2 Properties of a Probability Measure 128
4.1.3 Examples 128
4.2 Conditional Probability 131
4.2.1 Definition 132
4.2.2 Multiplication Rule 133
4.2.3 Examples 134
4.2.4 Theorem of Total Probability 135
4.2.5 Bayes' Theorem 137
4.2.6 Conditional-Probability Measure 138
4.3 Independence 141
4.3.1 Independence of Events 141
4.3.2 Independence of Set Systems 142
4.4 Conditional Independence Given an Event 144
4.4.1 Conditional Independence of Events Given an Event 144
4.4.2 Conditional Independence of Set Systems Given an Event . 144
4.5 Proofs 146

4.6 Exercises 148

5 Random Variable, Distribution, Density, and Distribution Function . . 153
5.1 Random Variable and its Distribution 153
5.2 Equivalence of Two Random Variables With Respect to a Probability Measure 158
5.2.1 Identical and P-Equivalent Random Variables 158
5.2.2 P-Equivalence, P^B-Equivalence, and Absolute Continuity . 161
5.3 Multivariate Random Variable 164
5.4 Independence of Random Variables 166
5.5 Probability Function of a Discrete Random Variable 171
5.6 Probability Density With Respect to a Measure 175
5.6.1 General Concepts and Properties 175
5.6.2 Density of a Discrete Random Variable 176
5.6.3 Density of a Bivariate Random Variable 177
5.7 Uni- or Multivariate Real-Valued Random Variable 178
5.7.1 Distribution Function of a Univariate Real-Valued Random Variable 179
5.7.2 Distribution Function of a Multivariate Real-Valued Random Variable 181
5.7.3 Density of a Continuous Univariate Real-Valued Random Variable 182
5.7.4 Density of a Continuous Multivariate Real-Valued Random Variable 184
5.8 Proofs 185
5.9 Exercises 193

6 Expectation, Variance, and Other Moments 197
6.1 Expectation 197
6.1.1 Definition 197
6.1.2 Expectation of Discrete Random Variables 198
6.1.3 Computing Expectations Using Densities 200
6.1.4 Transformation Theorem 201
6.1.5 Rules of Computation 204
6.2 Moments, Variance, and Standard Deviation 205
6.3 Proofs 209
6.4 Exercises 210

7 Linear Quasi-Regression, Covariance, and Correlation 213
7.1 Linear Quasi-Regression 213
7.2 Covariance 216
7.3 Correlation 220
7.4 Expectation Vector and Covariance Matrix 223
7.4.1 Random Vector and Random Matrix 223
7.4.2 Expectations of a Random Vector and a Random Matrix . . . 223
7.4.3 Covariance Matrix of two Multivariate Random Variables . . 224

7.5 Multiple Linear Quasi-Regression 227
7.6 Proofs 228
7.7 Exercises 232

8 Some Distributions 241
8.1 Some Distributions of Discrete Random Variables 241
8.1.1 Discrete Uniform Distribution 241
8.1.2 Bernoulli Distribution 242
8.1.3 Binomial Distribution 243
8.1.4 Poisson Distribution 246
8.1.5 Geometric Distribution 247
8.2 Some Distributions of Continuous Random Variables 250
8.2.1 Continuous Uniform Distribution 250
8.2.2 Normal Distribution 252
8.2.3 Central χ^2-Distribution 255
8.2.4 Central t-Distribution 256
8.2.5 Central F-Distribution 258
8.2.6 Multivariate Normal Distribution 260
8.3 Proofs 263
8.4 Exercises 266

Part III Conditional Expectation and Regression

9 Conditional Expectation Value and Discrete Conditional Expectation 271
9.1 Conditional Expectation Value 271
9.2 Transformation Theorem 274
9.3 Other Properties 276
9.4 Discrete Conditional Expectation 277
9.5 Discrete Regression 279
9.6 Examples 280
9.7 Proofs 284
9.8 Exercises 285

10 Conditional Expectation 289
10.1 Assumptions and Definitions 289
10.2 Existence and Uniqueness 291
10.2.1 Uniqueness With Respect to a Probability Measure 292
10.2.2 A Necessary and Sufficient Condition of Uniqueness 293
10.2.3 Examples 294
10.3 Rules of Computation and Other Properties 295
10.3.1 Rules of Computation 295
10.3.2 Monotonicity 296
10.3.3 Convergence Theorems 296
10.4 Factorization, Regression, and Conditional Expectation Value 300
10.4.1 Existence of a Factorization 300
10.4.2 Conditional Expectation and Mean-Squared Error 301

10.4.3 Uniqueness of a Factorization 302
10.4.4 Conditional Expectation Value 303
10.5 Characterizing a Conditional Expectation by the Joint Distribution 305
10.6 Conditional Mean Independence 306
10.7 Proofs 311
10.8 Exercises 313

11 Residual, Conditional Variance, and Conditional Covariance 321
11.1 Residual With Respect to a Conditional Expectation 321
11.2 Coefficient of Determination and Multiple Correlation 325
11.3 Conditional Variance and Covariance Given a σ-Algebra 329
11.4 Conditional Variance and Covariance Given a Value of a Random Variable 330
11.5 Properties of Conditional Variances and Covariances 333
11.6 Partial Correlation 336
11.7 Proofs 338
11.8 Exercises 338

12 Linear Regression 347
12.1 Basic Ideas 347
12.2 Assumptions and Definitions 349
12.3 Examples 351
12.4 Linear Quasi-Regression 356
12.5 Uniqueness and Identification of Regression Coefficients 357
12.6 Linear Regression 359
12.7 Parametrizations of a Discrete Conditional Expectation 360
12.8 Invariance of Regression Coefficients 363
12.9 Proofs 365
12.10 Exercises 367

13 Linear Logistic Regression 369
13.1 Logit Transformation of a Conditional Probability 369
13.2 Linear Logistic Parametrization 371
13.3 A Parametrization of a Discrete Conditional Probability 373
13.4 Identification of Coefficients of a Linear Logistic Parametrization . 375
13.5 Linear Logistic Regression and Linear Logit Regression 376
13.6 Proofs 378
13.7 Exercises 380

14 Conditional Expectation With Respect to a Conditional-Probability Measure 383
14.1 Introductory Examples 383
14.2 Assumptions and Definitions 387
14.2.1 Conditional Expectation With Respect to a Conditional-Probability Measure 388
14.2.2 Partial Conditional Expectation 392

14.2.3 Examples 394
14.3 Factorization 396
14.3.1 Conditional Expectation Value With Respect to P^B 396
14.3.2 Uniqueness of Factorizations 397
14.4 Uniqueness 398
14.4.1 A Necessary and Sufficient Condition of Uniqueness 398
14.4.2 Uniqueness w.r.t. P and Other Probability Measures 399
14.4.3 Necessary and Sufficient Conditions of P-Uniqueness 400
14.4.4 Properties Related to P-Uniqueness 402
14.5 Conditional Mean Independence With Respect to $P^{Z=z}$ 406
14.6 Proofs 409
14.7 Exercises 413

15 Conditional Effect Functions 419
15.1 Assumptions and Definitions 419
15.2 Conditional Intercept Function and Effect Functions 420
15.3 Adjusted Conditional Effect Functions 423
15.4 Mean Independence of the Conditional Effect Functions 424
15.5 Conditional Logit Effect Functions 426
15.6 Proofs 432
15.7 Exercises 434

Part IV Conditional Independence and Conditional Distribution

16 Conditional Independence 439
16.1 Assumptions and Definitions 439
16.1.1 Two Events 439
16.1.2 Two Sets of Events 441
16.1.3 Two Random Variables 442
16.2 Properties 443
16.3 Conditional Independence and Conditional Mean Independence . 450
16.4 Families of Events 453
16.5 Families of Set Systems 454
16.6 Families of Random Variables 455
16.7 Proofs 456
16.8 Exercises 462

17 Conditional Distribution 467
17.1 Assumptions and Definitions 467
17.1.1 Conditioning on a σ-Algebra or a Random Variable 467
17.1.2 Conditioning on a Value of a Random Variable 470
17.2 Existence and Uniqueness 473
17.2.1 Existence 473
17.2.2 Uniqueness of the Functions $P_{Y|\mathscr{C}}(\cdot, A')$ 474
17.2.3 Common Null Set (CNS) Uniqueness of a Conditional Distribution 474

17.3 Conditional-Probability Measure Given a Value of a Random Variable 478
17.4 Decomposing the Joint Distribution of Random Variables 479
17.5 Conditional Independence and Conditional Distributions 482
17.6 Expectations With Respect to a Conditional Distribution 486
17.7 Conditional Distribution Function and Probability Density 489
17.8 Conditional Distribution and Radon-Nikodym Density 492
17.9 Proofs 495
17.10 Exercises 511

References 515

List of Figures

1 VI

1.1 A Venn diagram of two sets and their intersection 4
1.2 Example of a tree representation of a Cartesian product 7
1.3 A half-open rectangle in the plane $\mathbb{R}^2$ 12
1.4 Approximation of an open egg-shaped set O from below 25
1.5 Approximation of an open egg-shaped set O from above 26

2.1 Rectangles and their images under a function 42
2.2 Rectangles and their inverse images under a function 43
2.3 A set and its image under a function 44
2.4 A set and its inverse image under a function 45
2.5 A composition of two mappings 57
2.6 Two step functions 61
2.7 The positive and negative parts of a function and the sign function . 64

3.1 A partition and a subset of Ω 87
3.2 Increasing nonnegative step functions 90
3.3 Integral of nonnegative step functions w.r.t. the Lebesgue measure . 92
3.4 Lebesgue integral of a function from -5 to 5 95
3.5 Illustrating the construction of the Riemann integral 3.22 105
3.6 Integral of a density for two intervals 108

4.1 An Example in which probabilities are proportional to areas 130
4.2 Probability tree illustrating the multiplication rule 135
4.3 Venn diagram illustrating a partition of a set 137

5.1 Two random variables that are P^B-equivalent if $P(\{\omega_3\}) = 0$ 160

6.1 Variance of an indicator of an event as a function of its probability . 207

7.1 Linear quasi-regression 214
7.2 The regressor X, the linear quasi-regression f and their composition $Q_{lin}(Y|X) = f(X)$ 215
7.3 Intercept and slope of a linear function $f\colon \mathbb{R} \to \mathbb{R}$ 216

8.1 Probability and distribution functions of a binomial distribution ... 243

8.2 Illustrating the probability function of the sum of i. i.d. Bernoulli variables 245
8.3 Probability functions of binomial and Poisson distributions 247
8.4 Probability functions of three geometric distributions 249
8.5 Density of a bivariate uniform distribution 251
8.6 Density functions of three normal distributions 253
8.7 Density and distribution function of the standard normal distribution 254
8.8 Densities of three central χ^2-distributions 255
8.9 Densities of the standard normal and three t-distributions 257
8.10 Density functions of four central F-distributions 259
8.11 Density function of a bivariate normal distribution 260

9.1 The conditional expectation $E(Y|X)$ as the composition of X and its factorization 279

12.1 $E(Y|X)$ as the composition of X and the linear regression g 359
12.2 A regression with a parametrization that is linear in (X, X^2) but not in X 363

13.1 Graph of the logit transformation of p 370
13.2 Graphs of three logistic functions 373
13.3 $P(Y=1|X)$ as the composition of X, the linear logit regression f, and the logistic function h. 377

List of Tables

2.1 Example of measurable sets represented by a mapping X 45
2.2 Joe and Ann With Random Assignment and Measurable Mappings . . 54

4.1 Joe and Ann With Random Assignment: Probability Measures 130
4.2 Joe and Ann With Self-Selection 134

5.1 Tom, Jim, and Kate 157
5.2 No Treatment for Joe 163

9.1 Joe and Ann With Random Assignment: Conditional Expectations . . 280
9.2 No Treatment for Joe With Conditional Expectations 282

10.1 Joe and Ann With no Treatment Effects 308

11.1 No Treatment for Joe With Conditional Expectations and Residuals . 324
11.2 Joe and Ann With Self-Selection and Residuals 332

12.1 Joe and Ann With Self-Selection: Conditional Expectations 349

14.1 Joe and Ann With Random Assignment: Conditional Expectations With Respect to $P^{X=x}$ 384
14.2 Joe and Ann With Self-Selection: Conditional Expectations With Respect to $P^{X=x}$ 385
14.3 No Treatment for Joe: Conditional Expectations With Respect to $P^{X=x}$ 387

15.1 Joe and Ann: Reversed Average Logit Effect 431

16.1 Z-Conditional Independence of X and U 450
16.2 Joe and Ann With no Individual Treatment Effect 451

17.1 Joe and Ann With Self-Selection: Conditional Distribution $P_{Y|X}$ 469

List of Boxes

0.1 A First List of Symbols IX
1.1 Rules of Computation for Measures 19
4.1 Rules of Computation for Probabilities 129
4.2 Independence and Conditional Independence of Events 145
6.1 Rules of Computation for Expectations 204
6.2 Rules of Computation for Variances 207
7.1 Rules of Computation for Covariances 218
7.2 Rules of Computation for Expectations of Random Matrices 225
7.3 Rules of Computation for Covariance Matrices 226
9.1 Rules of Computation for B- Conditional Expectation Values 276
9.2 Rules of Computation for $(X{=}x)$-Conditional Expectation Values ... 277
10.1 Rules of Computation for $\mathscr{C}$-Conditional Expectations 297
10.2 Rules of Computation for X-Conditional Expectations 298
10.3 Monotonicity of Conditional Expectations 299
11.1 Rules of Computation for a Residual 323
11.2 Rules of Computation for $\mathscr{C}$-Conditional Covariances 334
11.3 Rules of Computation for Conditional Variances 335
14.1 P-Uniqueness of $E^B(Y|\mathscr{C})$ 403
16.1 Notation 445
16.2 Conditional Independence of Set Systems 448
16.3 Conditional Independence of Random Variables 449

Part I
Measure-Theoretical Foundations of Probability Theory

Chapter 1
Measure

In this chapter, we introduce the concept of a measure and other closely related notions. We start with some examples and then introduce the concept of a σ-*algebra*, which is crucial in measure theory and probability theory. At first glance this concept seems to be a pure technical construction, which is usually not dealt with in textbooks on 'Probability and Statistics' for empirical sciences. However, a σ-algebra turned out to be the natural domain for a measure, including probability measures. Moreover, in probability theory, a σ-algebra is not only the domain of probability measures. The σ-algebra generated by a random variable can be interpreted as the set of events that is represented by this random variable. This is treated in more detail in chapter 2 on measurable mappings, which provides the general theory of random variables because random variables are measurable mappings. The virtues of σ-algebras will become fully apparent in chapter 10 on conditional expectations and its subsequent chapters. The pair $(\Omega, \mathscr{A})$ consisting of a nonempty set Ω and a σ-algebra $\mathscr{A}$ on Ω is called a *measurable space*. Such a measurable space is crucial for the definition of a *measure*. Next, we treat some important examples of measures, including the *counting measure*, the *Dirac measure*, and the *Lebesgue measure*. Finally, we turn to *continuity* and *uniqueness* properties of a measure.

1.1 Introductory Examples

Consider Figure 1.1 showing the set Ω of all points (x, y) inside the rectangle and the sets A and B of all points (x, y) inside the two ellipses, respectively. These three sets are subsets of the plane $\mathbb{R}^2 := \mathbb{R} \times \mathbb{R}$, where $\mathbb{R}$ denotes the *set of all real numbers*, and $\mathbb{R} \times \mathbb{R} := \{(a, b) \colon a, b \in \mathbb{R}\}$ is the set of all ordered pairs (a, b) with $a, b \in \mathbb{R}$, called the *Cartesian product* or *product set* of $\mathbb{R}$ with itself. In Figure 1.1, the sets A and B have a nonempty intersection. Now let $\mathit{area}(A)$ and $\mathit{area}(B)$ denote their areas and $\mathit{area}(A \cap B)$ the area of their intersection. Inspecting this figure reveals:

$$\mathit{area}(A \cup B) \;=\; \mathit{area}(A) + \mathit{area}(B) - \mathit{area}(A \cap B).$$

This example illustrates three important points:

(a) A measure such as *area* is a function on a *set system on* Ω, i. e., on a *set of subsets* of a set Ω such as A, B, and $A \cap B$.

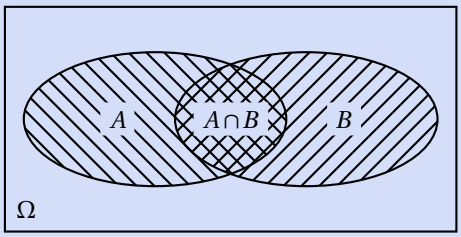


Figure 1.1. A Venn diagram of two sets and their intersection

(b) If *area* is defined for the *subsets* $A, B \subset \Omega$, then it is also defined for their *intersection* $A \cap B$ and for their *union* $A \cup B$.
(c) Measures are *additive*. In other words, if A and B are *disjoint* subsets of Ω, i. e., if $A \cap B = \emptyset$, then $\mathit{area}(A \cup B) = \mathit{area}(A) + \mathit{area}(B)$.

Note that, in the example presented in Figure 1.1, the sets A and B are *not disjoint*, and this is why $\mathit{area}(A \cap B)$ has to be subtracted in the equation displayed above. Points (a) to (c) also apply to other measures such as *length* and *volume* as well as to *probability measures*. Therefore, we adopt a more general language and talk about subsets A, B of a set Ω (or *measurable sets* A, B) and their *measure* μ instead of lines and their lengths, rectangles and their areas, cubes and their volume, or events and their probabilities.

For example, if $\Omega = \{1, \ldots, 6\}$ denotes the set of possible outcomes of tossing a fair dice, $A = \{1, 6\}$ and $B = \{2, 4, 6\}$ denote the events of tossing a 1 *or a* 6 and tossing an *even number*, respectively. Furthermore, $A \cap B = \{6\}$ and the probability of tossing a 1 *or a* 6 *or an even number* — the event $A \cup B$ — is

$$P(A \cup B) \;=\; P(A) + P(B) - P(A \cap B) \;=\; \frac{2}{6} + \frac{3}{6} - \frac{1}{6} \;=\; \frac{4}{6}.$$

In the first example, the measure *area* assigns a real number to a subset of $\mathbb{R}^2$. In the second example, the measure P assigns a real number to a subset of $\Omega = \{1, \ldots, 6\}$. This suggests that a measure should be defined such that it assigns a real number *to all subsets* of a set, i. e., to all elements of the power set. Unfortunately, this may lead to contradictions (see Rem. 1.70 and, e. g., Georgii, 2008, p. 9-10). In contrast, defining a measure on a σ-algebra such contradictions can be avoided.

1.2 σ-Algebra and Measurable Space

In the following definition, we consider a set system $\mathscr{A}$ on Ω, a sequence $A_1, A_2, \ldots$ of subsets of Ω, and their countable union. Remember, a *set system on a set* Ω is a set of subsets of Ω presuming that Ω is not empty. A *sequence of subsets of a set* Ω is a function from the set $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ or $\mathbb{N} = \{1, 2, \ldots\}$ or a subset of these sets

to $\mathscr{P}(\Omega)$, the *power set* of Ω. Furthermore, the *finite union* of the sets $A_1, \ldots, A_n$ and the *countable union* of the sets $A_1, A_2, \ldots$ are defined by

$$\bigcup_{i=1}^{n} A_i := \{a \in \Omega \colon \exists\, i \in \{1, \ldots, n\} \colon a \in A_i\} \tag{1.1}$$

and

$$\bigcup_{i=1}^{\infty} A_i := \{a \in \Omega \colon \exists\, i \in \mathbb{N} \colon a \in A_i\}, \tag{1.2}$$

respectively. Hence, by definition, $\bigcup_{i=1}^{n} A_i$ is the set of all elements that are an element of at least one of the sets A_i, $i = 1, \ldots, n$, and $\bigcup_{i=1}^{\infty} A_i$ is the set of all elements that are an element of at least one of the sets A_i, $i \in \mathbb{N}$. Finally, $A^c := \Omega \setminus A$ denotes the complement of A (with respect to Ω).

Definition 1.1 (σ-Algebra)
A set $\mathscr{A}$ of subsets of a nonempty set Ω is called a σ-algebra (or σ-field) on Ω, if the following three conditions hold:

(a) $\Omega \in \mathscr{A}$.
(b) If $A \in \mathscr{A}$, *then* $A^c \in \mathscr{A}$.
(c) If $A_1, A_2, \ldots \in \mathscr{A}$, *then* $\bigcup_{i=1}^{\infty} A_i \in \mathscr{A}$.

An element of a σ-algebra is called a measurable set.

Remark 1.2 (Closure With Respect to Set Operations) Condition (c) postulates that σ-algebras are closed with respect to *countable* unions of sets $A_1, A_2, \ldots \in \mathscr{A}$. However, in conjunction with (a) and (b) this implies that a σ-algebra is also closed with respect to *finite* unions of sets $A_1, \ldots, A_n \in \mathscr{A}$, because every finite union of sets $A_1, \ldots, A_n \in \mathscr{A}$ can be represented as a countable union of the sets that are elements of $\mathscr{A}$, e. g.,

$$\bigcup_{i=1}^{n} A_i = A_1 \cup \ldots \cup A_n \cup \emptyset \cup \emptyset \cup \ldots \ . \tag{1.3}$$

Note that (a) and (b) imply $\emptyset \in \mathscr{A}$, because $\Omega^c = \emptyset$.

Furthermore, although condition (c) only requires explicitly that σ-algebras are closed with respect to countable unions, Definition 1.1 implies that a σ-algebra is closed also with respect to intersections such as $A_1 \cap A_2$ and set differences $A_1 \setminus A_2$. In other words, if A_1 and A_2 are elements of $\mathscr{A}$, then $A_1 \cup A_2$, $A_1 \cap A_2$, and $A_1 \setminus A_2$ are elements of $\mathscr{A}$ as well, provided that $\mathscr{A}$ is a σ-algebra. The same is true for countable intersections $A_1 \cap A_2 \cap \ldots$ of elements of $\mathscr{A}$. In more formal terms: If $\mathscr{A}$ is a σ-algebra, then

$$A_1, A_2, \ldots \in \mathscr{A} \quad \Rightarrow \quad \bigcap_{i=1}^{\infty} A_i \in \mathscr{A} \tag{1.4}$$

(see Exercise 1-1), where $\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap \ldots$ is defined by

$$\bigcap_{i=1}^{\infty} A_i := \{a \in \Omega \colon \forall i \in \mathbb{N} \colon a \in A_i\}. \tag{1.5}$$

Because

$$\bigcap_{i=1}^{n} A_i = A_1 \cap \ldots \cap A_n \cap \Omega \cap \Omega \cap \ldots , \tag{1.6}$$

we can also conclude

$$A_1, \ldots, A_n \in \mathscr{A} \quad \Rightarrow \quad \bigcap_{i=1}^{n} A_i \in \mathscr{A}, \tag{1.7}$$

where $\bigcap_{i=1}^{n} A_i$, the *finite intersection* of the sets $A_1, \ldots, A_n$, is defined by

$$\bigcap_{i=1}^{n} A_i := \{a \in \Omega \colon \forall i \in \{1, \ldots, n\} \colon a \in A_i\}. \tag{1.8}$$

◁

Remark 1.3 (Countable and Uncountable Unions) Defining a σ-*algebra* we use the symbol σ in order to emphasize that unions of finitely or countably many sets are considered, *but not other unions of sets*. For example, the *closed interval* $[a,b] := \{x \in \mathbb{R} \colon a \le x \le b, a, b \in \mathbb{R},\}$ on the real axis is identical to the union of singletons $\{x\}$ that contain only one single element $x \in \mathbb{R}$, i. e.,

$$[a,b] = \bigcup_{a \le x \le b} \{x\}. \tag{1.9}$$

This union is neither finite nor countable. Hence, condition (c) of Definition 1.1 does *not* imply that this union is necessarily an element of a σ-algebra $\mathscr{A}$ on $\mathbb{R}$, even if all singletons $\{x\}$, $x \in \mathbb{R}$, are elements of $\mathscr{A}$. ◁

The following notion of a *measurable space* proves to be convenient in measure theory.

Definition 1.4 (Measurable Space)
If Ω is a nonempty set and $\mathscr{A}$ a σ-algebra on Ω, then the pair $(\Omega, \mathscr{A})$ is called a measurable space.

Example 1.5 (The Smallest σ-Algebra) The smallest σ-algebra on a set Ω is $\mathscr{A} = \{\Omega, \emptyset\}$. It contains only the elements Ω and the empty set $\emptyset$. As is easily seen, $\Omega \cup \emptyset = \Omega$, $\Omega^c = \emptyset$, and $\emptyset^c = \Omega$ are elements of $\mathscr{A}$. This shows that $\mathscr{A} = \{\Omega, \emptyset\}$ is closed with respect to union and complement. ◁

Example 1.6 (Power Set) The *power set* $\mathscr{P}(\Omega)$ of Ω, i. e., the *set of all subsets* of Ω, is always a σ-algebra on Ω. It is the largest σ-algebra on a set Ω. All other σ-algebras on Ω are subsets of $\mathscr{P}(\Omega)$. ◁

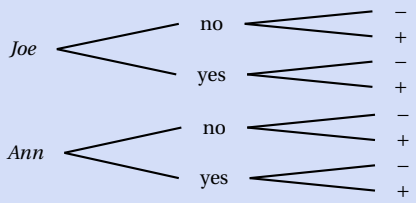


Figure 1.2. Example of a tree representation of a Cartesian product

Example 1.7 (A Small σ-Algebra) If A is a subset of Ω, then $\mathscr{A} = \{\Omega, \emptyset, A, A^c\}$ is always a σ-algebra on Ω (see Exercise 1-2). Again, it is easily seen that this set system is closed with respect to union and complement. ◁

Remark 1.8 (Motivation for σ-Algebras) These examples show that there can be many different σ-algebras on a set Ω. Why not simply always use the largest one, the power set $\mathscr{P}(\Omega)$? In fact, this would be possible as long as Ω is finite or countable. There are at least three reasons for using σ-algebras. *First*, there are important sets Ω (e. g., $\Omega = \mathbb{R}$) such that measures of interest (e. g., *length* — which is the Lebesgue measure pertaining to $\Omega = \mathbb{R}$) cannot be defined on $\mathscr{P}(\Omega)$ (see Rem. 1.70). These measures can be defined, however, on other σ-algebras, such as the Borel-σ-algebra [see Eq. (1.18)]. (For an example in which the power set is 'too large' see Georgii, 2008, p. 9-10). *Second*, in some sense, σ-algebras contain those elements of a larger set system that are relevant for a particular substantive question. In probability theory, together with Ω and a probability measure, each σ-algebra on Ω represents a random experiment that is in some sense contained in a (often larger) random experiment. For example, if we consider the random experiment of tossing a dice, then we may focus on whether or not the number is even. Together with Ω and the probability measure, the corresponding σ-algebra represents a 'new' random experiment contained in the random experiment of tossing a dice (see Exercise 1-3). *Third*, using different σ-algebras is indispensable for introducing conditional expectations, conditional independence, and conditional distributions (see chs. 9 to 17). ◁

Example 1.9 (Joe and Ann) Consider the following random experiment: *First*, we sample a unit u from the set $\Omega_U := \{Joe, Ann\}$. *Second*, each unit receives (*yes*) or does not receive a treatment (*no*). *Third*, it is observed whether (+) or not (−) a success criterion is reached (see Fig. 1.2). Defining $\Omega_X := \{yes, no\}$ and $\Omega_Y := \{+, -\}$, the Cartesian product

$$\Omega := \Omega_U \times \Omega_X \times \Omega_Y = \big\{ (Joe, no, -),\ (Joe, no, +), \ldots,\ (Ann, yes, +) \big\}$$

is the set of possible outcomes ω of this random experiment. It has eight elements, namely the triples $(Joe, no, -)$, $(Joe, no, +)$, …, $(Ann, yes, +)$ (see all eight leaves of Fig. 1.2 for a complete list of these elements).

In this example, a first σ-algebra $\mathscr{A}$ we may consider is *the set of all subsets of* Ω, the power set $\mathscr{P}(\Omega)$. This set has $2^8 = 256$ elements, where 8 is the number of elements, i. e., the cardinality of Ω (see Kheyfits, 2010, Th. 1.1.37, p. 22). Among these elements is the set

$$A := \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\} = \{Joe\} \times \Omega_X \times \Omega_Y.$$

In the context of probability theory it is also called the event that *Joe is drawn*. Other elements of $\mathscr{A}$ are the sets

$$B := \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\} = \Omega_U \times \{yes\} \times \Omega_Y.$$

that the *drawn person is treated*, and

$$C := \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\} = \Omega_U \times \Omega_X \times \{+\}$$

that $\{+\}$ *(success) occurs*, irrespective of which person is drawn and whether or not the person is treated.

Aside from the power set of Ω we could also consider the σ-algebras $\mathscr{A}_1 := \{\Omega, \emptyset, A, A^c\}$, $\mathscr{A}_2 := \{\Omega, \emptyset, B, B^c\}$, and $\mathscr{A}_3 := \{\Omega, \emptyset, C, C^c\}$, to name just three. (For another one see Exercise 1-4). In a sense, $\mathscr{A}_1$ represents the information which person is drawn. In contrast, $\mathscr{A}_2$ contains the information whether or not the drawn person is treated, and $\mathscr{A}_3$ whether or not the drawn person is successful. Of course, all these σ-algebras are subsets of $\mathscr{P}(\Omega)$, the power set of Ω. ◁

Example 1.10 (Trace of a Set System and Trace σ-Algebra) If $\mathscr{E}$ is a set system on Ω and $\Omega_0 \subset \Omega$, then

$$\mathscr{E}|_{\Omega_0} := \{\Omega_0 \cap A \colon A \in \mathscr{E}\}$$

is a set system on Ω_0. It is called the *trace of* $\mathscr{E}$ *in* Ω_0. Furthermore, if $\mathscr{A}$ is a σ-algebra on Ω and $\Omega_0 \subset \Omega$, then the set system

$$\mathscr{A}|_{\Omega_0} := \{\Omega_0 \cap A \colon A \in \mathscr{A}\}$$

is a σ-algebra on Ω_0 (see Exercise 1-5). If $\Omega \neq \Omega_0$, then the trace $\mathscr{A}|_{\Omega_0}$ is a σ-algebra on Ω_0, but not on Ω, because $\Omega \notin \mathscr{A}|_{\Omega_0}$. ◁

Example 1.11 (Joe and Ann – continued) In Example 1.9 we defined the event A that Joe is drawn, the event B that the drawn person is treated, and the σ-algebra $\mathscr{A}_2 = \{\Omega, \emptyset, B, B^c\}$. The trace of $\mathscr{A}_2$ in A is

$$\mathscr{A}_2|_A = \{A, \emptyset, A \cap B, A \cap B^c\}.$$

Obviously, just like all elements of $\mathscr{A}_2$ are subsets of Ω, all elements of $\mathscr{A}_2|_A$ are subsets of A. From a substantive point of view, considering $\mathscr{A}_2|_A$ means to presume that Joe is drawn and consider the events that he is treated or not treated, respectively. ◁

1.2.1 σ-Algebra Generated by a Set System

The concept of a σ-*algebra generated by a set system* is useful in order to define important σ-algebras. It is also useful for specifying certain measures (see section 1.6). The following theorem prepares Definition 1.13. Reading this theorem, remember that a σ-algebra on a set Ω is itself a set (of subsets of Ω), so that we can consider the intersection of σ-algebras.

Theorem 1.12 (Intersection of σ-Algebras is a σ-Algebra)
Let I be a nonempty (finite, countable, or uncountable) index set and let all $\mathcal{A}_i$, $i \in I$, be σ-algebras on Ω. Then $\bigcap_{i\in I} \mathcal{A}_i$ is also a σ-algebra on Ω.

(Proof p. 28)

This theorem allows us to define the σ-*algebra generated by a set system* on Ω.

Definition 1.13 (σ-Algebra Generated by a Set System)
Let $\mathcal{E}$ be a set system on Ω and let $(\mathcal{A}_i, i \in I)$ be the family of all σ-algebras on Ω that contain $\mathcal{E}$ as a subset. Then we define

$$\sigma(\mathcal{E}) := \bigcap_{i\in I} \mathcal{A}_i \tag{1.10}$$

and call it the σ-algebra generated by $\mathcal{E}$. The set $\mathcal{E}$ is also called a generating system of $\sigma(\mathcal{E})$.

Remark 1.14 (Smallest σ-Algebra Containing $\mathcal{E}$ as a Subset) According to Theorem 1.12, every set system $\mathcal{E}$ on Ω generates a uniquely defined σ-algebra $\sigma(\mathcal{E})$ on Ω. Note that the σ-algebra $\sigma(\mathcal{E})$ is the *smallest* σ-algebra on Ω containing $\mathcal{E}$ as a subset, i. e.,

$$\mathcal{C} \text{ is a σ-algebra on } \Omega \text{ and } \mathcal{E} \subset \mathcal{C} \quad \Rightarrow \quad \sigma(\mathcal{E}) \subset \mathcal{C}. \tag{1.11}$$

Furthermore,

$$\sigma[\sigma(\mathcal{E})] = \sigma(\mathcal{E}). \tag{1.12}$$

◁

The following lemma immediately follows from (1.11). It can be used in proofs of the identity of two σ-algebras.

Lemma 1.15 (Smallest σ-Algebra Containing $\mathcal{E}$ as a Subset)
Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathcal{E}$ a set system on Ω with $\sigma(\mathcal{E}) = \mathcal{A}$. If $\mathcal{C}$ is a σ-algebra on Ω with $\mathcal{E} \subset \mathcal{C} \subset \mathcal{A}$, then $\mathcal{C} = \mathcal{A}$.

(Proof p. 28)

Remark 1.16 (σ-Algebra Generated by Unions of Set Systems) Let $\mathscr{D}, \mathscr{E}, \mathscr{F}$ be set systems on a nonempty set Ω. Then

$$\sigma(\mathscr{D} \cup \mathscr{E} \cup \mathscr{F}) = \sigma[\mathscr{D} \cup \sigma(\mathscr{E} \cup \mathscr{F})] \tag{1.13}$$

(see Exercise 1-6). ◁

Example 1.17 (Several Set Systems May Generate the Same σ-Algebra) If A is a subset of Ω, then the set system $\{A\}$ generates the σ-algebra $\{\Omega, \emptyset, A, A^c\}$. Note that $\{\Omega, \emptyset, A, A^c\}$ is also generated by the set systems $\{A^c\}$ and $\{A, A^c\}$, for instance. Hence,

$$\sigma(\{A\}) = \sigma(\{A^c\}) = \sigma(\{A, A^c\}) = \sigma(\{\Omega, \emptyset, A, A^c\}) = \{\Omega, \emptyset, A, A^c\}.$$

In contrast, if $\emptyset \neq A \neq \Omega$, then the σ-algebra $\{\Omega, \emptyset, A, A^c\}$ is neither generated by the set system $\{\Omega\}$ nor by $\{\Omega, \emptyset\}$. Instead,

$$\sigma(\{\emptyset\}) = \sigma(\{\Omega\}) = \sigma(\{\Omega, \emptyset\}) = \{\Omega, \emptyset\},$$

i. e., $\{\Omega\}$, $\{\emptyset\}$, and $\{\Omega, \emptyset\}$ generate the σ-algebra $\{\Omega, \emptyset\}$. ◁

Example 1.18 (A Generator of the Power Set) Let Ω be finite or countable and let $\mathscr{E} := \big\{\{\omega\} : \omega \in \Omega\big\}$. Then $\sigma(\mathscr{E}) = \mathscr{P}(\Omega)$ (see Exercise 1-7). ◁

This example is generalized in Lemma 1.20.

Remark 1.19 (Partition) Reading the following lemma, remember that a set system $\mathscr{E}$ on Ω is called a *partition* of Ω if

(a) $\forall B \in \mathscr{E}: \ B \neq \emptyset$.
(b) $\forall B, C \in \mathscr{E}: \ B \neq C \ \Rightarrow \ B \cap C = \emptyset$.
(c) $\bigcup_{B \in \mathscr{E}} B = \Omega$.

◁

Lemma 1.20 (An Element of a σ-Algebra Generated by a Partition)
Let $\mathscr{E} := \{B_1, \ldots, B_n\}$ or $\mathscr{E} := \{B_1, B_2, \ldots\}$ be a finite or countable partition of Ω, respectively. Then for all $C \in \sigma(\mathscr{E})$ there is an $I(C) \subset \mathbb{N}$ such that

$$C = \bigcup_{i \in I(C)} B_i = \bigcup_{B_i \subset C} B_i, \tag{1.14}$$

where, by convention, $\bigcup_{i \in \emptyset} B_i := \emptyset$.

(Proof p. 29)

Remark 1.21 (Constructing a σ-Algebra) If $\mathscr{E} = \{A_1, \ldots, A_m\}$ is a finite set of subsets of Ω, then there is a finite partition $\mathscr{F} = \{B_1, \ldots, B_n\}$ of Ω with $\sigma(\mathscr{E}) = \sigma(\mathscr{F})$. Furthermore, if $\mathscr{E}$ is a *finite* set of subsets of Ω, then each element of $\sigma(\mathscr{E})$ is obtained by finitely many unions, intersections, or complements of elements of $\mathscr{E}$ (see Exercise 1-8). ◁

Example 1.22 (Joe and Ann – continued) In Example 1.11, we already considered the event A that Joe is drawn and noted that the trace of the σ-algebra $\mathscr{A}_2 = \{\Omega, \emptyset, B, B^c\}$ in A is $\mathscr{A}_2|_A = \{A, \emptyset, A\cap B, A\cap B^c\}$. In contrast, the σ-algebra on Ω generated by the trace $\mathscr{A}_2|_A$ is

$$\sigma(\mathscr{A}_2|_A) = \{\Omega, \emptyset, A, A^c, A\cap B, A\cap B^c, (A\cap B)\cup A^c, (A\cap B^c)\cup A^c\},$$

where $(A\cap B)\cup A^c = A^c\cup B$ and $(A\cap B^c)\cup A^c = A^c\cup B^c$. ◁

Remark 1.23 (Monotonicity of Generated σ-Algebras)
Let $\mathscr{E}_1, \mathscr{E}_2$ be set systems on Ω with $\mathscr{E}_1 \subset \mathscr{E}_2$. Then $\sigma(\mathscr{E}_1) \subset \sigma(\mathscr{E}_2)$ (see Exercise 1-9). ◁

An important kind of σ-algebras are those for which there is a countable set system that generates them.

Definition 1.24 (Countably Generated σ-Algebra)
Let $(\Omega, \mathscr{A})$ be a measurable space. Then $\mathscr{A}$ is called countably generated *if there is a finite or countable set $\mathscr{E} \subset \mathscr{A}$ such that $\sigma(\mathscr{E}) = \mathscr{A}$.*

Example 1.25 (Some Countably Generated σ-Algebras) Examples of countably generated σ-algebras are:

(a) All σ-algebras on a finite set Ω.
(b) $\mathscr{P}(\mathbb{N}_0^n)$, $n \in \mathbb{N}$.

(For a proof see Exercise 1-10). For another example, see Remark 1.28. ◁

Remark 1.26 (A Caveat) Note that there are countably generated σ-algebras for which not all of their elements can be constructed by countably many unions, intersections, or complements of elements of the generating system. An example in case are Borel σ-algebras on $\mathbb{R}$ or $\mathbb{R}^n$ (see, e. g., Michel, 1978, sect. I.4). ◁

Lemma 1.27 (σ-Algebra Generated by the Trace of a Set System)
Let $A \subset \Omega$ be nonempty, $\mathscr{E} \subset \mathscr{P}(\Omega)$, and $\mathscr{E}|_A := \{C\cap A\colon C \in \mathscr{E}\}$. Then

$$\sigma\big(\mathscr{E}|_A\big) = \sigma(\mathscr{E})|_A, \tag{1.15}$$

where $\sigma\big(\mathscr{E}|_A\big)$ denotes the σ-algebra generated on A, whereas $\sigma(\mathscr{E})$ is a σ-algebra on Ω. Furthermore, if $\mathscr{C}$ is a σ-algebra on Ω and $A \in \mathscr{E}$ such that

$$\forall C \in \mathscr{E}\colon \quad C \neq A \;\Rightarrow\; A\cap C = \emptyset, \tag{1.16}$$

i. e., A does not intersect with any other element of $\mathscr{E}$, then

$$\sigma(\mathscr{C}\cup\mathscr{E})|_A = \mathscr{C}|_A. \tag{1.17}$$

(Proof p. 29)

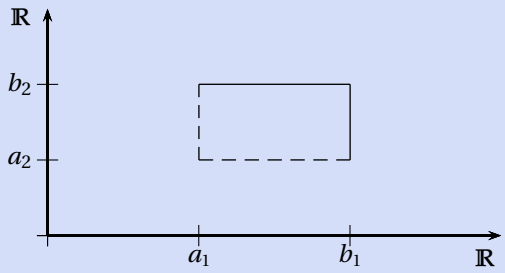


Figure 1.3. A half-open rectangle in the plane $\mathbb{R}^2$

Hence, according to Equation (1.15), the σ-algebra generated by the trace of a set system $\mathscr{E}$ is the trace of the σ-algebra generated by $\mathscr{E}$ and according to Equation (1.17), the trace of the σ-algebra $\sigma(\mathscr{C} \cup \mathscr{E})$ in the set A is identical to the trace of the σ-algebra $\mathscr{C}$ in A, if 1.16 holds.

1.2.2 σ-Algebra of Borel Sets on $\mathbb{R}^n$

For $a, b \in \mathbb{R}$ with $a < b$, let us consider a *half-open interval* $]a, b]$ in $\mathbb{R}$, which is defined by

$$]a, b] := \{x \in \mathbb{R} : a < x \leq b\},$$

and the *set system*

$$\mathscr{I}_1 := \{]a, b] : a, b \in \mathbb{R} \text{ and } a < b\}$$

of all half-open intervals in $\mathbb{R}$. The σ-algebra generated by this set system is called the *Borel σ-algebra* on $\mathbb{R}$. It is denoted by $\mathscr{B}$. The elements of $\mathscr{B}$ are called the *Borel sets* of $\mathbb{R}$. In formal terms,

$$\mathscr{B} := \mathscr{B}_1 := \sigma(\mathscr{I}_1). \tag{1.18}$$

Note that there are several sets systems generating the Borel σ-algebra (see, e.g., Klenke, 2008, Th. 1.23, p. 10). In particular,

$$\mathscr{B}_1 = \sigma\big(\{]-\infty, b] : b \in \mathbb{R}\}\big) \tag{1.19}$$

(see Georgii, 2008, p. 12). Similarly, we define the *Borel σ-algebra on* $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ to be the σ-algebra generated by the set system $\mathscr{I}_2$ of all half-open *rectangles* in $\mathbb{R}^2$, whose sides are parallel to the axes (see Fig. 1.3). These rectangles are defined by

$$]a_1, b_1] \times]a_2, b_2] = \{(x_1, x_2) \in \mathbb{R}^2 : a_1 < x_1 \leq b_1,\ a_2 < x_2 \leq b_2\}.$$

The σ-algebra $\sigma(\mathscr{I}_2)$ is denoted by $\mathscr{B}_2$, i. e., $\mathscr{B}_2 := \sigma(\mathscr{I}_2)$, and its elements are called the *Borel sets of* $\mathbb{R}^2$.

This definition is easily generalized: The *Borel σ-algebra on* $\mathbb{R}^n$ is defined by $\mathscr{B}_n := \sigma(\mathscr{I}_n)$, $n \in \mathbb{N}$, where $\mathscr{I}_n$ is the system of all half-open *cuboids* in $\mathbb{R}^n$, whose sides are parallel to the axes. Such a cuboid is a set

$$\begin{aligned} &]\,a_1, b_1\,] \times \ldots \times]\,a_n, b_n\,] \\ &= \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \colon a_1 < x_1 \le b_1, \ldots,\ a_n < x_n \le b_n\big\}, \end{aligned} \tag{1.20}$$

where $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$. Just like $\mathscr{B}_1$, the σ-algebra $\mathscr{B}_n$ has several generating systems, one of which is

$$\mathscr{B}_n \;=\; \sigma\big(\big\{]-\infty, b_1\,] \times \ldots \times]-\infty, b_n\,] \colon b_1, \ldots, b_n \in \mathbb{R}\big\}\big) \tag{1.21}$$

(see Exercise 1-11).

Note that not every subset of $\mathbb{R}^n$ is a Borel set. In other words, $\mathscr{B}_n$ is not the power set of $\mathbb{R}^n$. (The reason for using the Borel σ-algebra instead of the power set is illustrated in Remark 1.70.) However, for each $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the singleton $\{x\}$ *is* a Borel set of $\mathbb{R}^n$, i. e.,

$$\{x\} \in \mathscr{B}_n, \quad \forall\, x \in \mathbb{R}^n$$

(see Exercise 1-12).

Furthermore, if $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ denotes the *extended set of real numbers*, then

$$\overline{\mathscr{B}} \;:=\; \sigma\big(\mathscr{B} \cup \big\{\{-\infty\}, \{+\infty\}\big\}\big)$$

is a σ-algebra on $\overline{\mathbb{R}}$ and it is called the *Borel σ-algebra on* $\overline{\mathbb{R}}$. Similarly, $\overline{\mathscr{B}}_n$ is called the *Borel σ-algebra on* $\overline{\mathbb{R}}^n$. It is defined as the product of the σ-algebra $\overline{\mathscr{B}}$ with itself (n times) (see Def. 1.31). Finally, we may sometimes consider $\overline{\mathscr{B}}_n|_{\Omega_0}$, the *trace of the Borel σ-algebra on* $\overline{\mathbb{R}}^n$ *in* $\Omega_0 \subset \overline{\mathbb{R}}^n$.

Remark 1.28 (The Borel σ-Algebra is Countably Generated) Note that

$$\mathscr{B} \;=\; \sigma\big(\big\{]\,a, b] \colon a, b \in \mathbb{Q}, a < b\big\}\big),$$

where $\mathbb{Q}$ is the set of rational numbers. Because $\mathbb{Q}$ is countable, the set of intervals $\big\{]\,a, b] \colon a, b \in \mathbb{Q}, a < b\big\}$ is countable as well. Therefore, the Borel σ-algebra $\mathscr{B}$ is countably generated. This also holds for $\mathscr{B}_n$, $n \in \mathbb{N}$ (see Klenke, 2008, Th. 1.23, p. 10). ◁

Remark 1.29 (Trace of the Borel σ-Algebra in a Countable Subset of $\mathbb{R}$) Let $\mathscr{B}$ denote the Borel σ-algebra on $\mathbb{R}$. If $\Omega_0 \subset \mathbb{R}$ is finite or countable, then $\mathscr{B}|_{\Omega_0} = \mathscr{P}(\Omega_0)$, where $\mathscr{B}|_{\Omega_0}$ is the trace of the Borel σ-algebra on $\mathbb{R}$ in $\Omega_0 \subset \mathbb{R}$ (see Exercise 1-13). ◁

1.2.3 σ-Algebra on a Cartesian Product

In section 1.2.2 we defined a σ-algebra on $\mathbb{R}^n = \mathbb{R} \times \ldots \times \mathbb{R}$ (n-times). Now we consider σ-algebras on general Cartesian products. We start with an example.

Example 1.30 (Joe and Ann – continued) In Example 1.9 we already considered the Cartesian product

$$\Omega := \Omega_U \times \Omega_X \times \Omega_Y,$$

which consists of the eight triples $(Joe, no, -)$, $(Joe, no, +), \ldots,$ $(Ann, yes, +)$ (see again Fig. 1.2). Now consider the σ-algebras $\mathscr{A}_1 := \mathscr{P}(\Omega_U)$, $\mathscr{A}_2 := \mathscr{P}(\Omega_X)$, and $\mathscr{A}_3 := \mathscr{P}(\Omega_Y)$, as well as the set

$$\mathscr{E} := \big\{A_1 \times A_2 \times A_3 \colon A_1 \in \mathscr{A}_1,\, A_2 \in \mathscr{A}_2,\, A_3 \in \mathscr{A}_3\big\},$$

which is a set system on Ω consisting of $4 \cdot 4 \cdot 4 = 64$ elements. For example, the set system $\mathscr{E}$ contains the elements

$$A := \{Joe\} \times \{no\} \times \{-\} = \{(Joe, no, -)\}$$

and

$$B := \{Ann\} \times \{yes\} \times \{+\} = \{(Ann, yes, +)\}.$$

However, $\mathscr{E}$ does not contain

$$A \cup B = \big\{(Joe, no, -), (Ann, yes, +)\big\}$$

as an element. The only product set $A_1 \times A_2 \times A_3$ with $A_1 \in \mathscr{A}_1$, $A_2 \in \mathscr{A}_2$, $A_3 \in \mathscr{A}_3$ that contains $A \cup B$ as a subset is $\Omega_U \times \Omega_X \times \Omega_Y = \Omega$. However, $A \cup B \neq \Omega$. Therefore, $\mathscr{E}$ is not a σ-algebra [cf. condition (c) of Rem. 1.2]. In this example, the σ-algebra generated by $\mathscr{E}$ is the power set of Ω, i. e., $\sigma(\mathscr{E}) = \mathscr{P}(\Omega)$. It consists of $2^8 = 256$ elements. According to the following definition, $\sigma(\mathscr{E})$ is denoted by $\mathscr{A}_1 \otimes \mathscr{A}_2 \otimes \mathscr{A}_3$ and called the *product σ-algebra of* $\mathscr{A}_1$, $\mathscr{A}_2$, *and* $\mathscr{A}_3$. ◁

Definition 1.31 (Product σ-Algebra)
Let $(\Omega_1, \mathscr{A}_1), \ldots, (\Omega_n, \mathscr{A}_n)$ *be measurable spaces and* $\Omega := \Omega_1 \times \ldots \times \Omega_n$. *Then*

$$\mathscr{A}_1 \otimes \ldots \otimes \mathscr{A}_n := \bigotimes_{i=1}^{n} \mathscr{A}_i := \sigma\Big(\Big\{ \bigtimes_{i=1}^{n} A_i \colon A_i \in \mathscr{A}_i, i = 1, \ldots, n\Big\}\Big) \tag{1.22}$$

is called the product σ-algebra *of the* σ-algebras $\mathscr{A}_i$, $i = 1, \ldots, n$.

To emphasize, the product σ-algebra $\mathscr{A}_1 \otimes \ldots \otimes \mathscr{A}_n$ *is not* the Cartesian product $\mathscr{A}_1 \times \ldots \times \mathscr{A}_n$. Instead, the product σ-algebra is generated by the set system of all Cartesian products of elements of the σ-algebras $\mathscr{A}_1, \ldots, \mathscr{A}_n$. In Lemma 2.42 we give an equivalent specification of a product σ-algebra, using projection mappings.

The following lemma provides a relationship between the generating systems of the σ-algebras $\mathscr{A}_i$, $i = 1, \ldots, n$, and the generating system of the product σ-algebra.

Lemma 1.32 (Generating System of a Product σ-Algebra)
For $i = 1,\dots,n$, let $(\Omega_i, \mathscr{A}_i)$ be measurable spaces and $\mathscr{E}_i \subset \mathscr{A}_i$ with $\sigma(\mathscr{E}_i) = \mathscr{A}_i$. Then

$$\bigotimes_{i=1}^{n} \mathscr{A}_i \;=\; \sigma\Big(\Big\{ \bigtimes_{i=1}^{n} A_i \colon A_i \in \mathscr{E}_i, i = 1,\dots,n \Big\}\Big). \tag{1.23}$$

For a proof see Klenke (2008, Th. 14.12 (i), p. 274).

This lemma implies

$$\mathscr{B}_n \;=\; \bigotimes_{i=1}^{n} \mathscr{B} \;=\; \mathscr{B} \otimes \ldots \otimes \mathscr{B} \quad (n\text{-times})$$

for the Borel σ-algebra on $\mathbb{R}^n$. This lemma also implies the following corollary.

Corollary 1.33 (Countable Generating System of a Product σ-Algebra)
Let $(\Omega_i, \mathscr{A}_i)$, $i = 1,\dots,n$, be measurable spaces, where all $\mathscr{A}_i$ are countably generated. Then $\bigotimes_{i=1}^{n} \mathscr{A}_i$ is countably generated as well.

Example 1.34 (Countable Sets and Product σ-Algebra) Let $\Omega_1,\dots,\Omega_n$ be finite or countable sets and $\mathscr{A}_1,\dots,\mathscr{A}_n$ be their power sets. Then

$$\bigotimes_{i=1}^{n} \mathscr{A}_i \;=\; \mathscr{P}\Big(\bigtimes_{i=1}^{n} \Omega_i \Big),$$

i. e., $\bigotimes_{i=1}^{n} \mathscr{A}_i$ is the power set on $\Omega := \Omega_1 \times \ldots \times \Omega_n$ (see Exercise 1-14). ◁

Remark 1.35 (Complement of a Cartesian Product) Let $(\Omega_1 \times \Omega_2, \mathscr{A}_1 \otimes \mathscr{A}_2)$ be a measurable space, $A \in \mathscr{A}_1$, and $B \in \mathscr{A}_2$. Then $(A \times B)^c \in \mathscr{A}_1 \otimes \mathscr{A}_2$, and this set can be written

$$(A \times B)^c \;=\; (A^c \times B) \cup (\Omega_1 \times B^c), \tag{1.24}$$

which is a union of disjoint sets (see Exercise 1-15). ◁

1.2.4 ∩-Stable Set Systems That Generate a σ-Algebra

For many proofs, generating set systems are useful, which have the property of ∩-stability.

Definition 1.36 (∩-Stability)
Let Ω denote a nonempty set. A set $\mathscr{E}$ of subsets of Ω is called ∩-stable (or ∩-closed) if $A \cap B \in \mathscr{E}$ for all $A, B \in \mathscr{E}$.

Example 1.37 (Set System With One Single Element) A set system $\{A\}$ that has only a single element $A \subset \Omega$ is ∩-stable (cf. Example 1.17). ◁

Example 1.38 (Partition and $\cap$-Stability) If $\mathscr{E}$ is a partition of the set Ω, then $\mathscr{D} := \mathscr{E} \cup \{\emptyset\}$ is $\cap$-stable. $\lhd$

Example 1.39 (A $\cap$-Stable Generating System of a Product σ-Algebra) Consider the measurable spaces $(\Omega_i, \mathscr{A}_i)$, $i = 1, \ldots, n$. The set

$$\{A_1 \times \ldots \times A_n : A_i \in \mathscr{A}_i,\ i = 1, \ldots, n\},$$

is a $\cap$-stable generating system of $\bigotimes_{i=1}^n \mathscr{A}_i$ (see Exercise 1-16). $\lhd$

Another type of a set system is a Dynkin system. It can be used in order to show that a specific set system is a σ-algebra.

Definition 1.40 (Dynkin System)
A set $\mathscr{D}$ of subsets of a set Ω is called a Dynkin system on Ω, if the following three conditions hold:

(a) $\Omega \in \mathscr{D}$.
(b) *If* $A \in \mathscr{D}$, *then* $A^c \in \mathscr{D}$.
(c) *If* $A_1, A_2, \ldots \in \mathscr{D}$ *and they are pairwise disjoint, then* $\bigcup_{i=1}^\infty A_i \in \mathscr{D}$.

In the definition of a σ-algebra $\mathscr{A}$ we require $\bigcup_{i=1}^\infty A_i \in \mathscr{A}$ for all sequences $A_1, A_2, \ldots \in \mathscr{A}$, whereas for a Dynkin system the corresponding requirement is only made for all sequences $A_1, A_2, \ldots \in \mathscr{D}$ of pairwise disjoint sets. According to the following theorem a Dynkin system is also a σ-algebra if and only if it is $\cap$-stable.

Theorem 1.41 (Dynkin System and σ-Algebra)
Let $\mathscr{D}$ be a Dynkin system on Ω. Then $\mathscr{D}$ is a σ-algebra if and only if it is $\cap$-stable.

For a proof see Bauer (2001, Th. 2.3, p. 6). According to this theorem we can prove that a set system is a σ-algebra by showing that it is a $\cap$-stable Dynkin system.

1.3 Measure and Measure Space

A measure assigns to all elements of a σ-algebra an element of the closed interval

$$[0, \infty] := \{x \in \mathbb{R} : 0 \le x\} \cup \{\infty\},$$

i. e., a nonnegative real number or the element ∞.

Example 1.42 (A First Example) Let $\Omega = \mathbb{R}$ and assume that the closed interval $[3,9] = \{x \in \mathbb{R} : 3 \le x \le 9\}$ as well as the union $[3,9] \cup [10,12]$ are elements of a σ-algebra on Ω. If the measure is *length*, then

$$\mathit{length}([3,9]) = 9-3 = 6$$

and

$$\begin{aligned}\mathit{length}([3,9] \cup [10,12]) &= \mathit{length}([3,9]) + \mathit{length}([10,12]) \\ &= (9-3) + (12-10) = 6+2 = 8,\end{aligned}$$

because the two intervals are disjoint, i. e., their intersection is the empty set $\emptyset$. In this case the lengths of the intervals $[3,9]$ and $[10,12]$ add up to the length of their union $[3,9] \cup [10,12]$. In Definition 1.43 (c) we do not only require additivity but σ-additivity. ◁

Reading the following definition, remember that $\sum_{i=1}^{\infty} a_i$ is defined by

$$\sum_{i=1}^{\infty} a_i := \lim_{n\to\infty} \sum_{i=1}^{n} a_i.$$

Definition 1.43 (Measure and Measure Space)
Let $(\Omega, \mathscr{A})$ be a measurable space. A function $\mu\colon \mathscr{A} \to \overline{\mathbb{R}}$ is called a measure and the triple $(\Omega, \mathscr{A}, \mu)$ is called a measure space, if

(a) $\mu(\emptyset) = 0$.
(b) $\mu(A) \ge 0$, $\forall\, A \in \mathscr{A}$. *(nonnegativity)*
(c) If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint, then $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$. ($\sigma$-additivity)

1.3.1 σ-Additivity and Related Properties

Remark 1.44 (σ-Additivity Implies Finite Additivity) Note that σ-additivity of a measure implies finite additivity, i. e., it implies

$$\mu\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mu(A_i), \quad \text{if } A_1, \ldots, A_n \in \mathscr{A} \text{ are pairwise disjoint} \tag{1.25}$$

[see Rule (ii) of Box 1.1 and its proof in Exercise 1-18]. ◁

Remark 1.45 (σ-Additivity) Using the term *σ-additivity* signalizes that unions of finitely or countably many sets are considered, but not other unions of sets. If, instead of σ-additivity, we would require additivity for *any kind of unions*, including uncountable unions, then the Lebesgue measure λ on $(\mathbb{R}, \mathscr{B})$ — the measure of *length* — could not be constructed any more. This is explained in more detail in Remark 1.70. ◁

Remark 1.46 (Representation of a Union as a Union of Pairwise Disjoint Sets) Let $(\Omega, \mathscr{A})$ be a measurable space. If $A_1, A_2, \ldots \in \mathscr{A}$ is a sequence of subsets of Ω, then there is a sequence $B_1, B_2, \ldots \in \mathscr{A}$ of pairwise disjoint sets with

$$\bigcup_{i=1}^{\infty} A_i \;=\; \bigcup_{i=1}^{\infty} B_i. \tag{1.26}$$

One way to construct $B_1, B_2, \ldots$ is to define $B_1 := A_1$ and

$$B_i \;:=\; A_i \setminus \left(\bigcup_{j=1}^{i-1} A_j \right), \quad \text{for } i > 1, \tag{1.27}$$

(see Exercise 1-17). ◁

Remark 1.47 (Additivity of Measures for Partitions) Let $(\Omega, \mathscr{A}, \mu)$ be a measure space, $B \in \mathscr{A}$, and assume

(a) $A_1, \ldots, A_n \in \mathscr{A}$ are pairwise disjoint,
(b) $B \subset \bigcup_{i=1}^{n} A_i$.

Then

$$\mu(B) \;=\; \sum_{i=1}^{n} \mu(B \cap A_i). \tag{1.28}$$

Analogously, if

(c) $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint,
(d) $B \subset \bigcup_{i=1}^{\infty} A_i$,

then

$$\mu(B) \;=\; \sum_{i=1}^{\infty} \mu(B \cap A_i). \tag{1.29}$$

(see Exercise 1-19). ◁

1.3.2 Other Properties

Other important properties of a measure are displayed in Box 1.1. Some of these properties can intuitively be understood inspecting the Venn diagram presented in Figure 1.1. These properties always hold with the conventions $+\infty + \infty = +\infty$ and $\alpha + \infty = +\infty$, for $\alpha \in \mathbb{R}$. However, note that the term $+\infty - \infty$ cannot meaningfully be defined. Therefore, properties (vi) and (vii) only hold if we assume $\mu(A \cap B) < \infty$. For proofs of all theses properties see Exercise 1-18.

Remark 1.48 (Finite Additivity and σ-Additivity Applied to Singletons) If Ω is finite or countable, then each $A \subset \Omega$ is finite or countable as well. Hence, for any measure μ on the measurable space $\big(\Omega, \mathscr{P}(\Omega)\big)$,

$$\mu(A) \;=\; \mu\Big(\bigcup_{\omega \in A} \{\omega\} \Big) \;=\; \sum_{\omega \in A} \mu(\{\omega\}), \quad \forall\, A \subset \Omega. \tag{1.30}$$

Box 1.1 Rules of Computation for Measures

Let $(\Omega, \mathscr{A}, \mu)$ be a measure space.

If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint, then

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i). \qquad (\sigma\text{-}\textit{additivity}) \quad \text{(i)}$$

$$\mu\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mu(A_i), \quad \forall n \in \mathbb{N}. \qquad (\textit{finite additivity}) \quad \text{(ii)}$$

If $A, B \in \mathscr{A}$, then:

$$\mu(A) = \mu(A \cap B) + \mu(A \setminus B). \quad \text{(iii)}$$

$$\mu(\Omega) = \mu(B) + \mu(B^c). \quad \text{(iv)}$$

$$\mu(A) \leq \mu(B), \quad \text{if } A \subset B. \qquad (\textit{monotonicity}) \quad \text{(v)}$$

$$\mu(A \setminus B) = \mu(A) - \mu(A \cap B), \quad \text{if } \mu(A \cap B) < \infty. \quad \text{(vi)}$$

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B), \quad \text{if } \mu(A \cap B) < \infty. \quad \text{(vii)}$$

$$\mu(A) = \mu(\Omega) < \infty \;\Rightarrow\; \mu(A \cap B) = \mu(B). \quad \text{(viii)}$$

$$\mu(A) = 0 \;\Rightarrow\; \mu(A \cup B) = \mu(B). \quad \text{(ix)}$$

Let $A \in \mathscr{A}$ and let $\Omega_0 \subset \Omega$ and be finite or countable with $\mu(\Omega \setminus \Omega_0) = 0$.
If, for all $\omega \in \Omega_0$, $\{\omega\} \in \mathscr{A}$, then

$$\mu(A) = \sum_{\omega \in A \cap \Omega_0} \mu(\{\omega\}). \quad \text{(x)}$$

If $A_1, A_2, \ldots \in \mathscr{A}$, then

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mu(A_i). \qquad (\sigma\text{-}\textit{subadditivity}) \quad \text{(xi)}$$

This means that a measure on $\big(\Omega, \mathscr{P}(\Omega)\big)$ is already uniquely defined if its values $\mu(\{\omega\})$ are uniquely defined for all $\omega \in \Omega$, provided that Ω is finite or countable. Rule (x) of Box 1.1 extends this result to a more general measure space $(\Omega, \mathscr{A}, \mu)$. This rule shows that a measure on $(\Omega, \mathscr{A})$ is already uniquely defined if its values $\mu(\{\omega\})$ are uniquely defined for all $\omega \in \Omega_0$, provided that Ω_0 is finite or countable with $\mu(\Omega \setminus \Omega_0) = 0$ and $\{\omega\} \in \mathscr{A}$ for all $\omega \in \Omega_0$. $\lhd$

1.4 Specific Measures

Now we consider some examples of measures, all of which are used later on in order to introduce still other measures. For some of these examples we use the *indicator* of a set A.

Definition 1.49 (Indicator)
Let Ω be a set and $A \subset \Omega$. Then the function $1_A \colon \Omega \to \mathbb{R}$ defined by

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A, \end{cases} \tag{1.31}$$

is called the indicator of A.

Remark 1.50 (Sums and Products of Indicators) If $1_A, 1_B \colon \Omega \to \mathbb{R}$ are the indicators of two sets $A, B \subset \Omega$, then

$$1_A \cdot 1_B = 1_{A \cap B} \tag{1.32}$$

and

$$1_A + 1_B - 1_{A \cap B} = 1_A + 1_B - 1_A \cdot 1_B = 1_{A \cup B}. \tag{1.33}$$

Equation (1.32) immediately implies

$$1_A + 1_B = 1_{A \cup B}, \quad \text{if } A \cap B = \emptyset. \tag{1.34}$$

More generally, if $A_1, \ldots, A_n$ is a finite sequence of pairwise disjoint subsets of Ω, then

$$\sum_{i=1}^{n} 1_{A_i} = 1_{\bigcup_{i=1}^{n} A_i}, \tag{1.35}$$

i. e., then the sum of the indicators of the sets $A_1, \ldots, A_n$ is the indicator of the union $\bigcup_{i=1}^{n} A_i$. Finally, if $A_1, A_2, \ldots$ is a sequence of pairwise disjoint subsets of Ω, then

$$\sum_{i=1}^{\infty} 1_{A_i} = 1_{\bigcup_{i=1}^{\infty} A_i}. \tag{1.36}$$

◁

Remark 1.51 (Indicators of Products Sets) Let Ω_1, Ω_2 be nonempty sets, $A \subset \Omega_1$ and $B \subset \Omega_2$. Then

$$1_A(\omega_1) \cdot 1_B(\omega_2) = 1_{A \times B}(\omega_1, \omega_2), \quad \forall (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2. \tag{1.37}$$

This equation follows from the definitions of the product set and the indicator. ◁

1.4.1 Dirac Measure and Counting Measure

Example 1.52 (Dirac Measure) Let $(\Omega, \mathscr{A})$ be a measurable space, let $\omega \in \Omega$, and consider the function $\delta_\omega \colon \mathscr{A} \to \{0,1\}$ defined by

$$\delta_\omega(A) := 1_A(\omega), \quad \forall\, A \in \mathscr{A}. \tag{1.38}$$

Then δ_ω is a measure on $(\Omega, \mathscr{A})$ (see Exercise 1-20). ◁

Definition 1.53 (Dirac Measure)
The function δ_ω defined by Equation (1.38) *is called the Dirac measure at (point)* ω.

Example 1.54 (Counting Measure) Let $(\Omega, \mathscr{A})$ be a measurable space and define the function $\mu_\# \colon \mathscr{A} \to \overline{\mathbb{R}}$ by

$$\mu_\#(A) := \begin{cases} \sum_{\omega \in \Omega} 1_A(\omega), & \text{if } A \text{ is finite,} \\ \infty, & \text{if } A \text{ is infinite,} \end{cases} \qquad \forall\, A \in \mathscr{A}. \tag{1.39}$$

Then $\mu_\#$ is a measure on $(\Omega, \mathscr{A})$ (see Exercise 1-21). ◁

Definition 1.55 (Counting Measure)
The function $\mu_\#$ defined by Equation (1.39) *is called the counting measure on* $(\Omega, \mathscr{A})$.

Remark 1.56 (Cardinality of a Set) If A is finite, then $\mu_\#(A)$ is called the *cardinality* of A, i. e., $\mu_\#(A)$ simply counts the number of elements ω of the set A. Furthermore, for finite or countable Ω and $A \subset \Omega$,

$$\mu_\#(A) = \sum_{\omega \in \Omega} 1_A(\omega) = \sum_{\omega \in \Omega} \delta_\omega(A). \tag{1.40}$$

◁

Example 1.57 (Sum of Dirac Measures) Let $(\Omega, \mathscr{A})$ be a measurable space. If $B \subset \Omega$ is finite or countable and δ_ω is the Dirac measure on $(\Omega, \mathscr{A})$ at point ω, then $\sum_{\omega \in B} \delta_\omega \colon \mathscr{A} \to [0, \infty]$ defined by

$$\Big(\sum_{\omega \in B} \delta_\omega \Big)(A) := \sum_{\omega \in B} \delta_\omega(A), \quad \forall\, A \in \mathscr{A}, \tag{1.41}$$

is a measure on $(\Omega, \mathscr{A})$ (see Exercise 1-22). Hence, if Ω itself is finite or countable, then $\sum_{\omega \in \Omega} \delta_\omega$ is a measure on $(\Omega, \mathscr{A})$, and it is identical to the counting measure defined in Example 1.54, because, for $A \in \mathscr{A}$,

$$\begin{aligned}\Big(\sum_{\omega\in\Omega}\delta_\omega\Big)(A) &= \sum_{\omega\in\Omega}\delta_\omega(A) && [(1.41)]\\ &= \sum_{\omega\in\Omega}\mathbb{1}_A(\omega) && [(1.38)]\\ &= \mu_\#(A). && [(1.40)]\end{aligned} \tag{1.42}$$

◁

1.4.2 Lebesgue Measure

Consider the *half-open interval* $]a,b]$. Then

$$\lambda_1\big(]a,b]\big) = b-a \tag{1.43}$$

is the *length* of the interval $]a,b]$. Next consider a *rectangle* $]a_1,b_1] \times]a_2,b_2]$ in $\mathbb{R}^2$ with $a_1 < b_1$ and $a_2 < b_2$. This set can be visualized by the set of all points inside the rectangle presented in Figure 1.3 (excluding the lower and left boundary). Obviously,

$$\lambda_2\big(]a_1,b_1] \times]a_2,b_2]\big) = (b_1-a_1)\cdot(b_2-a_2) \tag{1.44}$$

is the *area* of this rectangle.

According to the following theorem, there is one and only one measure on $(\mathbb{R},\mathscr{B})$ satisfying (1.43) for all such intervals. This measure is called the *Lebesgue measure on* $(\mathbb{R},\mathscr{B})$ and is denoted by λ or λ_1. Similarly, there is one and only one measure on $(\mathbb{R}^2,\mathscr{B}_2)$ satisfying (1.44) for all such rectangles. It is called the *Lebesgue measure on* $(\mathbb{R}^2,\mathscr{B}_2)$ and is denoted by λ_2. The following theorem deals with the general case.

Theorem 1.58 (Existence and Uniqueness of the Lebesgue Measure)
For all $n \in \mathbb{N}$, there is a uniquely defined measure λ_n on $(\mathbb{R}^n,\mathscr{B}_n)$ satisfying

$$\begin{gathered}\lambda_n\big(]a_1,b_1] \times \ldots \times]a_n,b_n]\big) = \prod_{i=1}^{n}(b_i-a_i),\\ \forall\, a_i,b_i \in \mathbb{R} \text{ with } a_i < b_i,\ i=1,\ldots,n.\end{gathered} \tag{1.45}$$

For a proof see Klenke (2008, Th. 1.55, p. 25 and 26).

Definition 1.59 (Lebesgue Measure)
The measure λ_n satisfying Equation (1.45) *is called the* Lebesgue measure *on $(\mathbb{R}^n,\mathscr{B}_n)$.*

1.4.3 Other Examples of a Measure

Example 1.60 (Restriction of a Measure to a Sub-σ-Algebra) Suppose $(\Omega,\mathscr{A},\mu)$ is a measure space and $\mathscr{C} \subset \mathscr{A}$ a σ-algebra. Then the function $\nu\colon \mathscr{C} \to \overline{\mathbb{R}}$ defined

by

$$\nu(A) := \mu(A), \quad \forall A \in \mathscr{C}, \tag{1.46}$$

is a measure on $(\Omega, \mathscr{C})$ (see Exercise 1-23). ◁

Example 1.61 (Weighted Sum of Measures) If $\mu_1, \mu_2, \ldots$ are measures on $(\Omega, \mathscr{A})$ and $0 \le \alpha_1, \alpha_2, \ldots \in \mathbb{R}$, then $\sum_{i=1}^{\infty} \alpha_i \mu_i : \mathscr{A} \to [0, \infty]$ defined by

$$\Big(\sum_{i=1}^{\infty} \alpha_i \mu_i\Big)(A) := \sum_{i=1}^{\infty} \alpha_i \mu_i(A), \quad \forall A \in \mathscr{A}, \tag{1.47}$$

is again a measure on $(\Omega, \mathscr{A})$ (see Exercise 1-24). For $0 = \alpha_{n+1} = \alpha_{n+2} = \ldots$ this implies: If $\mu_1, \ldots, \mu_n$ are measures on $(\Omega, \mathscr{A})$ and $\alpha_1, \ldots, \alpha_n$ are nonnegative, then the function $\sum_{i=1}^{n} \alpha_i \mu_i$ defined by

$$\Big(\sum_{i=1}^{n} \alpha_i \mu_i\Big)(A) := \sum_{i=1}^{n} \alpha_i \mu_i(A), \quad \forall A \in \mathscr{A}, \tag{1.48}$$

is also a measure on $(\Omega, \mathscr{A})$. ◁

1.4.4 Finite and σ-Finite Measures

A measure μ on a measurable space $(\Omega, \mathscr{A})$ is called *finite* if $\mu(\Omega) < \infty$. Otherwise it is called *infinite*. Within the class of infinite measures there is a subclass with an important property, called σ-*finiteness*. Many fundamental propositions of measure and integration theory only hold for measures that are σ-finite.

Definition 1.62 (σ-Finite Measure)
Let μ be a measure on a measurable space $(\Omega, \mathscr{A})$. Then μ is called σ-finite if there is a sequence $A_1, A_2, \ldots \in \mathscr{A}$ with $\bigcup_{i=1}^{\infty} A_i = \Omega$ and, for all $i = 1, 2, \ldots$, $\mu(A_i) < \infty$.

To emphasize, even if $\mu(\Omega) = \infty$, the measure μ can be σ-finite (see Examples 1.63 and 1.64). Note that any finite measure is also σ-finite.

Example 1.63 (σ-Finiteness of the Lebesgue-Measure) The Lebesgue measure λ on $(\mathbb{R}, \mathscr{B})$ is σ-finite, because $\mathbb{R} = \bigcup_{i=1}^{\infty} [-i, i]$ and $\lambda\big([-i, i]\big) = 2 \cdot i < \infty$, for all $i \in \mathbb{N}$. ◁

Example 1.64 (A σ-Finite Counting Measure) Consider the measurable space $(\mathbb{R}, \mathscr{B})$ and the measure $\mu : \mathscr{B} \to [0, \infty]$, where $\mu = \sum_{i=0}^{\infty} \delta_i$ and δ_i denotes the Dirac measure at i on $(\mathbb{R}, \mathscr{B})$ with $\delta_i(A) = 1_A(i)$, $A \in \mathscr{B}$, $i \in \mathbb{N}_0$ (see Example 1.57). Then μ is σ-finite because $\mathbb{R} = \bigcup_{n=1}^{\infty} [-n, n]$ and $\mu\big([-n, n]\big) = n + 1$, for all $n \in \mathbb{N}_0$. This measure simply counts the number of elements $i \in \mathbb{N}_0$ in a Borel set A. In other words, for all finite $A \in \mathscr{B}$, $\mu(A)$ is the cardinality of the set $A \cap \mathbb{N}_0$. ◁

1.4.5 Product Measure

In section 1.4.2 we considered the Lebesgue measure on $(\mathbb{R}^n, \mathscr{B}_n)$ that is specified for n-dimensional cuboids by Equation (1.45) using the *product* of one-dimensional Lebesgue measures on $(\mathbb{R}, \mathscr{B})$. Now we introduce the general concept of a product measure. The following lemma shows that σ-finiteness of measures is sufficient for the existence and uniqueness of such a measure. Hence, this lemma shows that presuming finite measures is sufficient but not necessary for the definition of the product measure.

Lemma 1.65 (Existence and Uniqueness)
Let $(\Omega_i, \mathscr{A}_i, \mu_i)$ be measure spaces with σ-finite measures μ_i, $i = 1, \ldots, n$. Then there is a uniquely defined measure, denoted $\mu_1 \otimes \ldots \otimes \mu_n$, on the product space

$$\Big(\bigtimes_{i=1}^{n} \Omega_i, \bigotimes_{i=1}^{n} \mathscr{A}_i \Big),$$

satisfying

$$\begin{aligned} &\forall\, (A_1, \ldots, A_n) \in \mathscr{A}_1 \times \ldots \times \mathscr{A}_n\colon \\ &\mu_1 \otimes \ldots \otimes \mu_n (A_1 \times \ldots \times A_n) \;=\; \mu_1(A_1) \cdot \ldots \cdot \mu_n(A_n). \end{aligned} \tag{1.49}$$

This measure is σ-finite as well.

For a proof see Bauer (2001, Th. 23.9, p. 143). Hence, $\mu := \mu_1 \otimes \ldots \otimes \mu_n$ is a measure on the product space $\big(\bigtimes_{i=1}^{n} \Omega_i, \bigotimes_{i=1}^{n} \mathscr{A}_i\big)$ with

$$\mu(A_1 \times \ldots \times A_n) \;:=\; \mu_1(A_1) \cdot \ldots \cdot \mu_n(A_n), \quad \forall\, (A_1, \ldots, A_n) \in (\mathscr{A}_1 \times \ldots \times \mathscr{A}_n). \tag{1.50}$$

Definition 1.66 (Product Measure)
The measure $\mu_1 \otimes \ldots \otimes \mu_n$ defined by Equation (1.49) *is called the* product measure *of $\mu_1, \ldots, \mu_n$.*

1.5 Continuity of a Measure

The term σ-additivity refers to *countable* unions of pairwise disjoint sets and it implies finite additivity, which involves *finite* unions of pairwise disjoint sets. Furthermore, σ-additivity implies the following continuity properties of a measure, which is essential for the definition of the integral (see ch. 3).

Theorem 1.67 (Continuity of a Measure)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and let $A_1, A_2, \ldots \in \mathscr{A}$.

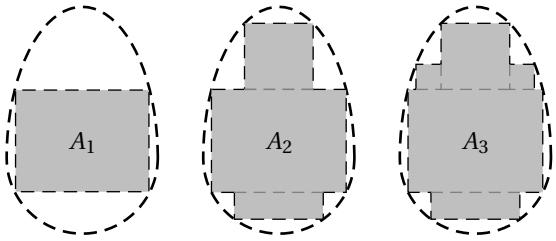


Figure 1.4. Approximation of an open egg-shaped set O from below

(i) If $A_1 \subset A_2 \subset \dots$, then

$$\lim_{i\to\infty} \mu(A_i) = \mu\Big(\bigcup_{i=1}^{\infty} A_i\Big). \qquad \textit{(continuity from below)}$$

(ii) If $A_1 \supset A_2 \supset \dots$ and there is an $n \in \mathbb{N}$ with $\mu(A_i) < \infty$, then

$$\lim_{i\to\infty} \mu(A_i) = \mu\Big(\bigcap_{i=1}^{\infty} A_i\Big). \qquad \textit{(continuity from above)}$$

For a proof see Klenke (2008, Theorem 1.36, p. 16 and 17).

Remark 1.68 (Finite Case) If $A_1, \dots, A_n \in \mathscr{A}$ is a finite sequence with $A_1 \subset \dots \subset A_n$, then $\bigcup_{i=1}^{n} A_i = A_n$ and

$$\mu\Big(\bigcup_{i=1}^{n} A_i\Big) = \mu(A_n). \tag{1.51}$$

This is a trivial case of Theorem 1.67 (with $A_n = A_{n+1} = A_{n+2} = \dots$). ◁

Example 1.69 (Geometric Examples) Figures 1.4 and 1.5 illustrate this theorem for the Lebesgue measure λ_2 on $(\mathbb{R}^2, \mathscr{B}_2)$, the *area* of a set O and the sets $A_i, i \in \mathbb{N}$. In this example, A_1 is the open rectangle in the open (i.e., the set without its boundary) egg-shaped set O displayed in Figure 1.4, A_2 the union of A_1 with two other rectangles in the middle figure, and A_3 the union of A_2 with two additional rectangles in the right figure. Adding more and more rectangles it is plausible that $A_1 \subset A_2 \subset \dots \subset O$ and that their union approximates O, i.e., $\bigcup_{i=1}^{\infty} A_i = O$. Under these premises Theorem 1.67 (i) yields the conclusion $\lim_{i\to\infty} \lambda_2(A_i) = \lambda_2\big(\bigcup_{i=1}^{\infty} A_i\big) = \lambda_2(O)$. Figure 1.5 illustrates the same principle. However, now the area of the egg-shaped set O is approximated from above by subtracting the areas of appropriate rectangles.

As a second example consider the Lebesgue measure λ on $(\mathbb{R}, \mathscr{B})$ and the intervals $A_i =]x - \frac{1}{i}, x]$, $i \in \mathbb{N}$. Obviously, $A_1 \supset A_2 \supset \dots$ and $\lambda(A_i) = \frac{1}{i} < \infty$, for all $i \in \mathbb{N}$ (see also Exercise 1-12). Hence, for all $x \in \mathbb{R}$,

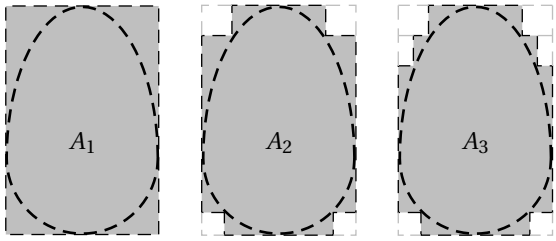


Figure 1.5. Approximation of an open egg-shaped set O from above

$$\lambda(\{x\}) = \lambda\Big(\bigcap_{i=1}^{\infty}]x-\frac{1}{i}, x]\Big) = \lim_{i\to\infty} \lambda\big(]x-\frac{1}{i}, x]\big) = \lim_{i\to\infty} \frac{1}{i} = 0. \tag{1.52}$$

This is an implication of continuity from above, and it implies

$$\begin{aligned}\forall a,b\in\mathbb{R}:\ a<b \ \Rightarrow\ \lambda\big(]a,b]\big) &= \lambda\big([a,b]\big) = \lambda\big([a,b[\big) = \lambda\big(]a,b[\big)\\ &= b-a.\end{aligned} \tag{1.53}$$

◁

Remark 1.70 (A Motivation for σ-Additivity) As already mentioned in Remark 1.45, σ-additivity refers to unions of finitely or countably many sets. Now consider $\bigcup_{1\le x\le 2}\{x\} = [1,2] \in \mathscr{B}$ [see Eq. (1.9)]. According to Equation (1.52), $\lambda\big(\{x\}\big) = 0$, for all $x\in[1,2]$, and hence $\lambda\big(\{x\in[1,2]\colon x\in\mathbb{Q}\}\big) = 0$, because the set of rational numbers is countable. In other words, the Lebesgue measure λ of the set of all rational number in the closed interval $[1,2]$ is zero, and this is not a contradiction to

$$\lambda\Big(\bigcup_{1\le x\le 2}\{x\}\Big) = \lambda\big([1,2]\big) = 2-1 = 1,$$

because $\bigcup_{1\le x\le 2}\{x\}$ is an uncountable union. This illustrates that additivity for uncountable unions can be meaningless. ◁

1.6 Specifying a Measure via a Generating System

Given a measurable space $(\Omega,\mathscr{A})$, a measure is a function that is defined on $\mathscr{A}$. In many situations, e. g., when $\mathscr{A} = \sigma(\mathscr{E})$ can only be described by a generating set system $\mathscr{E}$ (such as the set system $\mathscr{I}_1$ generating the Borel σ-algebra on $\mathbb{R}$), it is important to answer the following questions:

(a) *Existence*: If there is a set function $\tilde{\mu}\colon\mathscr{E}\to\overline{\mathbb{R}}$, is there also a measure $\mu\colon\sigma(\mathscr{E})\to\overline{\mathbb{R}}$ such that $\mu(A) = \tilde{\mu}(A)$, $\forall A\in\mathscr{E}$?

(b) *Uniqueness*: Is a measure μ on $\big(\Omega,\sigma(\mathscr{E})\big)$ already uniquely defined by its values $\mu(A)$, $A\in\mathscr{E}$?

(Sufficient conditions for the existence of such a measure μ are formulated in Theorem 1.53 of Klenke, 2008.)

The following uniqueness theorem for finite measures provides an answer to these questions, which suffices for our purposes. (A more general formulation for σ-finite measures with additional assumptions and a proof of Theorem 1.71 is found in Lemma 1.42 of Klenke, 2008.)

Theorem 1.71 (Generating System and Uniqueness of a Measure)
Let $(\Omega, \mathscr{A})$ be a measurable space and let $\mathscr{E} \subset \mathscr{A}$, where $\mathscr{E}$ is $\cap$-stable and $\sigma(\mathscr{E}) = \mathscr{A}$. If μ_1 and μ_2 are finite measures on $(\Omega, \mathscr{A})$, i.e., measures with $\mu_1(\Omega)$, $\mu_2(\Omega) < \infty$, then

$$\forall\, A \in \mathscr{E}\colon\ \mu_1(A) = \mu_2(A) \quad \Rightarrow \quad \forall\, A \in \mathscr{A}\colon\ \mu_1(A) = \mu_2(A).$$

Example 1.72 (Countable Ω) Let Ω be a finite or countable set and let $\mathscr{A} = \mathscr{P}(\Omega)$. Then the set system

$$\mathscr{E}_1 \;=\; \{\emptyset\} \cup \big\{\{\omega\}\colon \omega \in \Omega\big\}$$

is $\cap$-stable and $\sigma(\mathscr{E}_1) = \mathscr{A}$. As already noted in Remark 1.48, a finite measure μ on $(\Omega, \mathscr{A})$ is uniquely defined by its values $\mu(\{\omega\}), \omega \in \Omega$. ◁

Example 1.73 (Measures on $(\mathbb{R}, \mathscr{B})$) The set system

$$\mathscr{E}_2 \;=\; \big\{]a, b]\colon a < b,\ a, b \in \mathbb{R}\big\} \cup \{\emptyset\}$$

is $\cap$-stable and $\sigma(\mathscr{E}_2) = \mathscr{B}$ [see Eq. (1.18) and section 1.2.4]. Another $\cap$-stable set system $\mathscr{E}_3$ with $\sigma(\mathscr{E}_3) = \mathscr{B}$ is

$$\mathscr{E}_3 \;=\; \big\{]-\infty, b]\colon b \in \mathbb{R}\big\}$$

(cf. Klenke, 2008, p. 10). This set system is crucial for the definition of a cumulative distribution function (see section 5.7.1). ◁

1.7 σ-Algebra That is Trivial With Respect to a Measure

All σ-algebras treated in section 1.2 have been defined without reference to a measure. Now we define the concept of a *trivial σ-algebra*, which is defined referring to a measure. We start with a lemma about the set of all subsets of a set Ω with $\mu(A) = 0$ or $\mu(A) = \mu(\Omega)$, i. e., the set of all sets that are *trivial* with respect to the measure μ. Hence, the set of μ-trivial sets includes all *null sets*, i. e., all sets $A \subset \Omega$ with $\mu(A) = 0$, and all sets $A \subset \Omega$ with $\mu(A) = \mu(\Omega)$.

Lemma 1.74 (The Set of all Trivial Sets is a σ-Algebra)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and assume that μ is finite. Then

$$\mathscr{T}_\mu := \{A \in \mathscr{A} : \mu(A) = 0 \text{ or } \mu(A) = \mu(\Omega)\} \tag{1.54}$$

is a σ-algebra.

(Proof p. 30)

This lemma allows for the following definition:

Definition 1.75 (Trivial σ-Algebra With Respect to a Measure)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space, assume that μ is finite, and let $\mathscr{T}_\mu$ be defined by (1.54)*. Then each σ-algebra $\mathscr{C} \subset \mathscr{T}_\mu$ is called a μ-trivial σ-algebra and its elements μ-trivial sets.*

Obviously, $\{\Omega, \emptyset\}$ is a trivial σ-algebra with respect to all measures on $\{\Omega, \emptyset\}$. Hence, we can call it a *trivial σ-algebra*.

1.8 Proofs

Proof of Theorem 1.12

(a)

$$\begin{aligned} \forall\, i \in I\colon\ \mathscr{A}_i \text{ is a σ-algebra on } \Omega \ &\Rightarrow\ \forall\, i \in I\colon\ \Omega \in \mathscr{A}_i \qquad \text{[Def. 1.1 (a)]} \\ &\Rightarrow\ \Omega \in \bigcap_{i \in I} \mathscr{A}_i. \end{aligned}$$

(b)

$$\begin{aligned} A \in \bigcap_{i \in I} \mathscr{A}_i \ &\Rightarrow\ \forall\, i \in I\colon\ A \in \mathscr{A}_i \\ &\Rightarrow\ \forall\, i \in I\colon\ A^c \in \mathscr{A}_i \qquad \text{[Def. 1.1 (b)]} \\ &\Rightarrow\ A^c \in \bigcap_{i \in I} \mathscr{A}_i. \end{aligned}$$

(c)

$$\begin{aligned} A_1, A_2, \ldots \in \bigcap_{i \in I} \mathscr{A}_i \ &\Rightarrow\ \forall\, i \in I\colon\ A_1, A_2, \ldots \in \mathscr{A}_i \\ &\Rightarrow\ \forall\, i \in I\colon\ \bigcup_{j=1}^{\infty} A_j \in \mathscr{A}_i \qquad \text{[Def. 1.1 (c)]} \\ &\Rightarrow\ \bigcup_{j=1}^{\infty} A_j \in \bigcap_{i \in I} \mathscr{A}_i. \end{aligned}$$

Proof of Lemma 1.15

If $\mathscr{C}$ is a σ-algebra with $\mathscr{E} \subset \mathscr{C}$ and $\mathscr{A} = \sigma(\mathscr{E})$, then (1.11) and the assumption $\mathscr{C} \subset \mathscr{A}$ imply $\mathscr{A} = \sigma(\mathscr{E}) \subset \mathscr{C} \subset \mathscr{A}$. Hence, $\mathscr{C} = \mathscr{A}$.

Proof of Lemma 1.20

Define $\mathscr{D} := \{C = \bigcup_{i\in I(C)} B_i \colon I(C) \subset \mathbb{N}\}$.

$\mathscr{E} \subset \mathscr{D}$: For $B_j \in \mathscr{E}$ choose $I(B_j) = \{j\}$. Then $B_j = \bigcup_{i\in I(B_j)} B_i$.

$\mathscr{D} \subset \mathscr{E}$: Because $\mathbb{N}$ is countable, any $I(C) \subset \mathbb{N}$ is finite or countable, and this implies that $C = \bigcup_{i\in C} B_i$ is an element of $\sigma(\mathscr{E})$ [see Def. 1.1 (c), (1.3)].

Checking the three conditions defining a σ-algebra (see Def. 1.1), we show that $\mathscr{D}$ is a σ-algebra.

(a)

$$\Omega = \begin{cases} \bigcup_{i=1}^{n} B_i, & \text{if } \mathscr{E} = \{B_1, \dots, B_n\} \\ \bigcup_{i=1}^{\infty} B_i, & \text{if } \mathscr{E} = \{B_1, B_2, \dots\}, \end{cases}$$

because $\mathscr{E}$ is assumed to be a partition. This shows that $\Omega \in \mathscr{D}$.

(b) The equation for Ω in (a) also implies $I(C^c) = I(C)^c$. Therefore, $C^c \in \mathscr{D}$ if $C \in \mathscr{D}$.

(c) If $C_1, C_2, \dots \in \mathscr{D}$, then

$$\bigcup_{j=1}^{\infty} C_j = \bigcup_{j=1}^{\infty} \bigcup_{i\in I(C_j)} B_i = \bigcup_{i\in \bigcup_{j=1}^{\infty} I(C_j)} B_i \in \mathscr{D},$$

because $\bigcup_{j=1}^{\infty} I(C_j) \subset \mathbb{N}$.

Finally, we prove the second equation in (1.14). If $j \in I(C)$ and $C = \bigcup_{i\in I(C)} B_i$, then $B_j \subset C$, which implies

$$\bigcup_{i\in I(C)} B_i \subset \bigcup_{B_i\subset C} B_i.$$

Vice versa, if $B_j \subset C$, then $j \in I(C)$, because for any $\omega \in B_j$, there is no $i \neq j$ such that $\omega \in B_i$ [see condition (b) of Rem. 1.19]. Hence,

$$\bigcup_{B_i\subset C} B_i \subset \bigcup_{i\in I(C)} B_i,$$

which proves the second equation in (1.14).

Proof of Lemma 1.27

In this proof we use $\sigma_\Omega(\mathscr{E})$ to denote the σ-algebra on Ω generated by $\mathscr{E} \subset \mathscr{P}(\Omega)$. Similarly, $\sigma_A(\mathscr{D})$ denotes the σ-algebra on A generated by $\mathscr{D} \subset \mathscr{P}(A)$.

(1.15). $\sigma_\Omega(\mathscr{E})$ is a σ-algebra on Ω and $\mathscr{E} \subset \sigma_\Omega(\mathscr{E})$, by definition of $\sigma_\Omega(\mathscr{E})$. Hence, $\mathscr{E}\big|_A \subset \sigma_\Omega(\mathscr{E})\big|_A$, and $\sigma_\Omega(\mathscr{E})\big|_A$ is a σ-algebra on A (see Exercise 1-5). Therefore, the definition (1.10) yields

$$\sigma_A\big(\mathscr{E}\big|_A\big) \subset \sigma_\Omega(\mathscr{E})\big|_A.$$

Furthermore, $\mathscr{E} \subset \sigma_\Omega\big(\mathscr{E}\big|_A \cup \mathscr{E}\big|_{A^c}\big)$, which implies

$$\begin{aligned} \sigma_\Omega(\mathscr{E}) &\subset \sigma_\Omega\big(\mathscr{E}\big|_A \cup \mathscr{E}\big|_{A^c}\big) && \text{[Rem. 1.23]} \\ &\subset \sigma_\Omega\Big(\sigma_A\big(\mathscr{E}\big|_A\big) \cup \sigma_{A^c}\big(\mathscr{E}\big|_{A^c}\big)\Big) && \text{[Rem. 1.23]} \\ &= \Big\{C \cup D \colon C \in \sigma_A\big(\mathscr{E}\big|_A\big),\, D \in \sigma_{A^c}\big(\mathscr{E}\big|_{A^c}\big)\Big\}. && \text{[this set system is a σ-algebra]} \end{aligned}$$

Therefore,

$$\begin{aligned}
\sigma_\Omega(\mathscr{E})\big|_A &\subset \Big\{C\cup D\colon C\in\sigma_A\big(\mathscr{E}\big|_A\big),\, D\in\sigma_{A^c}\big(\mathscr{E}\big|_{A^c}\big)\Big\}\Big|_A \\
&= \Big\{(C\cup D)\cap A\colon C\in\sigma_A\big(\mathscr{E}\big|_A\big),\, D\in\sigma_{A^c}\big(\mathscr{E}\big|_{A^c}\big)\Big\} \\
&= \big\{C\cap A\colon C\in\sigma_A\big(\mathscr{E}\big|_A\big)\big\} && [D\subset A^c] \\
&= \sigma_A\big(\mathscr{E}\big|_A\big). && [C\subset A]
\end{aligned}$$

Hence, we have shown $\sigma_A\big(\mathscr{E}\big|_A\big)\subset\sigma_\Omega(\mathscr{E})\big|_A$ and $\sigma_\Omega(\mathscr{E})\big|_A\subset\sigma_A\big(\mathscr{E}\big|_A\big)$, which is equivalent to $\sigma_A\big(\mathscr{E}\big|_A\big)=\sigma_\Omega(\mathscr{E})\big|_A$.

(1.17).

$$\begin{aligned}
\sigma_\Omega(\mathscr{C}\cup\mathscr{E})\big|_A &= \sigma_A\big(\mathscr{C}\cup\mathscr{E}\big|_A\big) && [(1.15)] \\
&= \sigma_A\big(\mathscr{C}\big|_A\cup\mathscr{E}\big|_A\big) && [\text{see def. of the trace in Example 1.10}] \\
&= \sigma_A\big(\mathscr{C}\big|_A\cup\{\emptyset,A\}\big) && [(1.16)] \\
&= \sigma_A\big(\mathscr{C}\big|_A\big) && \big[\{\emptyset,A\}\subset\mathscr{C}\big|_A\big] \\
&= \mathscr{C}\big|_A. && [\text{Exercise 1-5, (1.12)}]
\end{aligned}$$

Proof of Lemma 1.74

(a) $\Omega\in\mathscr{T}_\mu$ by definition of $\mathscr{T}_\mu$.
(b) If $A\in\mathscr{T}_\mu$, then Rules (iv) and (v) of Box 1.1 and finiteness of μ yield

$$\mu(A^c)=\mu(\Omega)-\mu(A)=\begin{cases}\mu(\Omega), & \text{if } \mu(A)=0,\\ 0, & \text{if } \mu(A)=\mu(\Omega),\end{cases}$$

which implies $A^c\in\mathscr{T}_\mu$.

(c) Let $A_1,A_2,\ldots\in\mathscr{A}$. We consider two cases. *First*, if $\mu(A_i)=0$, for all $A_i, i\in\mathbb{N}$, then Rule (xi) of Box 1.1 yields $\left(\bigcup_{i=1}^\infty A_i\right)\le\sum_{i=1}^\infty\mu(A_i)=0$, i. e., $\bigcup_{i=1}^\infty A_i\in\mathscr{T}_\mu$. *Second*, if there is a $j\in\mathbb{N}$ such that $\mu(A_j)=\mu(\Omega)$, then Rule (v) of Box 1.1 yields

$$\mu(\Omega)=\mu(A_j)\le\mu\left(\bigcup_{i=1}^\infty A_i\right)\le\mu(\Omega),$$

which implies $\mu\left(\bigcup_{i=1}^\infty A_i\right)=\mu(\Omega)$. Therefore, $\bigcup_{i=1}^\infty A_i\in\mathscr{T}_\mu$.

1.9 Exercises

▷ **Exercise 1-1** Let $\mathscr{A}$ be a σ-algebra of subsets of a nonempty set Ω and let $A_1,A_2,\ldots\in\mathscr{A}$. Show: (a) $A_1\cap A_2\cap\ldots\in\mathscr{A}$, (b) $A_1\cap A_2\in\mathscr{A}$, and (c) $A_1\setminus A_2\in\mathscr{A}$.

▷ **Exercise 1-2** Show that the set system $\mathscr{A}=\{\Omega,\,\emptyset,A,A^c\}$ is stable (closed) with respect to union of elements of $\mathscr{A}$.

▷ **Exercise 1-3** Consider the set $\Omega=\{\omega_1,\ldots,\omega_6\}$ representing the set of all possible outcomes of tossing a dice and the power set $\mathscr{P}(\Omega)$, which, in probability theory, represents

the set of all possible events (including the 'impossible' event $\emptyset$) in this random experiment. Specify the σ-algebra on Ω that represents all possible events if we only distinguish between even and uneven number of points.

▷ **Exercise 1-4** Consider the random experiment that has been described in Example 1.9. Aside from the power set of Ω we already considered the σ-algebras $\mathscr{A}_1 = \{\Omega, \emptyset, A, A^c\}$, $\mathscr{A}_2 = \{\Omega, \emptyset, B, B^c\}$, and $\mathscr{A}_3 = \{\Omega, \emptyset, C, C^c\}$. Define another σ-algebra not yet mentioned.

▷ **Exercise 1-5** Prove: If $\mathscr{A}$ is a σ-algebra on Ω and $\Omega_0 \subset \Omega$, then $\mathscr{A}|_{\Omega_0} = \{\Omega_0 \cap A \colon A \in \mathscr{A}\}$ is a σ-algebra on Ω_0.

▷ **Exercise 1-6** Prove the proposition of Remark 1.16.

▷ **Exercise 1-7** Show that $\sigma(\mathscr{E}) = \mathscr{P}(\Omega)$ if Ω is finite or countable and $\mathscr{E} := \big\{\{\omega\} \colon \omega \in \Omega\big\}$.

▷ **Exercise 1-8** Prove the proposition of Remark 1.21.

▷ **Exercise 1-9** Let $\mathscr{E}_1, \mathscr{E}_2$ be set systems on Ω with $\mathscr{E}_1 \subset \mathscr{E}_2$. Show that $\sigma(\mathscr{E}_1) \subset \sigma(\mathscr{E}_2)$.

▷ **Exercise 1-10** Prove propositions (a) and (b) of Example 1.25.

▷ **Exercise 1-11** Prove Equation (1.21).

▷ **Exercise 1-12** Show that $\{x\} \in \mathscr{B}_n$ for all $x \in \mathbb{R}^n$, where $\mathscr{B}_n$ is the Borel σ-algebra on $\mathbb{R}^n$.

▷ **Exercise 1-13** Let $\mathscr{B}$ be the Borel σ-algebra on $\mathbb{R}$ and let $\Omega_0 \subset \mathbb{R}$ be finite or countable. Show that $\mathscr{B}|_{\Omega_0} = \mathscr{P}(\Omega_0)$.

▷ **Exercise 1-14** Prove the proposition of Example 1.34.

▷ **Exercise 1-15** Prove the proposition of Remark 1.35.

▷ **Exercise 1-16** Let $(\Omega_i, \mathscr{A}_i)$, $i = 1, \ldots, n$, be measurable spaces. Show that the set system $\mathscr{E} := \big\{A_1 \times \ldots \times A_n \colon A_i \in \mathscr{A}_i,\ i = 1, \ldots, n\big\}$ is $\cap$-stable.

▷ **Exercise 1-17** Prove the proposition of Remark 1.46.

▷ **Exercise 1-18** Prove the rules of Box 1.1.

▷ **Exercise 1-19** Prove the propositions of Remark 1.47.

▷ **Exercise 1-20** Show that $\delta_\omega \colon \mathscr{A} \to \{0, 1\}$ in Example 1.52 is a measure.

▷ **Exercise 1-21** Prove that the function defined by Equation (1.39) is a measure on $(\Omega, \mathscr{A})$.

▷ **Exercise 1-22** Show that $\sum_{\omega \in B} \delta_\omega$ in Example 1.57 is a measure.

▷ **Exercise 1-23** Show that $\nu \colon \mathscr{C} \to \overline{\mathbb{R}}$ defined in Example 1.60 is a measure on $(\Omega, \mathscr{C})$.

▷ **Exercise 1-24** Prove that the function $\sum_{i=1}^{\infty} \alpha_i \mu_i$ defined in Example 1.61 is a measure on $(\Omega, \mathscr{A})$.

Solutions

▷ **Solution 1-1** (a) If $A_1, A_2, \ldots \in \mathscr{A}$, then $A_1^c, A_2^c, \ldots \in \mathscr{A}$ [see Def. 1.1 (b)]. Hence,

$$\begin{aligned} \bigcap_{i=1}^{\infty} A_i \;&=\; \left[\left(\bigcap_{i=1}^{\infty} A_i\right)^c\right]^c \;=\; \left[\bigcup_{i=1}^{\infty} A_i^c\right]^c && \text{[de Morgan]} \\ &\in \mathscr{A}. && \text{[Def. 1.1 (c), (b)]} \end{aligned}$$

(b) Let $A_1, A_2 \in \mathscr{A}$ and choose $A_3, A_4, \ldots$ such that $\Omega = A_i$, for all $i \geq 3$, $i \in \mathbb{N}$. Then, according to Definition 1.1 (a),

$$A_1 \cap A_2 \;=\; A_1 \cap A_2 \cap \Omega \;=\; \bigcap_{i=1}^{\infty} A_i \;\in \mathscr{A}.$$

(c) $A_1 \setminus A_2 \;=\; A_1 \cap A_2^c \;\in \mathscr{A}$ [see (b) and Def. 1.1 (b)].

▷ **Solution 1-2** The unions $\Omega \cup A = \Omega$, $\Omega \cup A^c = \Omega$, and $\Omega \cup \emptyset = \Omega$ are all elements of $\mathscr{A}$ and the same is true for $\emptyset \cup A = A$, $\emptyset \cup A^c = A^c$, and $A \cup A^c = \Omega$. Furthermore, $B \cup B = B$ for all $B \in \mathscr{A}$.

▷ **Solution 1-3** The σ-algebra on Ω that only distinguishes between even and uneven number of points is $\mathscr{A}_1 := \big\{\{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}, \Omega, \emptyset\big\}$. This is a sub-σ-algebra of $\mathscr{P}(\Omega)$. Therefore, $\mathscr{A}_1$ represents the set of all possible events of a random experiment that is, in a sense, contained in the original random experiment.

▷ **Solution 1-4** Consider the set system that contains as elements $A, A^c, B, B^c, \Omega, \emptyset$, all unions and all intersections of these sets as well as the unions and intersections of the resulting sets such as $(A^c \cup B^c) \cap (A \cup B)$ and $(A^c \cup B^c) \cup (A \cup B)$. Altogether these are 16 sets. This is $\sigma(\mathscr{A}_1 \cup \mathscr{A}_2)$, the σ-algebra generated by $\mathscr{A}_1 \cup \mathscr{A}_2 = \{A, A^c, B, B^c, \Omega, \emptyset\}$ (see Def. 1.13 and Rem. 1.21).

▷ **Solution 1-5** (a) $\Omega_0 \cap \Omega = \Omega_0$. This implies $\Omega_0 \in \mathscr{A}\big|_{\Omega_0}$.

(b)

$$A^* \in \mathscr{A}\big|_{\Omega_0} \;\Rightarrow\; \exists\, A \in \mathscr{A} \colon A^* = \Omega_0 \cap A.$$

With this set A and using B^c for the complement of a set B with respect to Ω,

$$\begin{aligned} \Omega_0 \setminus A^* \;&=\; \Omega_0 \setminus (\Omega_0 \cap A) \\ &=\; \Omega_0 \cap (\Omega_0 \cap A)^c \\ &=\; \Omega_0 \cap (\Omega_0^c \cup A^c) \\ &=\; (\Omega_0 \cap \Omega_0^c) \cup (\Omega_0 \cap A^c) \\ &=\; \Omega_0 \cap A^c \in \mathscr{A}\big|_{\Omega_0}. \end{aligned}$$

(c)

$$A_1^*, A_2^*, \ldots \in \mathscr{A}\big|_{\Omega_0} \;\Rightarrow\; \exists\, A_1, A_2, \ldots \in \mathscr{A} \colon A_i^* = \Omega_0 \cap A_i, i \in \mathbb{N}.$$

Hence,

$$A_1^* \cup A_2^* \cup \ldots \;=\; (\Omega_0 \cap A_1) \cup (\Omega_0 \cap A_2) \cup \ldots \;=\; \Omega_0 \cap (A_1 \cup A_2 \cup \ldots) \in \mathscr{A}\big|_{\Omega_0}.$$

▷ **Solution 1-6** If $\mathscr{G}$ is a σ-algebra on Ω, then

$$\mathscr{E} \cup \mathscr{F} \subset \mathscr{G} \quad \Leftrightarrow \quad \sigma(\mathscr{E} \cup \mathscr{F}) \subset \mathscr{G}. \qquad [(1.11)] \tag{1.55}$$

Furthermore, for three sets A, B, C,

$$A \cup B \subset C \quad \Leftrightarrow \quad A \subset C \wedge B \subset C. \tag{1.56}$$

Hence,

$$\begin{aligned} \mathscr{D} \cup \mathscr{E} \cup \mathscr{F} \subset \mathscr{G} \quad &\Leftrightarrow \quad (\mathscr{D} \subset \mathscr{G}) \wedge \big(\mathscr{E} \cup \mathscr{F} \subset \mathscr{G}\big) && [(1.56)] \\ &\Leftrightarrow \quad (\mathscr{D} \subset \mathscr{G}) \wedge \big(\sigma(\mathscr{E} \cup \mathscr{F}) \subset \mathscr{G}\big) && [(1.55)] \\ &\Leftrightarrow \quad \mathscr{D} \cup \sigma(\mathscr{E} \cup \mathscr{F}) \subset \mathscr{G}. && [(1.56)] \end{aligned}$$

Now Definition 1.13 yields the proposition.

▷ **Solution 1-7** If Ω is finite or countable, then each of its subsets A is finite or countable as well. Therefore,

$$\forall A \subset \Omega\colon\ A = \bigcup_{\omega \in A} \{\omega\} \in \sigma(\mathscr{E}). \qquad \text{[Def. 1.1 (c), Rem. 1.2]}$$

Because each element A of $\mathscr{P}(\Omega)$ is a union $\bigcup_{\omega \in A}\{\omega\}$ of singletons $\{\omega\}$, $\omega \in A$, this implies $\mathscr{P}(\Omega) \subset \sigma(\mathscr{E})$. Hence, $\mathscr{E} \subset \mathscr{P}(\Omega) \subset \sigma(\mathscr{E})$. Therefore, Lemma 1.15 implies $\sigma(\mathscr{E}) = \mathscr{P}(\Omega)$.

▷ **Solution 1-8** Suppose that $\mathscr{E} = \{A_1, \ldots, A_m\}$ and $A_j^1 := A_j$ and let A_j^c denote the complement of A_j. Then, for all $(k_1, \ldots, k_m) \in \{1, c\}^m$ define

$$B_{(k_1,\ldots,k_m)} := \bigcap_{j=1}^{m} A_j^{k_j}.$$

Then

$$\mathscr{F} := \big\{B_{(k_1,\ldots,k_m)} \colon (k_1, \ldots, k_m) \in \{1, c\}^m, B_{(k_1,\ldots,k_m)} \neq \emptyset\big\}$$

is a finite partition of Ω. Note that $\mathscr{F}$ contains all nonempty intersections of sets A_j or their complements, respectively, where $j = 1, \ldots, m$. Now Lemma 1.20 implies the proposition.

▷ **Solution 1-9** If $\mathscr{E}_1 \subset \mathscr{E}_2 \subset \mathscr{P}(\Omega)$, then for any σ-algebra $\mathscr{A}$ on Ω with $\mathscr{E}_2 \subset \mathscr{A}$ also $\mathscr{E}_1 \subset \mathscr{A}$. Remember, if $J \subset I$, then $\bigcap_{i \in I} B_i \subset \bigcap_{i \in J} B_i$, for any sets $B_i, i \in I$. Therefore, $\sigma(\mathscr{E}_1)$, which is the intersection of all σ-algebras containing $\mathscr{E}_1$, is a subset of the intersection of all σ-algebras containing $\mathscr{E}_2$, which is $\sigma(\mathscr{E}_2)$.

▷ **Solution 1-10** (a) If Ω is finite, then $\mathscr{P}(\Omega)$ is a finite set system. Therefore, each σ-algebra $\mathscr{A}$ on Ω is a finite set system. Because $\mathscr{A} = \sigma(\mathscr{A})$, this σ-algebra is countably generated.

(b) The set $\mathbb{N}_0$ is countable and therefore also $\mathbb{N}_0^n$ for $n \in \mathbb{N}$. Example 1.18 then implies that $\mathscr{P}(\mathbb{N}_0^n)$ is countably generated.

▷ **Solution 1-11** Let $\mathscr{H}_n = \big\{]-\infty, b_1] \times \ldots \times]-\infty, b_n] \colon b_1, \ldots, b_n \in \mathbb{R}\big\}$.

(i) For all $(b_1, \ldots, b_n) \in \mathbb{R}^n$ and all $m \in \mathbb{N}$ with $m < b_i$, $i = 1, \ldots, n$,

$$B_m :=]-m, b_1] \times \ldots \times]-m, b_n] \in \mathscr{I}_n.$$

According to Definition 1.1 (c) this implies

$$\bigcup_{\substack{m\in\mathbb{N}\\ m<b_i,\,i=1,\dots,n}} B_m \;=\;]-\infty,b_1]\times\ldots\times]-\infty,b_n] \;\in\; \sigma(\mathscr{I}_n).$$

Hence, $\mathscr{H}_n \subset \sigma(\mathscr{I}_n)$, which, according to (1.11) and (1.12), implies

$$\sigma(\mathscr{H}_n) \;\subset\; \sigma(\mathscr{I}_n) = \mathscr{B}_n.$$

(ii) For all $a_1,\dots,a_n,\, b_1,\dots,b_n \in \mathbb{R}$, with $a_i < b_i,\ i=1,\dots,n$,

$$]a_1,b_1]\times\ldots\times]a_n,b_n] \;=\;]-\infty,b_1]\times\ldots\times]-\infty,b_n] \;\setminus\; \left(\bigcup_{j=1}^{n} H_j\right),$$

where $H_j :=]-\infty,b_1]\times\ldots\times]-\infty,b_{j-1}]\times]-\infty,a_j]\times]-\infty,b_{j+1}]\times]-\infty,b_n]$. Hence, according to Remark 1.2, $]a_1,b_1]\times\ldots\times]a_n,b_n] \in \sigma(\mathscr{H}_n)$ and $\mathscr{I}_n \subset \sigma(\mathscr{H}_n)$, which, according to (1.11) and (1.12), implies

$$\mathscr{B}_n \;=\; \sigma(\mathscr{I}_n) \;\subset\; \sigma(\mathscr{H}_n).$$

▷ **Solution 1-12** If $x\in\mathbb{R}$, then $\{x\} = \bigcap_{i=1}^{\infty}]x-1/i,\ x]$. According to Equation (1.18), the intervals $]x-1/i,\ x]$ are elements of the generating set system of $\mathscr{B}$, the Borel σ-algebra on $\mathbb{R}$. Therefore, their countable intersection is an element of $\mathscr{B}$. If $x=(x_1,\dots,x_n)\in\mathbb{R}^n$, then

$$\{x\} \;=\; \bigcap_{i=1}^{\infty}\left(\bigtimes_{j=1}^{n}]x_j-\frac{1}{i},x_j]\right).$$

According to Equation (1.20), the cuboids $\displaystyle\bigtimes_{j=1}^{n}]x_j-\frac{1}{i},x_j]$ are elements of the set system $\mathscr{I}_n$ and $\sigma(\mathscr{I}_n)=\mathscr{B}_n$.

▷ **Solution 1-13** Because $\{x\}\in\mathscr{B}$ for all $x\in\mathbb{R}$ (see Exercise 1-12) we can conclude: $\{x\}\in \mathscr{B}\big|_{\Omega_0}$ for all $x\in\Omega_0$. Hence, if Ω_0 is finite or countable, Example 1.18 implies $\mathscr{B}\big|_{\Omega_0} = \mathscr{P}(\Omega_0)$.

▷ **Solution 1-14** Let $\Omega_1,\dots,\Omega_n$ be finite or countable sets and let $\mathscr{A}_1,\dots,\mathscr{A}_n$ be their power sets. Then $\omega_1\in\Omega_1,\dots,\omega_n\in\Omega_n$ implies $\{\omega_1\}\in\mathscr{A}_1,\dots,\{\omega_n\}\in\mathscr{A}_n$. Therefore,

$$\{(\omega_1,\dots,\omega_n)\} \;=\; \{\omega_1\}\times\ldots\times\{\omega_n\} \;\in\; \left\{\bigtimes_{i=1}^{n} A_i\colon A_i\in\mathscr{A}_i, i\in\{1,\dots,n\}\right\}.$$

Hence,

$$\sigma\big(\{(\omega_1,\dots,\omega_n)\}\colon \omega_1\in\Omega_1,\dots,\omega_n\in\Omega_n\big) \;\subset\; \bigotimes_{i=1}^{n}\mathscr{A}_i.$$

With Ω_i being finite or countable, $\Omega=\Omega_1\times\ldots\times\Omega_n$ is finite or countable. Therefore,

$$\sigma\big(\{(\omega_1,\dots,\omega_n)\}\colon \omega_1\in\Omega_1,\dots,\omega_n\in\Omega_n\big) \;=\; \mathscr{P}(\Omega)$$

(see Example 1.18). Because $\bigotimes_{i=1}^{n}\mathscr{A}_i \subset \mathscr{P}(\Omega)$, we can conclude

$$\bigotimes_{i=1}^{n}\mathscr{A}_i \;=\; \mathscr{P}(\Omega_1\times\ldots\times\Omega_n) \;=\; \mathscr{P}\Big(\bigtimes_{i=1}^{n}\Omega_i\Big).$$

▷ **Solution 1-15**

$$\begin{aligned}(A\times B)^c &= \{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2\colon \omega_1\notin A \text{ or } \omega_2\notin B\}\\ &= \{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2\colon (\omega_1\notin A, \omega_2\in B) \text{ or } \omega_2\notin B\}\\ &= (A^c\times B)\cup(\Omega_1\times B^c)\end{aligned}$$

and

$$\begin{aligned}&(A^c\times B)\cap(\Omega_1\times B^c)\\ &= \{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2\colon \omega_1\notin A, \omega_2\in B, \omega_2\notin B\}\\ &= \{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2\colon \omega_1\notin A, \omega_2\in B\cap B^c=\emptyset\}\\ &= \emptyset.\end{aligned}$$

▷ **Solution 1-16** Remember that $(a\in A,\ b\in B)$ means $(a\in A \text{ and } b\in B)$ and that $(a\in A$ and $b\in B)$ and $(b\in B \text{ and } a\in A)$ are equivalent. Let $A_1,B_1\in\mathscr{A}_1,\dots,A_n,B_n\in\mathscr{A}_n$. Then $A_1\cap B_1\in\mathscr{A}_1,\dots,A_n\cap B_n\in\mathscr{A}_n$. Hence, $A_1\times\ldots\times A_n\in\mathscr{E}, B_1\times\ldots\times B_n\in\mathscr{E}$ and $(A_1\cap B_1)\times\ldots\times(A_n\cap B_n)\in\mathscr{E}$. Furthermore,

$$\begin{aligned}&(A_1\times\ldots\times A_n)\cap(B_1\times\ldots\times B_n)\\ &= \{(\omega_1,\dots,\omega_n)\colon \omega_1\in A_1,\dots,\omega_n\in A_n,\omega_1\in B_1,\dots,\omega_n\in B_n\}\\ &= \{(\omega_1,\dots,\omega_n)\colon \omega_1\in(A_1\cap B_1),\dots,\omega_n\in(A_n\cap B_n)\}\\ &= (A_1\cap B_1)\times\ldots\times(A_n\cap B_n)\ \in\ \mathscr{E}.\end{aligned}$$

▷ **Solution 1-17** Let B_i denote the sets defined in Remark 1.46.

(i) $B_1=A_1\in\mathscr{A}$. For all $i\in\mathbb{N}, i>1, B_i\in\mathscr{A}$:

$$B_i = A_i\setminus\left(\bigcup_{j=1}^{i-1}A_j\right) = A_i\cap\left(\bigcup_{j=1}^{i-1}A_j\right)^c\in\mathscr{A}. \qquad \text{[Def. 1.1 (b), Rem. 1.2]}$$

(ii) For any sequence $C_1,C_2,\ldots\subset\Omega$, define

$$\bigcup_{j=m}^{n}C_j := \emptyset, \quad \text{if } m>n, \qquad \text{and} \qquad \bigcap_{j=m}^{n}C_j := \Omega, \quad \text{if } m>n.$$

Then, using associativity and commutativity of the intersection, for $1\le k<l$,

$$\begin{aligned}B_k\cap B_l &= \left[A_k\setminus\left(\bigcup_{j=1}^{k-1}A_j\right)\right]\cap\left[A_l\setminus\left(\bigcup_{j=1}^{l-1}A_j\right)\right]\\ &= A_k\cap\left(\bigcup_{j=1}^{k-1}A_j\right)^c\cap A_l\cap\left(\bigcup_{j=1}^{l-1}A_j\right)^c && [A\setminus B=A\cap B^c]\\ &= A_k\cap\left(\bigcap_{j=1}^{k-1}A_j^c\right)\cap A_l\ \cap\ \left(\bigcap_{j=1}^{l-1}A_j^c\right) && \text{[de Morgan]}\\ &= A_k\cap A_l\cap\left(\bigcap_{j=1}^{k-1}A_j^c\right)\ \cap\ \left(\bigcap_{j=1}^{k-1}A_j^c\right)\cap A_k^c\cap\left(\bigcap_{j=k+1}^{l-1}A_j^c\right)\\ &= \emptyset. && [A_k\cap A_k^c=\emptyset]\end{aligned}$$

(iii) The sets B_i are defined such that $B_i\subset A_i$, for all $i\in I$. Therefore, $\bigcup_{i=1}^{\infty}B_i\subset\bigcup_{i=1}^{\infty}A_i$. Furthermore, for all $\omega\in\Omega$,

$$\begin{aligned}\omega \in \bigcup_{i=1}^{\infty} A_i \;&\Rightarrow\; \exists\, i \in \mathbb{N}\colon \omega \in A_i \wedge (\forall\, j < i\colon \omega \notin A_j)\\ &\Rightarrow\; \exists\, i \in \mathbb{N}\colon \omega \in A_1^c \cap \ldots \cap A_{i-1}^c \cap A_i = B_i\\ &\Rightarrow\; \omega \in \bigcup_{i=1}^{\infty} B_i.\end{aligned}$$

Hence, $\bigcup_{i=1}^{\infty} A_i \subset \bigcup_{i=1}^{\infty} B_i$, and this implies $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$.

▷ **Solution 1-18** (i) This is condition (c) of Definition 1.43.

(ii) If $A_1, \ldots, A_n \in \mathscr{A}$ are pairwise disjoint, then $A_1, A_2, \ldots$ with $\emptyset = A_{n+1} = A_{n+2} = \ldots$ is a sequence of pairwise disjoint measurable sets. Therefore, conditions (a) and (c) of Def. 1.43 imply

$$\mu\Big(\bigcup_{i=1}^{n} A_i\Big) = \mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{n} \mu(A_i) + \sum_{i=n+1}^{\infty} \mu(\emptyset) = \sum_{i=1}^{n} \mu(A_i).$$

(iii) For $A, B \subset \Omega$,

$$A = (A \cap B) \cup (A \cap B^c) = (A \cap B) \cup (A \setminus B)$$

and

$$(A \cap B) \cap (A \cap B^c) = A \cap B \cap B^c = \emptyset.$$

Hence, for sets $A, B \in \mathscr{A}$, Rule (ii) (finite additivity of μ) implies proposition (iii).

(iv) This proposition is a special case of (iii) with $A = \Omega$.

(v) Exchanging the roles of A and B in (iii) we obtain

$$\mu(B) = \mu(A \cap B) + \mu(B \setminus A).$$

If $A \subset B$, then $A \cap B = A$ and, because $\mu(B \setminus A) \geq 0$,

$$\mu(A) = \mu(A \cap B) \leq \mu(A \cap B) + \mu(B \setminus A) = \mu(B).$$

(vi) This rule immediately follows from proposition (iv) for $\mu(A \cap B) < \infty$. [Note that $\mu(A) - \mu(A \cap B)$ is not defined if $\mu(A) = \mu(A \cap B) = \infty$.]

(vii) For $A, B \subset \Omega$,

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A).$$

Because the right-hand side is a union of pairwise disjoint sets, finite additivity of μ yields

$$\begin{aligned}\mu(A \cup B) + \mu(A \cap B) &= \mu(A \setminus B) + \mu(A \cap B) + \mu(B \setminus A) + \mu(A \cap B)\\ &= \mu(A) + \mu(B). \qquad \text{[Box 1.1 (iii)]}\end{aligned}$$

(viii) $\mu(\Omega) = \mu(A \cup A^c) = \mu(A) + \mu(A^c)$. Hence, if $\mu(\Omega) = \mu(A) < \infty$, then $\mu(A^c) = 0$. Therefore, for all $B \in \mathscr{A}$, (v) implies $\mu(A^c \cap B) = 0$. Furthermore, $B = (A \cap B) \cup (A^c \cap B)$ and $(A \cap B) \cap (A^c \cap B) = \emptyset$. Hence, $\mu(B) = \mu(A \cap B) + \mu(A^c \cap B) = \mu(A \cap B)$. Note that, in general, $\mu(A) = \mu(\Omega)$ does not imply $A = \Omega$.

(ix) $\mu(A) = 0$ implies

$$\begin{aligned}\mu(B) &= \mu(A) + \mu(B)\\ &\geq \mu(A \cup B) \qquad \text{[(xi)]}\\ &\geq \mu(B). \qquad \text{[(v)]}\end{aligned}$$

Note that, in general, $\mu(A) = 0$ does not imply $A = \emptyset$.

(x) Let $B := \Omega \setminus \Omega_0$. Then $\mu(B) = 0$ as well as $\mu(A \cap B) = 0$ for all $A \in \mathscr{A}$ [see Box Box 1.1 (v)]. Furthermore, for $A \in \mathscr{A}$: $A = (A \cap \Omega_0) \cup (A \cap B)$, where $A \cap \Omega_0$ and $A \cap B$ are disjoint. Now, the sets $A \cap \Omega_0$, $A \in \mathscr{A}$, are the elements of the trace σ-algebra and $(\Omega_0, \mathscr{A}\big|_{\Omega_0}) = [\Omega_0, \mathscr{P}(\Omega_0)]$. Therefore we can apply Equation (1.30). Hence, for all $A \in \mathscr{A}$,

$$\begin{aligned} \mu(A) &= \mu(A \cap \Omega_0) + \mu(A \cap B) && \text{[Box 1.1 (ii)]} \\ &= \sum_{\omega \in A \cap \Omega_0} \mu(\{\omega\}) + \mu(A \cap B) && \text{[(1.30)]} \\ &= \sum_{\omega \in A \cap \Omega_0} \mu(\{\omega\}). && \text{[}\mu(A \cap B) = 0\text{]} \end{aligned}$$

(xi) Let $A_1, A_2, \ldots \in \mathscr{A}$ and define $B_1, B_2, \ldots \in \mathscr{A}$ by $B_1 = A_1$, and $B_i = A_i \setminus \bigcup_{j=1}^{i-1} B_j$ for $i > 1$ (see Rem. 1.46). Then $B_1, B_2, \ldots$ is a sequence of pairwise disjoint sets with $B_i \subset A_i$ for all $i \in \mathbb{N}$ and $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. Hence,

$$\begin{aligned} \mu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(\bigcup_{i=1}^{\infty} B_i\right) \\ &= \sum_{i=1}^{\infty} \mu(B_i) && \text{[Def. 1.43 (c)]} \\ &\leq \sum_{i=1}^{\infty} \mu(A_i). && \text{[Box 1.1 (v)]} \end{aligned}$$

▷ **Solution 1-19** If the $A_1, \ldots, A_n \in \mathscr{A}$ are pairwise disjoint and $B \in \mathscr{A}$, then, for $i \neq j$, $i, j = 1, \ldots, n$,

$$(B \cap A_i) \cap (B \cap A_j) = B \cap (A_i \cap A_j) = B \cap \emptyset = \emptyset.$$

Hence, the sets $B \cap A_1, \ldots, B \cap A_n$ are pairwise disjoint. Furthermore, condition (b) of Remark 1.47 implies

$$\bigcup_{i=1}^{n} (B \cap A_i) = B \cap \bigcup_{i=1}^{n} A_i = B.$$

Therefore, additivity of μ yields

$$\mu(B) = \mu\left(\bigcup_{i=1}^{n} (B \cap A_i)\right) = \sum_{i=1}^{n} \mu(B \cap A_i),$$

which is Equation (1.28) The proof of Equation (1.29) is literally the same except for replacing $\bigcup_{i=1}^{n}$ by $\bigcup_{i=1}^{\infty}$, $\sum_{i=1}^{n}$ by $\sum_{i=1}^{\infty}$, and additivity of μ by σ-additivity.

▷ **Solution 1-20** Let $\omega \in \Omega$.

(a) According to Equation (1.31), $\delta_\omega(\emptyset) = 1_\emptyset(\omega) = 0$.

(b) According to Equation (1.31), $\delta_\omega(A) = 1_A(\omega) \in \{0, 1\}$, for all $A \in \mathscr{A}$, and this implies $\delta_\omega(A) \geq 0$, for all $A \in \mathscr{A}$.

(c) If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint, then

$$\begin{aligned} \delta_\omega\Big(\bigcup_{i=1}^{\infty} A_i\Big) &= 1_{\bigcup_{i=1}^{\infty} A_i}(\omega) && \text{[(1.31)]} \\ &= \sum_{i=1}^{\infty} 1_{A_i}(\omega) && \text{[(1.36)]} \\ &= \sum_{i=1}^{\infty} \delta_\omega(A_i). && \text{[(1.31)]} \end{aligned}$$

▷ **Solution 1-21** (a) According to Equation (1.39), $\mu_\#(\emptyset) = \sum_{\omega\in\Omega} 1_\emptyset(\omega) = 0$.

(b) According to Equation (1.39), $\mu_\#(A) = \sum_{\omega\in\Omega} 1_A(\omega)$, for all finite $A \in \mathscr{A}$, and $\mu_\#(A) = \infty$, if A is infinite. This implies $\mu_\#(A) \geq 0$, for all $A \in \mathscr{A}$.

(c) If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint and all A_i are finite, then

$$\begin{aligned}
\mu_\#\Big(\bigcup_{i=1}^{\infty} A_i\Big) &= \sum_{\omega\in\Omega} 1_{\bigcup_{i=1}^{\infty} A_i}(\omega) && \text{[(1.39)]}\\
&= \sum_{\omega\in\Omega} \sum_{i=1}^{\infty} 1_{A_i}(\omega) && \text{[(1.36)]}\\
&= \sum_{i=1}^{\infty} \sum_{\omega\in\Omega} 1_{A_i}(\omega)\\
&= \sum_{i=1}^{\infty} \mu_\#(A_i). && \text{[(1.39)]}
\end{aligned}$$

Note that the set $\bigcup_{i=1}^{\infty} A_i$ can be countably infinite even if all A_i are finite. In this case $\mu_\#\big(\bigcup_{i=1}^{\infty} A_i\big) = \infty = \sum_{i=1}^{\infty} \mu_\#(A_i)$. If at least one of the A_i is infinite, then $\bigcup_{j=1}^{\infty} A_j \supset A_i$ is an infinite set and $\mu_\#\big(\bigcup_{j=1}^{\infty} A_j\big) \geq \mu_\#(A_i)$ is infinite as well.

▷ **Solution 1-22** (a) Using Equations (1.41) and (1.38),

$$\Big(\sum_{\omega\in B} \delta_\omega\Big)(\emptyset) = \sum_{\omega\in B} \delta_\omega(\emptyset) = \sum_{\omega\in B} 1_\emptyset(\omega) = \sum_{\omega\in B} 0 = 0.$$

(b) Using Equations (1.41) and (1.38),

$$\forall\, A \in \mathscr{A}\colon\ \Big(\sum_{\omega\in B} \delta_\omega\Big)(A) = \sum_{\omega\in B} \delta_\omega(A) = \sum_{\omega\in B} 1_A(\omega) \geq 0.$$

(c) If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint, then

$$\begin{aligned}
\Big(\sum_{\omega\in B} \delta_\omega\Big)\Big(\bigcup_{i=1}^{\infty} A_i\Big) &= \sum_{\omega\in B} \delta_\omega\Big(\bigcup_{i=1}^{\infty} A_i\Big) && \text{[(1.41)]}\\
&= \sum_{\omega\in B} 1_{\bigcup_{i=1}^{\infty} A_i}(\omega) && \text{[(1.38)]}\\
&= \sum_{\omega\in B} \sum_{i=1}^{\infty} 1_{A_i}(\omega) && \text{[(1.36)]}\\
&= \sum_{\omega\in B} \sum_{i=1}^{\infty} \delta_\omega(A_i) && \text{[(1.38)]}\\
&= \sum_{i=1}^{\infty} \left(\Big(\sum_{\omega\in B} \delta_\omega\Big)(A_i)\right). && \text{[(1.41)]}
\end{aligned}$$

▷ **Solution 1-23** (a) Equation (1.46) yields: $\nu(\emptyset) = \mu(\emptyset) = 0$.

(b) Equation (1.46) also yields: $\nu(A) = \mu(A) \geq 0$, for all $A \in \mathscr{C}$.

(c) If $A_1, A_2, \ldots \in \mathscr{C}$ are pairwise disjoint, then

$$\begin{aligned}
\nu\Big(\bigcup_{i=1}^{\infty} A_i\Big) &= \mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) && \text{[Def. 1.1 (c), (1.46)]}\\
&= \sum_{i=1}^{\infty} \mu(A_i) && \text{[Def. 1.43 (c)]}\\
&= \sum_{i=1}^{\infty} \nu(A_i). && \text{[(1.46)]}
\end{aligned}$$

▷ **Solution 1-24** (a) Using Equation (1.47) and Definition Def. 1.43 (a) yields

$$\Big(\sum_{i=1}^{\infty} \alpha_i \mu_i\Big)(\emptyset) = \sum_{i=1}^{\infty} \alpha_i \mu_i(\emptyset) = \sum_{i=1}^{\infty} 0 = 0.$$

(b) Similarly, using Equation (1.47) yields, for all $A \in \mathscr{A}$,

$$\Big(\sum_{i=1}^{\infty} \alpha_i \mu_i\Big)(A) = \sum_{i=1}^{\infty} \alpha_i \mu_i(A) = \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i \mu_i(A) \geq 0,$$

because $\mu_i(A) \geq 0$ and we assume $\alpha_i \geq 0$.

(c) If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint, then

$$\begin{aligned}
\Big(\sum_{i=1}^{\infty} \alpha_i \mu_i\Big)\Big(\bigcup_{j=1}^{\infty} A_j\Big) &= \sum_{i=1}^{\infty} \alpha_i \mu_i\Big(\bigcup_{j=1}^{\infty} A_j\Big) && [(1.47)] \\
&= \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{\infty} \mu_i(A_j) && [\text{Def. 1.43 (c)}] \\
&= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \alpha_i \mu_i(A_j) \\
&= \sum_{j=1}^{\infty} \Big(\Big(\sum_{i=1}^{\infty} \alpha_i \mu_i\Big)(A_j)\Big). && [(1.47)]
\end{aligned}$$

Note that the last but one equation holds, because rearranging summands does not change the sum if the terms α_i and $\mu_i(A_j)$ are nonnegative.

Chapter 2
Measurable Mapping

In chapter 1 we treated the concepts of a σ-*algebra* and a σ-*algebra generated by a set system* on a set Ω. An element A of a σ-algebra $\mathscr{A}$ has been called a *measurable set*. We also introduced the concept of a *measure*, which assigns a nonnegative real number or ∞ to all elements of a σ-algebra. This chapter is devoted to the concept of a *measurable mapping*, related concepts such as the σ-*algebra generated by a mapping*, and the *image measure* of μ under f, the measure induced by a measurable mapping f on its codomain space. All these concepts play an important role in integration and probability theory. In probability theory, a measurable set is called an event, a measurable mapping f is called a *random variable* and the image measure of the probability measure P under f is called the *distribution of* f.

2.1 Image and Inverse Image

Two key concepts of this chapter are the *image* of a set $A \subset \Omega$ and the *inverse image* of a set $A' \subset \Omega'$ under a mapping $f: \Omega \to \Omega'$. We start with the formal definitions and then illustrate these concepts in section 2.2.

Definition 2.1 (Image and Inverse Image)
Let Ω, Ω' denote two sets and $f: \Omega \to \Omega'$ a mapping. Then we call

$$f(A) := \{f(\omega): \omega \in A\}, \quad A \subset \Omega, \tag{2.1}$$

the image of A under f, and

$$f^{-1}(A') := \{\omega \in \Omega: f(\omega) \in A'\}, \quad A' \subset \Omega', \tag{2.2}$$

the inverse image of A' under f.

Whereas the image $f(A)$ is a subset of Ω', the inverse image $f^{-1}(A')$ is the set of all elements of the *domain* Ω for which f takes on a value in the subset A' of its *codomain* Ω'. For convenience, we also use the notation

$$\{f \in A'\} := f^{-1}(A') \qquad \text{and} \qquad \{f = \omega'\} := f^{-1}(\{\omega'\}). \tag{2.3}$$

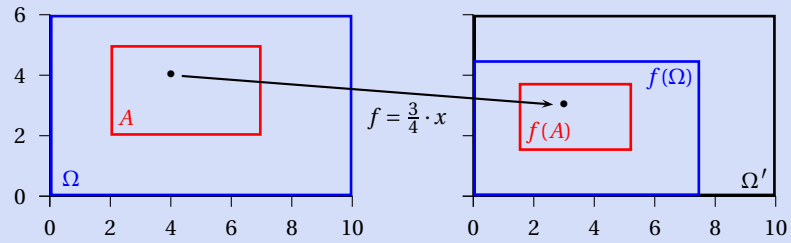


Figure 2.1. Rectangles and their images under a function

Remark 2.2 (Properties of Inverse Images) Let $f: \Omega \to \Omega'$ be a mapping, I be an index set, $A' \subset \Omega'$, and $(A'_i,\ i \in I)$ a family of subsets A'_i of Ω'. Then

$$f^{-1}[(A')^c] = [f^{-1}(A')]^c, \tag{2.4}$$

$$f^{-1}\Big(\bigcap_{i\in I} A'_i\Big) = \bigcap_{i\in I} f^{-1}(A'_i), \tag{2.5}$$

$$f^{-1}\Big(\bigcup_{i\in I} A'_i\Big) = \bigcup_{i\in I} f^{-1}(A'_i) \tag{2.6}$$

(see Exercise 2-1). Note that, in general, the corresponding properties do not necessarily hold for the image $f(A),\ A \subset \Omega$. ◁

2.2 Introductory Examples

2.2.1 Example 1: Rectangles

Our first example deals with rectangles, their *images*, and their *inverse images* under a mapping f.

The Measurable Space

Let $[a,b]$, $a,b \in \mathbb{R}$, denote the closed in interval between a and b, inclusively, and consider the two rectangles

$$\Omega = [0,10] \times [0,6] \qquad \text{and} \qquad A = [2,7] \times [2,5]$$

depicted on the left-hand side of Figure 2.1. The elements of Ω and A are points $x = (x_1, x_2)$ in these rectangles with coordinates x_1 on the horizontal axis and x_2 on the vertical axis. Furthermore, let us consider a σ-algebra on Ω,

$$\mathscr{A} = \{\Omega, \emptyset, A, A^c\}.$$

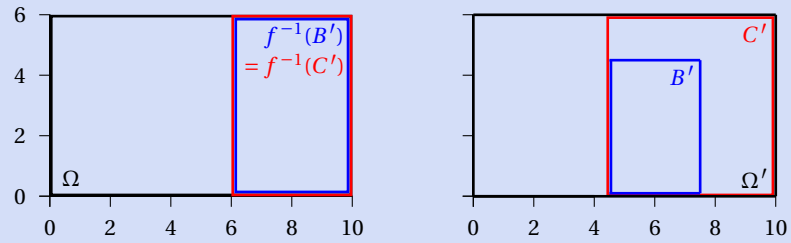


Figure 2.2. Rectangles and their inverse images under a function

The Mapping and the Image

Consider the set $\Omega' = \Omega$ and the function $f\colon \Omega \to \Omega'$ defined by

$$f(x) = \frac{3}{4} \cdot x = \left(\frac{3}{4} \cdot x_1, \frac{3}{4} \cdot x_2 \right), \quad \forall x \in \Omega. \tag{2.7}$$

Hence, f maps all points $x = (x_1, x_2) \in \Omega$ to the points $f(x_1, x_2) \in \Omega'$. This is illustrated by Figure 2.1 for the point $x = (4,4)$, which is mapped to $f(x) = (3,3)$. The right-hand side of Figure 2.1 also depicts the *image of A under f*, i. e., $f(A) = \{f(x)\colon x \in A\}$, as well as the image $f(\Omega)$ of Ω under f.

The Inverse Images

We specify the σ-algebra

$$\mathscr{A}' = \{\Omega', \emptyset, B', (B')^c\}$$

on Ω', where

$$B' =]4.5, 7.5] \times [0, 4.5]$$

is the rectangle depicted on the right-hand side of Figure 2.2, and $(B')^c = \Omega' \setminus B'$ is its complement.

Now we consider the *inverse image* of B' under f [see Eq. (2.7)], i. e.,

$$f^{-1}(B') =]6, 10] \times [0, 6]$$

(see Fig. 2.2). It is the rectangle on the right side of Ω. For further examples see Exercises 2-2 and 2-3.

Also consider the inverse image of the rectangle

$$C' =]4.5, 10] \times [0, 6]$$

(see Fig. 2.2). Its inverse image under f is identical to the inverse image $f^{-1}(B')$, i. e.,

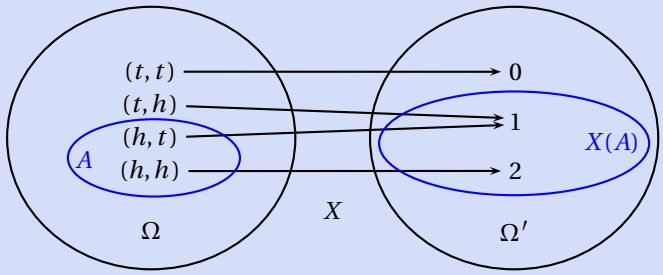


Figure 2.3. A set and its image under a function

$$f^{-1}(C') = f^{-1}(B'),$$

which follows from

$$\begin{aligned} f^{-1}(C') &= f^{-1}[B' \cup (C' \setminus B')] && [B' \subset C', \text{ Fig. 2.2}] \\ &= f^{-1}(B') \cup f^{-1}(C' \setminus B') && [(2.6)] \\ &= f^{-1}(B') \cup \emptyset = f^{-1}(B'). \end{aligned}$$

Note that $f^{-1}(C' \setminus B') = \emptyset$, because f has been defined on $\Omega = [0,10] \times [0,6]$. If we would define f on $\Omega = \mathbb{R}^2$, then the set $f^{-1}(C' \setminus B')$ would *not* be empty. (See also Exercise 2-4.)

2.2.2 Example 2: Flipping two Coins

Now we consider the random experiment of *flipping two coins*.

The Measurable Space

In this random experiment, the set of possible outcomes is

$$\Omega = \{(h,h), (h,t), (t,h), (t,t)\}.$$

This set consists of four elements (pairs). For example, the first component of the pair (h,t) represents the outcome of flipping $h = \textit{heads}$ with the first coin and the second component represents the outcome of flipping $t = \textit{tails}$ with the second coin. As a σ-algebra on Ω we consider the power set $\mathscr{A} = \mathscr{P}(\Omega)$.

The Mapping

Consider the function $X\colon \Omega \to \Omega' = \{0,1,2\}$ defined by

$$X[(t,t)] = 0, \qquad X[(t,h)] = 1, \qquad X[(h,t)] = 1, \qquad \text{and} \qquad X[(h,h)] = 2.$$

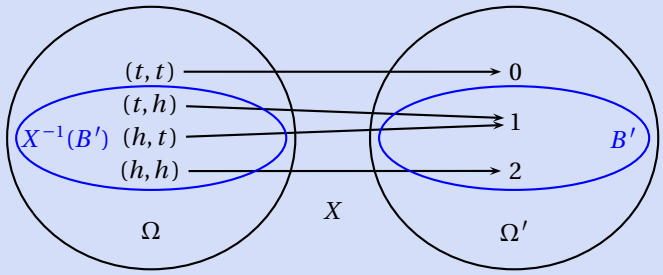


Figure 2.4. A set and its inverse image under a function

Table 2.1. Example of measurable sets represented by a mapping X

$\{X \in \Omega'\}$	$= X^{-1}(\Omega')$	$= \Omega$	0, 1, or 2 heads are flipped
$\{X \in \emptyset\}$	$= X^{-1}(\emptyset)$	$= \emptyset$	neither 0, 1, nor 2 *heads* are flipped
$\{X = 0\}$	$= X^{-1}(\{0\})$	$= \{(t,t)\}$	no *heads* are flipped.
$\{X = 1\}$	$= X^{-1}(\{1\})$	$= \{(h,t),(t,h)\}$	*heads* are flipped exactly once
$\{X = 2\}$	$= X^{-1}(\{2\})$	$= \{(h,h)\}$	two heads are flipped
$\{X \in \{0,1\}\}$	$= X^{-1}(\{0,1\})$	$= \{(h,t),(t,h),(t,t)$	not more than one *heads* are flipped.
$\{X \in \{0,2\}\}$	$= X^{-1}(\{0,2\})$	$= \{(h,h),(t,t)\}$	either two heads or no *heads* at all are flipped
$\{X \in \{1,2\}\}$	$= X^{-1}(\{1,2\})$	$= \{(h,h),(h,t),(t,h)\}$	at least one *heads* is flipped

Looking at this assignment rule shows that this function may be called *number of flipping heads*. Again, we consider the *image of a set* $A \subset \Omega$ *under* X, i. e., $X(A) = \{X(\omega)\colon \omega \in A\}$, $A \subset \Omega$. For example, for $A = \{(h,h),(h,t)\}$, the image under X is $X(A) = \{1,2\}$ (see Fig. 2.3).

The Inverse Images

Suppose $\mathscr{A}' = \mathscr{P}(\Omega')$ is the power set of $\Omega' = \{0,1,2\}$. In this example there are also $2^3 = 8$ *inverse images* $X^{-1}(A') = \{\omega \in \Omega\colon X(\omega) \in A'\}$, $A' \in \mathscr{A}'$. Three of these eight inverse images are:

$$X^{-1}(\{0\}) = \{(t,t)\}, \quad X^{-1}(\{1\}) = \{(h,t),(t,h)\}, \quad X^{-1}(\{2\}) = \{(h,h)\}.$$

These are the events that X takes on the value 0, 1, and 2, respectively. (In order to identify the inverse images listed above, trace back the arrows from right to left in Figure 2.4.) Furthermore, consider the inverse images

$$\begin{aligned} X^{-1}(\{0,1\}) &= \{(t,t),(h,t),(t,h)\},\\ X^{-1}(\{0,2\}) &= \{(t,t),(h,h)\},\\ X^{-1}(\{1,2\}) &= \{(h,t),(t,h),(h,h)\}. \end{aligned}$$

These are the events that X takes on a value in the sets $\{0,1\},\{0,2\}$, and $\{1,2\}$, respectively. One of these inverse images, namely $X^{-1}(B')$, with $B' := \{1,2\}$, is represented in Figure 2.4. Finally,

$$X^{-1}(\Omega') = \Omega \quad \text{and} \quad X^{-1}(\emptyset) = \emptyset.$$

Hence, we listed all eight inverse images $X^{-1}(A')$, $A' \in \mathscr{A}'$. They are the eight measurable sets that can be represented by the mapping X and the σ-algebra $\mathscr{A}' = \mathscr{P}(\Omega')$. These sets are listed in Table 2.1, using the notation $\{X \in A'\} := X^{-1}(A')$, $A' \in \mathscr{A}'$, and $\{X = x\} := X^{-1}(\{x\})$, $\{x\} \in \mathscr{A}'$ [see Eq. (2.3)].

2.3 Measurable Mapping

Now we define the concept of a *measurable mapping* and related concepts such as the σ*-algebra generated by a mapping* and *measurability of a mapping with respect to a mapping.*

Remark 2.3 (Mapping) Remember, a mapping $f\colon \Omega \to \Omega'$ assigns to *all* $\omega \in \Omega$ a unique $f(\omega) \in \Omega'$. Hence, f is, by definition, a subset of the Cartesian product $\Omega \times \Omega'$, i. e., $f = \big\{\big(\omega, f(\omega)\big)\colon \omega \in \Omega\big\}$. This implies that, instead of $f\colon \Omega \to \Omega'$, we can also write $f\colon \Omega \to \Omega''$ for the *same mapping*, provided that $f(\Omega) \subset \Omega''$. ◁

Remark 2.4 (Identical Mappings) If $f, g\colon \Omega \to \Omega'$ are two mappings, then

$$f = g \quad \Leftrightarrow \quad \big\{\big(\omega, f(\omega)\big)\colon \omega \in \Omega\big\} = \big\{\big(\omega, g(\omega)\big)\colon \omega \in \Omega\big\}. \tag{2.8}$$

If $f = g$ we say that the two mappings are *identical*. Hence, even if $f\colon \Omega \to \Omega'$ and $g\colon \Omega \to \Omega''$ are mappings with $\Omega' \neq \Omega''$, it is still possible that f and g are identical. Note that (2.8) also implies: If, for $f\colon \Omega \to \Omega'$ and $g\colon \Omega \to \Omega''$, we write $f, g\colon \Omega \to \Omega'''$ with $\Omega''' := \Omega' \cup \Omega''$, then f and g remain unchanged. ◁

2.3.1 Measurable Mapping

Now the core concept of this chapter is defined as follows:

Definition 2.5 (Measurable Mapping)
Let $(\Omega, \mathscr{A})$, $(\Omega', \mathscr{A}')$ *be measurable spaces and let* $f\colon \Omega \to \Omega'$ *be a mapping. Then* f *is called* $(\mathscr{A}, \mathscr{A}')$*-measurable if*

$$f^{-1}(A') \in \mathscr{A}, \quad \forall\, A' \in \mathscr{A}'.$$

Remark 2.6 (Notation) We use the notation

$$f\colon (\Omega, \mathscr{A}) \to (\Omega', \mathscr{A}')$$

to express that the mapping $f\colon \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable. If there is no ambiguity about $\mathscr{A}'$, we also say that f is *$\mathscr{A}$-measurable* or *measurable with respect to $\mathscr{A}$*. ◁

Examples

Example 2.7 (Rectangles – continued) In Example 2.2.1, we considered the mapping $f\colon \Omega \to \Omega' = \Omega$ defined by $f(x) = \frac{3}{4}x$. Furthermore, we considered the rectangle $B' =]4.5, 7.5] \times [0, 4.5]$ and the inverse image

$$f^{-1}(B') =]6, 10] \times [0, 6].$$

If $A = [2,7] \times [2,5]$, then the inverse image $f^{-1}(B')$ is not an element of the σ-algebra $\mathscr{A} = \{\Omega, \emptyset, A, A^c\}$. In this example, we also specified the σ-algebra $\mathscr{A}' = \{\Omega', \emptyset, B', (B')^c\}$. Hence, f is not $(\mathscr{A}, \mathscr{A}')$-measurable. However, if we specify a σ-algebra $\mathscr{C}$ such that $f^{-1}(B') \in \mathscr{C}$, then f *is* $(\mathscr{C}, \mathscr{A}')$-measurable. As we see later on, this condition is sufficient for f to be $(\mathscr{C}, \mathscr{A}')$-measurable (see Th. 2.20). ◁

Example 2.8 (Flipping two Coins – continued) In Example 2.2.2, we considered the mapping $X =$ *number of flipping heads* and in Table 2.1 (p. 45) we listed all inverse images $X^{-1}(A')$, $A' \in \mathscr{A}' = \mathscr{P}(\{0,1,2\})$. Of course, $\mathscr{A} = \mathscr{P}(\Omega)$ ensures that all inverse images $X^{-1}(A')$, $A' \in \mathscr{A}'$, are elements of $\mathscr{A}$.

However, instead of $\mathscr{A} = \mathscr{P}(\Omega)$, we might consider the σ-algebra

$$\mathscr{A}_0 = \big\{\Omega,\ \emptyset,\ \{(h,h),(h,t)\},\ \{(t,h),(t,t)\}\big\}.$$

The element $\{(h,h),(h,t)\}$ represents the event that *heads* are flipped in the first flip and $\{(t,h),(t,t)\}$ is the event that *tails* are flipped in the first flip. Hence, the σ-algebra $\mathscr{A}_0$ contains the events that refer to the outcome of the *first flip* only, whereas X represents the number of heads in *both coin flips*. If we choose $\mathscr{A}'$ to be the power set of $\Omega' = \{0,1,2\}$, then it is *not* true that all eight inverse images $X^{-1}(A')$, $A' \in \mathscr{A}'$, are elements of $\mathscr{A}_0$. The inverse image $X^{-1}(\{2\}) = \{(h,h)\}$, e. g., is not an element of $\mathscr{A}_0$. Hence, if we consider the measurable spaces $(\Omega, \mathscr{A}_0)$ and $\big(\Omega', \mathscr{P}(\Omega')\big)$, then the mapping X is *not* $\big(\mathscr{A}_0, \mathscr{P}(\Omega')\big)$-measurable. Hence, in some sense $\mathscr{A}_0$ is 'not well-adapted' to X. ◁

Example 2.9 (Two Trivial Cases) If (a) $\mathscr{A} = \mathscr{P}(\Omega)$ is the power set of Ω or if (b) $\mathscr{A}' = \{\Omega', \emptyset\}$, then every mapping $f\colon \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable. This is easily seen as follows: (a) If $\mathscr{A} = \mathscr{P}(\Omega)$ is the power set of Ω, then all inverse images $f^{-1}(A')$, $A' \subset \Omega'$, are elements in $\mathscr{A} = \mathscr{P}(\Omega)$, because it is the set of *all* subsets of Ω. (b) If $\mathscr{A}' = \{\Omega', \emptyset\}$, then every mapping $f\colon \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable, because $f^{-1}(\Omega') = \Omega$ and $f^{-1}(\emptyset) = \emptyset$. Again, the inverse images Ω and $\emptyset$ are both elements in every σ-algebra on Ω. Hence, in both cases, (a) and (b), every mapping $f\colon \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable. ◁

Example 2.10 (Constant Mapping) A constant mapping $f\colon \Omega \to \Omega'$ is defined by

$$f(\omega) = \omega', \quad \forall \omega \in \Omega,$$

where ω' is a fixed element of Ω'. Such a constant mapping is $(\mathscr{A}, \mathscr{A}')$-measurable for any σ-algebra $\mathscr{A}$ on Ω and any σ-algebra $\mathscr{A}'$ on Ω'. This is true, because for all subsets A' of Ω': If $\omega' \in A'$, then $f^{-1}(A') = \Omega$. If, in contrast, $\omega' \notin A'$, then $f^{-1}(A') = \emptyset$. However, Ω and $\emptyset$ are elements of *all* σ-algebras on Ω. ◁

Example 2.11 (Identity Mapping) The identity mapping $id\colon \Omega \to \Omega$ defined by

$$id(\omega) = \omega, \quad \forall \omega \in \Omega,$$

is $(\mathscr{A}, \mathscr{A}_0)$-measurable for any pair of σ-algebras on Ω with $\mathscr{A}_0 \subset \mathscr{A}$. This is easily seen as follows:

$$id^{-1}(A) = A, \quad \forall A \in \mathscr{A}_0.$$

Because we assume $\mathscr{A}_0 \subset \mathscr{A}$, we can conclude that id is $(\mathscr{A}, \mathscr{A}_0)$-measurable. ◁

Example 2.12 (Indicator of a Measurable Set) Let $(\Omega, \mathscr{A}), (\Omega', \mathscr{A}')$ be two measurable spaces, where $\mathscr{A}'$ is any σ-algebra on $\Omega' \subset \mathbb{R}$ with $\{0\}, \{1\} \in \mathscr{A}'$. Then the indicator $1_A\colon \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable *if and only if* $A \in \mathscr{A}$. Note that the requirement $\{0\}, \{1\} \in \mathscr{A}'$ is not only satisfied by $(\Omega', \mathscr{A}') := \big(\{0,1\}, \mathscr{P}(\{0,1\})\big)$, but also by $(\Omega', \mathscr{A}') = (\mathbb{R}, \mathscr{B})$ and by $(\Omega', \mathscr{A}') = (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, where $\mathscr{B}$ denotes the Borel σ-algebra on $\mathbb{R}$ and $\overline{\mathscr{B}}$ the Borel σ-algebra on $\overline{\mathbb{R}}$. ◁

Example 2.13 (Indicators of Unions and Intersections) If $(\Omega, \mathscr{A})$ is a measurable space and $A, B \in \mathscr{A}$, then $1_{A \cap B}$ and $1_{A \cup B}$ are $(\mathscr{A}, \mathscr{B})$-measurable. This follows from the fact that $A \cap B \in \mathscr{A}$ and $A \cup B \in \mathscr{A}$. For the same reason, $A_1, A_2, \ldots \in \mathscr{A}$ implies that $1_{\bigcup_{i=1}^{\infty} A_i}$ is $(\mathscr{A}, \mathscr{B})$-measurable. ◁

Example 2.14 (Constant Function) Assume that $(\Omega, \mathscr{A})$ and $(\Omega', \mathscr{A}')$ are measurable spaces such that $\{\omega'\} \in \mathscr{A}'$, for all $\omega' \in \Omega'$. Furthermore, let $f\colon \Omega \to \Omega'$. If $\mathscr{A} = \{\Omega, \emptyset\}$, then f is $(\mathscr{A}, \mathscr{A}')$-measurable if and only if f is a constant function, i. e., if and only if there is an $\omega' \in \Omega'$ such that $f(\omega) = \omega'$, for all $\omega \in \Omega$ (see Exercise 2-5). Note that for $(\Omega', \mathscr{A}') = (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, $\{x\} \in \overline{\mathscr{B}}$, for all $x \in \overline{\mathbb{R}}$. ◁

Example 2.15 (Dichotomous Function) If $\mathscr{A} = \{\Omega, \emptyset, A, A^c\}$ with $A \subset \Omega$, then $f\colon \Omega \to \mathbb{R}$ is $(\mathscr{A}, \mathscr{B})$-measurable if and only if $f = \alpha_1 1_A + \alpha_2 1_{A^c}$ for $\alpha_1, \alpha_2 \in \mathbb{R}$ (see Exercise 2-6). ◁

Step Function

Another important example of a measurable function is a step function, which is defined as follows:

Definition 2.16 (Step Function)
Let $A_1,\dots,A_n$, $n\in\mathbb{N}$, be a finite sequence of subsets of a set Ω. Then a finite linear combination

$$f = \sum_{i=1}^{n} \alpha_i 1_{A_i}, \quad \alpha_1,\dots,\alpha_n \in \mathbb{R}, \tag{2.9}$$

is called a step function.

Remark 2.17 (Step Function and a Partition of Ω) If the sets $A_1,\dots,A_n$ are pairwise disjoint, if we define $A_{n+1} := \Omega \setminus \left(\bigcup_{i=1}^{n} A_i\right)$, then $\{A_1,\dots,A_n,A_{n+1}\}$ is a finite partition of Ω. Furthermore, if f satisfies (2.9), and $\alpha_{n+1} := 0$, then, for all $A'\subset\mathbb{R}$,

$$f^{-1}(A') = \bigcup_{\substack{i=1,\dots,n+1,\\ \alpha_i\in A'}} A_i \tag{2.10}$$

(see Exercise 2-7). ◁

Remark 2.18 (Measurability of a Step Function) If $(\Omega,\mathscr{A})$ is a measurable space and $A_1,\dots,A_n\in\mathscr{A}$, then the step function $f:\Omega\to\mathbb{R}$ defined by Equation (2.9) is $(\mathscr{A},\mathscr{B})$-measurable (see Exercise 2-8). ◁

Lemma 2.19 (Measurability if $\mathscr{A}$ is Countably Generated)
Let $(\Omega,\mathscr{A})$ be a measurable space and let $\mathscr{A}=\sigma(\mathscr{E})$, where $\mathscr{E}$ is a finite (i.e., $\mathscr{E}=\{A_1,\dots,A_n\}$) or countable (i.e., $\mathscr{E}=\{A_1,A_2,\dots\}$) partition of Ω. Then $f:\Omega\to\overline{\mathbb{R}}$ is $(\mathscr{A},\overline{\mathscr{B}})$-measurable if and only if there are $\alpha_1,\alpha_2,\dots\in\overline{\mathbb{R}}$ such that $f=\sum_{i=1}^{\infty}\alpha_i 1_{A_i}$.

(Proof p. 70)

If $\mathscr{E}=\{A_1,\dots,A_n\}$ is a finite partition of Ω, then choosing $\alpha_{n+1}=\alpha_{n+2}=\dots=0$ implies $f=\sum_{i=1}^{\infty}\alpha_i 1_{A_i}=\sum_{i=1}^{n}\alpha_i 1_{A_i}$.

A Necessary and Sufficient Condition of Measurability

Let $(\Omega',\mathscr{A}')$ be a measurable space and $\mathscr{E}'\subset\mathscr{A}'$. Then we denote

$$f^{-1}(\mathscr{E}') := \{f^{-1}(A'): A'\in\mathscr{E}'\}. \tag{2.11}$$

This notation is used in the following theorem, which can be utilized for proving $(\mathscr{A},\mathscr{A}')$-measurability of a mapping $f:\Omega\to\Omega'$.

Theorem 2.20 (Measurable Mapping and Generating Systems)
Let $(\Omega,\mathscr{A})$, $(\Omega',\mathscr{A}')$ denote measurable spaces, let $\mathscr{E}'\subset\mathscr{A}'$, and $f\colon\Omega\to\Omega'$. Then

$$\sigma[f^{-1}(\mathscr{E}')] = f^{-1}[\sigma(\mathscr{E}')]. \tag{2.12}$$

Furthermore, if $\sigma(\mathscr{E}')=\mathscr{A}'$, then f is $(\mathscr{A},\mathscr{A}')$-measurable if and only if $f^{-1}(A')\in\mathscr{A}$, for all $A'\in\mathscr{E}'$.

For a proof see Klenke (2008, Theorem 1.81, p. 36).

Now consider a finite or countable set Ω'. Then Theorem 2.20 and Example 1.18 immediately imply the following corollary:

Corollary 2.21 (Finite or Countable Generating Systems)
Let $(\Omega,\mathscr{A})$, $\big(\Omega',\mathscr{P}(\Omega')\big)$ be measurable spaces, where Ω' is finite or countable, and let $\mathscr{E}'=\big\{\{\omega'\}\colon\omega'\in\Omega'\big\}$. Then a mapping $f\colon\Omega\to\Omega'$ is $\big(\mathscr{A},\mathscr{P}(\Omega')\big)$-measurable if and only if $f^{-1}(\{\omega'\})\in\mathscr{A}$, for all $\omega'\in\Omega'$.

Example 2.22 (Rectangles – continued) In Example 2.2.1 we considered the mapping $f\colon\Omega\to\Omega'=\Omega$ defined by $f(x)=\frac{3}{4}x$. Furthermore, we considered the rectangle $B'=\,]4.5,7.5]\times[0,4.5]$. The set system

$$\mathscr{E}' = \{B'\},$$

which contains B' as the only element, generates the σ-algebra

$$\mathscr{A}' = \{\Omega',\emptyset,B',(B')^c\}.$$

Hence according to Theorem 2.20, the mapping f is $(\mathscr{A},\mathscr{A}')$-measurable provided that $f^{-1}(B')\in\mathscr{A}$. ◁

Example 2.23 (Flipping two Coins – continued) In Example 2.2.2, we defined the mapping $X=$ *number of flipping heads* with codomain $\Omega'=\{0,1,2\}$. Now consider the system

$$\mathscr{E}' = \big\{\{0\},\{1\}\big\}$$

of subsets of Ω'. First, note that $\sigma(\mathscr{E}')=\mathscr{P}(\Omega')$. Therefore, Theorem 2.20 implies that X is $\big(\mathscr{A},\mathscr{P}(\Omega')\big)$-measurable for each σ-algebra $\mathscr{A}$ on

$$\Omega = \{(h,h),\,(h,t),\,(t,h),\,(t,t)\}$$

for which

$$X^{-1}(\{0\}) = \{(t,t)\}\in\mathscr{A} \quad\text{and}\quad X^{-1}(\{1\}) = \{(h,t),(t,h)\}\in\mathscr{A}.$$

This does not only hold for $\mathscr{A}_1=\mathscr{P}(\Omega)$, but also for the σ-algebra

$$\mathscr{A}_2 = \big\{\Omega,\ \emptyset,\ \{(t,t)\},\ \{(h,t),(t,h)\},\ \{(h,h)\},\\ \{(h,h),(h,t),(t,h)\},\ \{(h,h),(t,t)\},\ \{(h,t),(t,h),(t,t)\}\big\}.$$

As mentioned before, $\mathscr{A}_2$ contains all events that can be represented by X (see Table 2.1, p. 45). In contrast, this does not hold for the σ-algebra

$$\mathscr{A}_0 = \big\{\Omega, \emptyset,\ \{(h,h),(h,t)\},\ \{(t,h),(t,t)\}\big\}$$

(see Example 2.8). Hence, X is measurable with respect to $\mathscr{A}_1$ and $\mathscr{A}_2$, but it is not measurable with respect to $\mathscr{A}_0$. From a substantive point of view this means that the events $\{(h,h),(h,t)\}$ and $\{(t,h),(t,t)\}$ cannot be formulated in terms of X. Furthermore, some of the events that *can* be formulated in terms of X are not elements of $\mathscr{A}_0$. For example, $X^{-1}(\{0\}) = \{(t,t)\}$ is *not* an element of $\mathscr{A}_0$. ◁

2.3.2 σ-Algebra Generated by a Mapping

Let us consider again Example 2.2.2 and the mapping $X = \textit{number of flipping heads}$. The set that consists of the eight inverse images $X^{-1}(A')$, $A' \in \mathscr{A}'$, is again a σ-algebra on Ω. In a sense, this σ-algebra carries the information associated with the mapping X; it contains all events that can be represented by X (see Table 2.1). In the following theorem we formulate the general proposition.

Theorem 2.24 (σ-Algebra Generated by a Mapping)
Let $f\colon \Omega \to \Omega'$ be a mapping and let $(\Omega', \mathscr{A}')$ be a measurable space. Then

$$f^{-1}(\mathscr{A}') := \{f^{-1}(A')\colon A' \in \mathscr{A}'\} \tag{2.13}$$

is a σ-algebra on Ω.

For a proof see Klenke (2008, Theorem 1.81, p. 36)).

Remark 2.25 (Smallest σ-Algebra) Note that $f^{-1}(\mathscr{A}')$ is the smallest σ-algebra $\mathscr{C}$ on Ω such that f is $(\mathscr{C}, \mathscr{A}')$-measurable, i. e.,

$$\mathscr{C} \text{ is a σ-algebra on } \Omega \text{ and } f \text{ is } (\mathscr{C}, \mathscr{A}')\text{-measurable} \quad \Rightarrow \quad f^{-1}(\mathscr{A}') \subset \mathscr{C}.$$

◁

The set $f^{-1}(\mathscr{A}')$ contains all sets in $\mathscr{A}$ that can be represented by f and elements of $\mathscr{A}'$. Because $f^{-1}(\mathscr{A}')$ is important, it has an own name and an alternative notation, which is sometimes more convenient.

Definition 2.26 (σ-Algebra Generated by a Mapping)
The set $f^{-1}(\mathscr{A}')$ defined by Equation (2.13) is called the σ-algebra generated by f and $\mathscr{A}'$. If there is no ambiguity about $\mathscr{A}'$, then we also say that $f^{-1}(\mathscr{A}')$ is generated by f and use the notation

$$\sigma(f) := f^{-1}(\mathscr{A}'). \tag{2.14}$$

Remark 2.27 (Monotonicity) Note that, for two set systems $\mathscr{C}' \subset \mathscr{A}'$,

$$f^{-1}(\mathscr{C}') \subset f^{-1}(\mathscr{A}'), \tag{2.15}$$

because $f^{-1}(\mathscr{C}') = \{f^{-1}(A') \colon A' \in \mathscr{C}'\} \subset \{f^{-1}(A') \colon A' \in \mathscr{A}'\} = f^{-1}(\mathscr{A}')$. ◁

The following corollary immediately follows from Definition 2.26 and the definition of $(\mathscr{A},\mathscr{A}')$-measurability (see Def. 2.5).

Corollary 2.28 (A Condition Equivalent to Measurability)
Let $f\colon \Omega \to \Omega'$ be a mapping and let $(\Omega',\mathscr{A}')$ be a measurable space. Then f is $(\mathscr{A},\mathscr{A}')$-measurable if and only if $\sigma(f) \subset \mathscr{A}$.

In the following lemma and the subsequent remark, we treat a $\cap$-stable generating system (see Def. 1.36).

Lemma 2.29 ($\cap$-Stable Generating System)
Let $(\Omega,\mathscr{A})$ be a measurable space and let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra. Furthermore, assume that Ω' is finite or countable and let $f\colon (\Omega,\mathscr{A}) \to \big(\Omega', \mathscr{P}(\Omega')\big)$ be a measurable mapping. Then the set

$$\mathscr{D} := \big\{C \cap f^{-1}(\{\omega'\}) \colon \omega' \in \Omega' \text{ and } C \in \mathscr{C}\big\}$$

is a $\cap$-stable generating system of $\sigma(\mathscr{C}, f) := \sigma\big(\mathscr{C} \cup f^{-1}[\mathscr{P}(\Omega')]\big)$.

(Proof p. 71)

Remark 2.30 (A Special Case) Let us consider the special case in which $\mathscr{C} = \{\Omega,\emptyset\}$. In this case, Lemma 2.29 simplifies as follows: Let $(\Omega,\mathscr{A})$ be a measurable space and let $f\colon (\Omega,\mathscr{A}) \to \big(\Omega', \mathscr{P}(\Omega')\big)$ be a measurable mapping, where Ω' is finite or countable. Then the set $\big\{ f^{-1}(\{\omega'\}) \colon \omega' \in \Omega' \big\} \cup \{\emptyset\}$ is a $\cap$-stable generating system of $\sigma(f) := f^{-1}[\mathscr{P}(\Omega')]$. ◁

Example 2.31 (σ-Algebra Generated by an Indicator) Let $1_A\colon (\Omega,\mathscr{A}) \to (\mathbb{R},\mathscr{B})$ be the indicator of $A \in \mathscr{A}$. Then $\sigma(1_A) = \{\Omega, \emptyset, A, A^c\}$. The same σ-algebra is generated by $1_A\colon (\Omega,\mathscr{A}) \to \big(\{0,1\}, \mathscr{P}(\{0,1\})\big)$ (see Remark 2.33 for the general proposition). ◁

Example 2.32 (Flipping two Coins – continued) In Example 2.2.2, we considered flipping two coins and the measurable mapping X = *number of flipping heads* with codomain $\Omega' = \{0,1,2\}$. In this example, all elements of $X^{-1}[\mathscr{P}(\Omega')]$

have been listed in Table 2.1 (p. 45) as the inverse images $X^{-1}(A')$ of the eight sets $A' \in \mathscr{P}(\Omega')$. Furthermore, $X^{-1}[\mathscr{P}(\Omega')] = \mathscr{A}_2$, where $\mathscr{A}_2$ is the σ-algebra defined in Example 2.23.

Instead of choosing $\Omega' = \{0,1,2\}$ as the codomain of X, we may also choose the set $\mathbb{R}$ of real numbers, i. e., $X\colon \Omega \to \mathbb{R}$ is then considered to be a function into $\mathbb{R}$. In this case we use the *Borel σ-algebra* $\mathscr{B}$ on $\mathbb{R}$. However, according to the following remark, the σ-algebra $X^{-1}(\mathscr{B})$ generated by X and $\mathscr{B}$ is the same as the σ-algebra $X^{-1}[\mathscr{P}(\Omega')]$ generated by X and the power set of $\Omega' = \{0,1,2\}$. ◁

Remark 2.33 (σ-Algebra Generated by a Function Into a Countable Set) Let us consider a function $f\colon \Omega \to \Omega' \subset \mathbb{R}$ and let $\mathscr{B}$ denote the Borel σ-algebra on $\mathbb{R}$. If Ω' is finite or countable, then $f^{-1}[\mathscr{P}(\Omega')] = f^{-1}(\mathscr{B})$ (see Exercise 2-9). ◁

Example 2.34 (Joe and Ann – continued) Table 2.2 (p. 54) displays mappings on $(\Omega, \mathscr{A})$, all components of which have already been specified in Example 1.9. The *first* mapping displayed in Table 2.2 is the *person variable U* that assigns to each possible outcome $\omega \in \Omega$ the value *Joe* if $\omega \in \{Joe\} \times \Omega_X \times \Omega_Y$ and the value *Ann* if $\omega \in \{Ann\} \times \Omega_X \times \Omega_Y$. Hence, $U\colon \Omega \to \Omega_U$ is a mapping with domain $\Omega = \Omega_U \times \Omega_X \times \Omega_Y$ and codomain Ω_U. It projects the first component u of $\omega = (u, \omega_X, \omega_Y)$ onto the set Ω_U. Therefore, it is also called the first *projection mapping*.

The *second* mapping in this table is the *treatment variable X*. It assigns to each possible outcome $\omega \in \Omega$ the value 0 if $\omega \in \Omega_U \times \{no\} \times \Omega_Y$ and the value 1 if $\omega \in \Omega_U \times \{yes\} \times \Omega_Y$. Hence, $X\colon \Omega \to \Omega'$ is a function with domain Ω and codomain $\Omega' = \{0,1\}$.

The *third* mapping is the *outcome variable Y*. It assigns to each $\omega \in \Omega$ the value 0 if $\omega \in \Omega_U \times \Omega_X \times \{-\}$ and the value 1 if $\omega \in \Omega_U \times \Omega_X \times \{+\}$. Therefore, $Y\colon \Omega \to \Omega'$ is a function with domain Ω and codomain $\Omega' = \{0,1\}$. Hence, all three mapping U, X, and Y have the same domain Ω.

Considering $U\colon \Omega \to \Omega_U$ and the σ-algebra $\mathscr{A}_U := \{\Omega_U, \emptyset, \{Joe\}, \{Ann\}\}$, the σ-algebra $U^{-1}(\mathscr{A}_U)$ consists of the following four inverse images: the event

$$U^{-1}(\{Joe\}) = \{(Joe, no, -),\ (Joe, no, +),\ (Joe, yes, -),\ (Joe, yes, +)\}$$

that *Joe is drawn*, the event

$$U^{-1}(\{Ann\}) = \{(Ann, no, -),\ (Ann, no, +),\ (Ann, yes, -),\ (Ann, yes, +)\}$$

that *Ann is drawn*, the sure event $U^{-1}(\Omega_U) = \Omega$ that *Joe or Ann are drawn*, and the impossible event $U^{-1}(\emptyset) = \emptyset$ that *neither Joe nor Ann are drawn*. ◁

2.3.3 *Final σ-Algebra*

Consider the mapping $f\colon \Omega \to \Omega'$. As noted in Remark 2.25, for a σ-algebra $\mathscr{A}'$ on Ω', $\sigma(f) = f^{-1}(\mathscr{A}')$ is the smallest σ-algebra on Ω for which f is measurable. In contrast, now we consider a σ-algebra $\mathscr{C}$ on Ω and look for the largest σ-algebra $\mathscr{C}'$ on Ω' such that f is $(\mathscr{C}, \mathscr{C}')$-measurable. This σ-algebra is specified in the following lemma. It is called the *final σ-algebra*.

Table 2.2. Joe and Ann With Random Assignment and Measurable Mappings

Elements of Ω			Measurable mappings		
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y
(*Joe*, *no*, −)			.09	*Joe*	0	0
(*Joe*, *no*, +)			.21	*Joe*	0	1
(*Joe*, *yes*, −)			.04	*Joe*	1	0
(*Joe*, *yes*, +)			.16	*Joe*	1	1
(*Ann*, *no*, −)			.24	*Ann*	0	0
(*Ann*, *no*, +)			.06	*Ann*	0	1
(*Ann*, *yes*, −)			.12	*Ann*	1	0
(*Ann*, *yes*, +)			.08	*Ann*	1	1

Lemma 2.35 (Final σ-Algebra)
Let $f: \Omega \to \Omega'$ be a mapping and $\mathscr{C}$ a σ-algebra on Ω.

(i) Then

$$\mathscr{C}'_f := \left\{ A' \subset \Omega' : f^{-1}(A') \in \mathscr{C} \right\} \tag{2.16}$$

is a σ-algebra on Ω'.

(ii) Furthermore, if $f: (\Omega, \mathscr{C}) \to (\Omega', \mathscr{A}')$ is a measurable mapping, then $\mathscr{A}' \subset \mathscr{C}'_f$.

(Proof p. 71)

Note that (ii) is a formal way of saying that $\mathscr{C}'_f$ is the largest σ-algebra on Ω' such that f is $\mathscr{C}$-measurable.

Definition 2.36 (Final σ-Algebra)
The σ-algebra $\mathscr{C}'_f$ defined by Equation (2.16) *is called the final σ-algebra of $\mathscr{C}$ under f.*

2.3.4 Multivariate Mapping

Now consider the measurable space $\left(\bigtimes_{i=1}^n \Omega'_i, \bigotimes_{i=1}^n \mathscr{A}'_i\right)$ and note that the definitions of measurable mappings and of the σ-algebra generated by a mapping also apply to n-variate mappings $f: \Omega \to \Omega'_1 \times \ldots \times \Omega'_n$ and in particular to functions for which $\Omega'_1 \times \ldots \times \Omega'_n = \mathbb{R}^n$.

Lemma 2.37 (σ-Algebra Generated by a Multivariate Mapping)
Let Ω be a nonempty set, $(\Omega_i', \mathscr{A}_i')$, $i = 1,\dots,n$, $n \in \mathbb{N}$, be measurable spaces, and $f = (f_1,\dots,f_n)$ be a multivariate mapping with $f_i\colon \Omega \to \Omega_i'$, $i = 1,\dots,n$, i. e., $f\colon \Omega \to \bigtimes_{i=1}^n \Omega_i'$. Then

$$\sigma(f_1,\dots,f_n) := \sigma(f) = f^{-1}\Big(\bigotimes_{i=1}^n \mathscr{A}_i'\Big) = \sigma\Big(\bigcup_{i=1}^n \sigma(f_i)\Big). \tag{2.17}$$

(Proof p. 72)

According to the following theorem, a multivariate mapping is measurable if and only if all its components are measurable.

Theorem 2.38 (Measurability of Multivariate Mappings)
Under the assumptions of Lemma 2.37, the following two propositions are equivalent to each other:

(a) $f\colon (\Omega,\mathscr{A}) \to \big(\bigtimes_{i=1}^n \Omega_i', \bigotimes_{i=1}^n \mathscr{A}_i'\big)$ *is a measurable mapping.*
(b) $\forall\ i = 1,\dots,n\colon\ f_i\colon (\Omega,\mathscr{A}) \to (\Omega_i', \mathscr{A}_i')$ *is a measurable mapping.*

(Proof p. 73)

Remark 2.39 (σ-Algebra Generated by a Family of Mappings) Let I be a (finite, countable, or uncountable) index set and let $(f_i, i \in I)$ be a family of mappings $f_i\colon (\Omega,\mathscr{A}) \to (\Omega_i', \mathscr{A}_i')$. The σ-algebra generated by this family is defined as

$$\sigma(f_i, i \in I) := \sigma\Big(\bigcup_{i\in I} \sigma(f_i)\Big). \tag{2.18}$$

Then Equation (2.17) implies

$$\sigma(f) = \sigma(f_i, i \in I), \quad \text{where } I = \{i = 1,\dots,n\}. \tag{2.19}$$

◁

Example 2.40 (Joe and Ann – continued) In Example 2.34 we already considered the function $X\colon \Omega \to \mathbb{R}$ indicating with its values 1 and 0 whether or not the drawn person is treated and the function $Y\colon \Omega \to \mathbb{R}$ indicating with its values 1 and 0 whether or not the drawn person is successful. If we specify the σ-algebra $\mathscr{A}$ on Ω such that X and Y are both $(\mathscr{A},\mathscr{B})$-measurable, then the bivariate function $(X,Y)\colon \Omega \to \mathbb{R}^2$ is $(\mathscr{A},\mathscr{B}_2)$-measurable. And vice versa, if we specify the σ-algebra $\mathscr{A}$ on Ω such that the bivariate function $(X,Y)\colon \Omega \to \mathbb{R}^2$ is $(\mathscr{A},\mathscr{B}_2)$-measurable then X and Y are both $(\mathscr{A},\mathscr{B})$-measurable. In this example X, Y, and (X,Y) are measurable with respect to $\mathscr{A}$ whenever the two inverse images $X^{-1}(\{1\})$ and $Y^{-1}(\{1\})$ are elements of $\mathscr{A}$ (see Exercise 2-10). ◁

Remark 2.41 (Lower Dimension Multivariate Mappings) Lemma 2.37 and Remark 1.23 imply

$$\sigma(f_i, i \in J) \subset f^{-1}\left(\bigotimes_{i=1}^{n} \mathscr{A}_i'\right), \quad \forall J \subset \{1,\dots,n\}.$$

Furthermore, Theorem 2.38 implies: If

$$f = (f_1,\dots,f_n)\colon (\Omega,\mathscr{A}) \to \left(\bigtimes_{i=1}^{n} \Omega_i', \bigotimes_{i=1}^{n} \mathscr{A}_i'\right)$$

is a measurable mapping and $J = \{i_1,\dots,i_k\} \subset \{1,\dots,n\}$, $k \le n$, then

$$f_J := (f_{i_1},\dots,f_{i_k})\colon (\Omega,\mathscr{A}) \to \left(\bigtimes_{j=1}^{k} \Omega_{i_j}', \bigotimes_{j=1}^{k} \mathscr{A}_{i_j}'\right)$$

is measurable as well. ◁

2.3.5 Projection Mapping

In Definition 1.31 we introduced the product σ-algebra $\bigotimes_{i=1}^{n} \mathscr{A}_i$ for a finite number of measurable spaces $(\Omega_i,\mathscr{A}_i)$. Now we give an equivalent characterization. Let $(\Omega_i,\mathscr{A}_i)$, $i = 1,\dots,n$, be measurable spaces. Then, for $j = 1,\dots,n$, the jth *projection mapping* $\pi_j\colon \bigtimes_{i=1}^{n} \Omega_i \to \Omega_j$ is defined by

$$\pi_j(\omega_1,\dots,\omega_n) = \omega_j, \quad \forall (\omega_1,\dots,\omega_n) \in \bigtimes_{i=1}^{n} \Omega_i. \tag{2.20}$$

The inverse images are

$$\pi_j^{-1}(A_j) = \Omega_1 \times \dots \times \Omega_{j-1} \times A_j \times \Omega_{j+1} \times \dots \times \Omega_n, \quad \text{for } A_j \subset \Omega_j. \tag{2.21}$$

Lemma 2.42 (Product σ-Algebra)
If $(\Omega_i,\mathscr{A}_i)$, $i = 1,\dots,n$, are measurable spaces, then

$$\bigotimes_{i=1}^{n} \mathscr{A}_i = \sigma(\pi_1,\dots,\pi_n). \tag{2.22}$$

(Proof p. 73)

2.3.6 Measurability With Respect to a Mapping

In the next definition, we consider two mappings and the concept of a mapping being measurable with respect to another mapping.

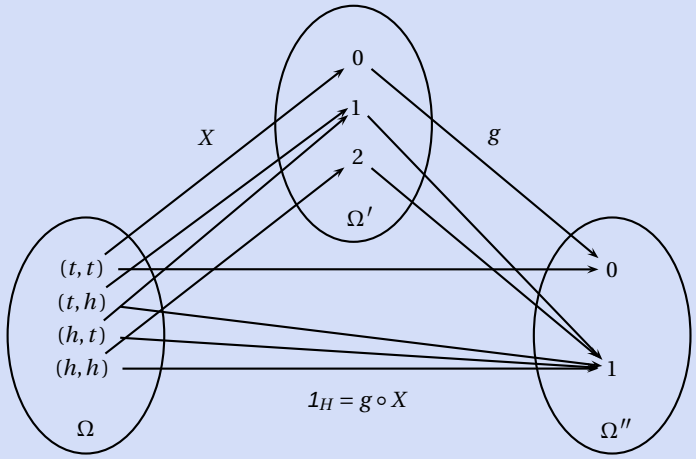


Figure 2.5. A composition of two mappings

Definition 2.43 (Measurability With Respect to a Mapping)
Let $f\colon \Omega \to \Omega'$ and $h\colon \Omega \to \Omega''$ be mappings, and let $(\Omega', \mathcal{A}')$ and $(\Omega'', \mathcal{A}'')$ be measurable spaces. Then h is called measurable w.r.t. f (or f-measurable) if

$$h^{-1}(\mathcal{A}'') \subset f^{-1}(\mathcal{A}'). \tag{2.23}$$

If Ω' is finite or countable, then the following corollary provides a representation for all functions that are measurable with respect to f.

Corollary 2.44 (Measurability With Respect to a Discrete Function)
Let $f\colon \Omega \to \Omega'$ be a mapping, let $\left(\Omega', \mathcal{P}(\Omega')\right)$ be a measurable space, where Ω' is finite or countable, and let $h\colon \Omega \to \bar{\mathbb{R}}$ be a function. Then h is measurable w.r.t. f if and only if for all $\omega' \in \Omega'$ there are $\alpha_{\omega'} \in \bar{\mathbb{R}}$ such that

$$h = \sum_{\omega' \in \Omega'} \alpha_{\omega'} \cdot 1_{f^{-1}(\{\omega'\})}. \tag{2.24}$$

(Proof p. 74)

Example 2.45 (Flipping Two Coins – continued) Consider the mapping $X =$ *number of flipping heads* with codomain $\Omega' = \{0,1,2\}$, let $H := \{(h,t),(h,h),(t,h)\}$, and let $1_H\colon \Omega \to \Omega''$ denote the indicator of H, with $\Omega'' = \{0,1\}$. Hence, 1_H indicates with its values 1 and 0 whether or not at least one *heads* is flipped. If we

consider the σ-algebra $\mathscr{A}' = \mathscr{P}(\Omega')$ on Ω' and the σ-algebra $\mathscr{A}'' = \mathscr{P}(\Omega'')$ on Ω'', then

$$X^{-1}(\mathscr{A}') = \big\{\Omega,\ \emptyset,\ \{(h,h)\},\ \{(h,t),(t,h)\},\ \{(t,t)\},\\ \{(h,h),(h,t),(t,h)\},\ \{(h,h),(t,t)\},\ \{(h,t),(t,h),(t,t)\}\big\}$$

and

$$1_H^{-1}(\mathscr{A}'') = \big\{\Omega,\ \emptyset,\ \{(h,h),(h,t),(t,h)\},\ \{(t,t)\}\big\}.$$

Obviously, $1_H^{-1}(\mathscr{A}'') \subset X^{-1}(\mathscr{A}')$. Therefore, 1_H is measurable with respect to X, but not vice versa. That is, X represents a more detailed information about the outcome of the random experiment than 1_H. Hence, if the value of X is known, then we can compute the value of 1_H, but not vice versa. In our example, Figure 2.5 shows: if $X(\omega) = 1$, then $1_H(\omega) = 1$. However, if $1_H(\omega) = 1$, then $X(\omega) = 1$ or $X(\omega) = 2$. (For a more general presentation of this property see Lemma 2.52.) ◁

2.4 Theorems on Measurable Mappings

In this section we consider compositions of mappings, which are defined as follows: Let Ω, Ω', and Ω'' be nonempty sets and let $f\colon \Omega \to \Omega'$ and $g\colon \Omega' \to \Omega''$ be mappings. Then the *composition of f and g* is the mapping $g \circ f\colon \Omega \to \Omega''$ defined by:

$$g \circ f(\omega) := g[f(\omega)], \quad \forall \omega \in \Omega, \tag{2.25}$$

(see Fig. 2.5), where $g \circ f(\omega)$ denotes the value of the mapping $g \circ f$ for the argument ω. Instead of $g \circ f$, we often use the notation $g(f)$ and say that $g(f)$ *is a function of* f. Using this notation, Equation (2.25) can be written

$$g(f)(\omega) = g[f(\omega)], \quad \forall \omega \in \Omega. \tag{2.26}$$

Lemma 2.46 (Compositions With a Finite or Countable Number of Values) *Let $f\colon \Omega \to \Omega'$ be a mapping, where Ω' is finite or countable, and let $g\colon \Omega' \to \overline{\mathbb{R}}$ be a function. Furthermore, for $\omega' \in \Omega'$, define $1_{f=\omega'} := 1_{f^{-1}(\{\omega'\})}$. Then*

$$g \circ f = g(f) = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f^{-1}(\{\omega'\})} = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f=\omega'}. \tag{2.27}$$

(Proof p. 74)

Hence, under the assumptions of Lemma 2.46, for all $\omega \in \Omega$,

$$g \circ f(\omega) = g[f(\omega)] = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f^{-1}(\{\omega'\})}(\omega) = \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f=\omega'}(\omega). \tag{2.28}$$

Example 2.47 (Flipping two Coins – continued) Let us consider $X =$ *number of flipping heads* and the mapping $g\colon \Omega' \to \Omega''$ defined by $g(x) := 1_{\{1,2\}}(x)$, for all $x \in \Omega'$ (see Fig. 2.5). Then the composition $g \circ X$ defines a new mapping $g \circ X\colon \Omega \to$

Ω'', where $\Omega'' = \{0,1\}$. In this example, the composition $g \circ X$ is identical to the indicator 1_H of the event $H = \{(h,h),(t,h),(h,t)\}$ that heads are flipped at least once. ◁

Example 2.48 (Joe and Ann – continued) In Example 2.34 we already considered the mapping $U: \Omega \to \Omega_U = \{Joe, Ann\}$ showing which person is drawn and the mapping $X: \Omega \to \Omega' = \{0,1\}$ indicating whether or not the drawn person is treated. Now we can consider the bivariate random variable $(U,X): \Omega \to \Omega_U \times \Omega'$ and we can write

$$X = g \circ (U,X) = g(U,X)$$

as the composition of (U,X) and a (projection) mapping g,

$$g[(u,x)] = x, \quad \forall (u,x) \in \Omega_U \times \Omega'.$$

◁

2.4.1 Measurability of a Composition

The following theorem shows that measurability is preserved by the composition of mappings.

Theorem 2.49 (Measurability of a Composition)
If $f: (\Omega, \mathscr{A}) \to (\Omega', \mathscr{A}')$ and $g: (\Omega', \mathscr{A}') \to (\Omega'', \mathscr{A}'')$ are measurable mappings, then the composition $g \circ f$ is $(\mathscr{A}, \mathscr{A}'')$-measurable.

(Proof p. 74)

Remark 2.50 (σ-Algebra Generated by a Composition) Note that

$$(g \circ f)^{-1}(\mathscr{A}'') = f^{-1}[g^{-1}(\mathscr{A}'')] \tag{2.29}$$

(see the proof of Theorem 2.49). ◁

Example 2.51 (Flipping two Coins – continued) Figure 2.5 illustrates Theorem 2.49. If (a) X is $(\mathscr{A}, \mathscr{A}')$-measurable and (b) g is $(\mathscr{A}', \mathscr{A}'')$-measurable, then $1_H = g \circ X$ is $(\mathscr{A}, \mathscr{A}'')$-measurable. Suppose $\mathscr{A}' = \mathscr{P}(\Omega')$ and $\mathscr{A}'' = \mathscr{P}(\Omega'')$, where $\Omega' = \{0,1,2\}$ and $\Omega'' = \{0,1\}$. Then the premise '(a) and (b)' is satisfied if $\mathscr{A}$ is such that $X^{-1}(\mathscr{A}') \subset \mathscr{A}$. If the premise '(a) and (b)' is *not* satisfied, then we cannot conclude that 1_H is $(\mathscr{A}, \mathscr{A}'')$-measurable. Note that in this example 1_H can be $(\mathscr{A}, \mathscr{A}'')$-measurable even if (a) and (b) do not hold. A sufficient requirement is that $\{(t,t)\}$ and $\{(t,h),(h,t),(h,h)\}$, the inverse images of $\{0\}$ and $\{1\}$ under 1_H, respectively, are elements of $\mathscr{A}$ (see Cor. 2.21). ◁

If a mapping h is measurable with respect to a mapping f, then each element in the σ-algebra generated by h is an element in the σ-algebra generated by f. If h is measurable with respect to f, then in a sense, the information represented

by h is already contained in f (cf. section 2.3.2). This is expressed in more formal terms in the following lemma, which is crucial, e. g., in the general definition of conditional expectations $E(Y|X{=}x)$ (see ch. 10).

Lemma 2.52 (Factorization Lemma of Measurable Functions)
Let $f\colon \Omega \to \Omega'$ be a mapping, let $(\Omega', \mathscr{A}')$ be a measurable space, and let $h\colon \Omega \to \overline{\mathbb{R}}$ be a function. Then h is measurable w.r.t. f, i. e., $h^{-1}(\overline{\mathscr{B}}) \subset f^{-1}(\mathscr{A}')$, if and only if there is a measurable function $g\colon (\Omega', \mathscr{A}') \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ such that

$$h = g \circ f \tag{2.30}$$

is the composition of f and g. We call g a factorization of h w.r.t. f.

For a proof see Klenke (2008, Corollary 1.97, pp. 41, 42).

If, instead of $(\overline{\mathbb{R}}, \overline{\mathscr{B}})$ we consider a measurable space $\big(\Omega'', \mathscr{P}(\Omega'')\big)$, where Ω'' is finite or countable, then the elements $\omega'' \in \Omega''$ can be renamed by real numbers such as $1, 2$, etc. Renaming is a one-to-one measurable function, because the σ-algebra on Ω'' is the power set of Ω'' (see Example 2.9). Hence, Lemma 2.52 implies the following corollary:

Corollary 2.53 (Factorization of a Mapping into a Finite or Countable Set)
Let $f\colon \Omega \to \Omega'$ be a mapping, $(\Omega', \mathscr{A}')$ a measurable space, and $h\colon \Omega \to \Omega''$ a mapping, where Ω'' is finite or countable. Then h is measurable w.r.t. f, i. e., $h^{-1}[\mathscr{P}(\Omega'')] \subset f^{-1}(\mathscr{A}')$, if and only if there is a measurable mapping $g\colon (\Omega', \mathscr{A}') \to \big(\Omega'', \mathscr{P}(\Omega'')\big)$ such that $h = g \circ f$.

Example 2.54 (Flipping two Coins – continued) If we specify $\Omega' = \{0, 1, 2\}$, the σ-algebra $\mathscr{A}' = \mathscr{P}(\Omega')$, the set $\Omega'' = \{0, 1\}$, the σ-algebra $\mathscr{A}'' = \mathscr{P}(\Omega'')$, and the function $h = 1_H$, then the example depicted in Figure 2.5 can be used to illustrate this corollary. The mapping g in this figure is such that $1_H = g \circ X$. ◁

Example 2.55 (Two Step Functions) Figure 2.6 presents an example in which $\Omega = [0, 4]$, $A_1 = [0, 1]$, $A_2 = \,]1, 2]$, $A_3 = \,]2, 3]$, and $A_4 = \,]3, 4]$. Note that the sets $A_1, \ldots, A_4$ are pairwise disjoint. The measurable function $f\colon (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ is defined by

$$f = \sum_{i=1}^{4} \alpha_i 1_{A_i},$$

where $\alpha_1 = 1$, $\alpha_2 = 2.5$, $\alpha_3 = 2$, and $\alpha_4 = 0.5$. Furthermore, the function $h\colon \Omega \to \mathbb{R}$ is defined by

$$h = \sum_{j \in \{1,3\}} \beta_j 1_{A_j \cup A_{j+1}}$$

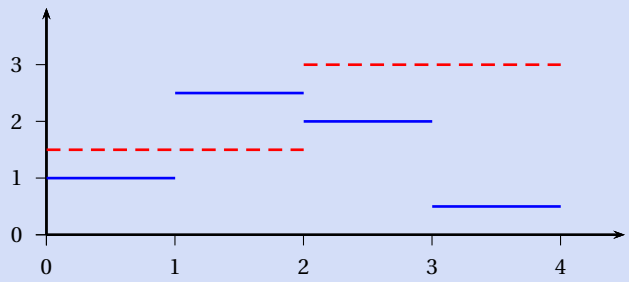


Figure 2.6. Two step functions

with $\beta_1 = 1.5$ and $\beta_3 = 3$. Note that $\sigma(h) = \sigma\big(\{A_j \cup A_{j+1} \colon j \in \{1,3\}\}\big)$, whereas $\sigma(f) = \sigma(\{A_i \colon i = 1,\dots,4\})$ (see Exercise 2-11). Therefore, h is measurable with respect to f, i. e., $\sigma(h) \subset \sigma(f)$.

According to Lemma 2.52, there is a function $g \colon \mathbb{R} \to \mathbb{R}$ such that $h = g \circ f$. In fact, if we define g by

$$g(x) = \sum_{j \in \{1,3\}} \beta_j 1_{\{\alpha_j, \alpha_{j+1}\}}(x), \quad \forall x \in \mathbb{R},$$

then $h = g \circ f$. The function g takes on the value 1.5 if $x = \alpha_1 = 1$ or $x = \alpha_2 = 2.5$ and the value 3 if $x = \alpha_3 = 2$ or $x = \alpha_4 = 0.5$. For all other $x \in \mathbb{R}$ its value is 0. ◁

Example 2.56 (Square of a Real-Valued Function) Suppose $f \colon (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ is a real-valued measurable function and $f^2(\omega) := f(\omega)^2$, for all $\omega \in \Omega$.

(i) If f is *nonnegative*, i. e., if $f(\omega) \geq 0$, for all $\omega \in \Omega$, then f and f^2 are measurable with respect to each other, i. e., $\sigma(f) = \sigma(f^2)$.

(ii) If there are $\omega_1, \omega_2 \in \Omega$ with $f(\omega_1) < 0 < f(\omega_2)$ and $f^2(\omega_1) = f^2(\omega_2)$, then $\sigma(f^2) \subset \sigma(f)$, but $\sigma(f) \neq \sigma(f^2)$

(see Exercise 2-13.) In a sense, $\sigma(f) = \sigma(f^2)$ means that f and f^2 contain the same information, whereas $\sigma(f^2) \subset \sigma(f)$, $\sigma(f) \neq \sigma(f^2)$ means that f^2 contains less information than f. If, e. g., $f^2(\omega) = 4$, then $f(\omega) = 2$ or $f(\omega) = -2$. ◁

2.4.2 Theorems on Measurable Functions

In the first theorem we consider sums and differences as well as products and ratios of measurable functions. The *sum* of two functions $f, h \colon \Omega \to \mathbb{R}^n$ is again a function $(f + h) \colon \Omega \to \mathbb{R}^n$ defined by

$$(f+h)(\omega) := \begin{pmatrix} f_1+h_1 \\ \vdots \\ f_n+h_n \end{pmatrix}(\omega) := \begin{pmatrix} f_1(\omega)+h_1(\omega) \\ \vdots \\ f_n(\omega)+h_n(\omega) \end{pmatrix}, \quad \forall \omega \in \Omega.$$

The first parentheses in the term $(f+h)(\omega)$ are used to make clear that $f+h$ is a symbol of a new function on Ω. Of course, the difference $f-h$ is defined in the same way as $f+h$ replacing + by −.

Similarly, the *product* $f \cdot h$ of two functions $f, h\colon \Omega \to \mathbb{R}$ is again a function $(f \cdot h) : \Omega \to \mathbb{R}$ defined by

$$(f \cdot h)(\omega) := f(\omega) \cdot h(\omega), \quad \forall \omega \in \Omega.$$

Correspondingly, $f/h\colon \Omega \to \mathbb{R}$ is defined by

$$(f/h)(\omega) := f(\omega)/h(\omega), \quad \forall \omega \in \Omega,$$

provided that $h(\omega) \neq 0$ for all $\omega \in \Omega$.

Theorem 2.57 (Sums and Products of Measurable Functions)
If $f, h\colon (\Omega, \mathcal{A}) \to (\mathbb{R}^n, \mathcal{B}_n)$ are measurable functions, then $f+h$ and $f-h$ are $(\mathcal{A}, \mathcal{B}_n)$-measurable as well. Furthermore, if $f, h\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ are measurable functions, then $f \cdot h$ and f/h (with $h(\omega) \neq 0$, for all $\omega \in \Omega$) are also $(\mathcal{A}, \mathcal{B})$-measurable.

For a proof see Klenke (2008, Theorem 1.91, p. 39).

Remark 2.58 (Squared Function) If $f\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is a measurable function, then $f^2 := f \cdot f$ is also $(\mathcal{A}, \mathcal{B})$-measurable. Obviously, this also applies to f^n, $n \in \mathbb{N}$. Hence, if f is $(\mathcal{A}, \mathcal{B})$-measurable, then f^n is also $(\mathcal{A}, \mathcal{B})$-measurable. ◁

Example 2.59 (Scale Transformations and Translations) Remember that a constant real number can always be interpreted as a measurable function (see Example 2.10). Therefore, Theorem 2.57 implies that, for all $\alpha \in \mathbb{R}$, the functions $f+\alpha$, $f-\alpha$, and $\alpha \cdot f$ are $(\mathbb{R}, \mathcal{B})$-measurable if $f\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is a measurable function. ◁

Example 2.60 (Number of Flipping *Heads*) Consider flipping a coin n times, let $\Omega = \{h, t\}^n$, and let $1_{A_i} : \Omega \to \mathbb{R}$ denote the indicators of flipping *heads* at the ith flip of the coin. Then

$$X = \sum_{i=1}^{n} 1_{A_i}$$

is the *number of flipping heads*. If $A_i \in \mathcal{A} = \mathcal{P}(\Omega)$, $i = 1, \ldots, n$, then $(\Omega, \mathcal{A})$ is a measurable space and X is measurable for any σ-algebra on $\mathbb{R}$ (see Example 2.9). In the case $\mathcal{A} = \mathcal{P}(\Omega)$ it is not necessary to apply Theorem 2.57. ◁

Example 2.61 (Linear Combination of two Functions) Let $f, h\colon (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ be measurable functions and $\alpha, \beta \in \mathbb{R}$. Then, according to Theorem 2.57, the function $(\alpha \cdot f + \beta \cdot g)\colon (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ defined by

$$(\alpha \cdot f + \beta \cdot h)(\omega) = \alpha \cdot f(\omega) + \beta \cdot h(\omega), \quad \forall \omega \in \Omega. \tag{2.31}$$

is $(\mathscr{A}, \mathscr{B})$-measurable. ◁

Remark 2.62 (Positive and Negative Parts of a Function) In Theorem 2.66 we consider the positive and the negative parts of a function $f\colon \Omega \to \overline{\mathbb{R}}$. The *positive part* $f^+\colon \Omega \to \overline{\mathbb{R}}$ is defined by

$$f^+(\omega) := \max\big(f(\omega), 0\big), \quad \forall \omega \in \Omega,$$

and the *negative part* $f^-\colon \Omega \to \overline{\mathbb{R}}$ by

$$f^-(\omega) := -\min\big(f(\omega), 0\big), \quad \forall \omega \in \Omega.$$

Hence, the value $f^+(\omega)$ of the positive part of f is defined to be the *greater* one of the two numbers $f(\omega)$ and 0 if they differ and $f^+(\omega) = 0$ if $f(\omega) = 0$. In contrast, the value $f^-(\omega)$ of the negative part of f is defined to be the *smaller* one of the two numbers $f(\omega)$ and 0 *multiplied by* -1 if they differ and $f^-(\omega) = 0$ if $f(\omega) = 0$. Note that f^+ and f^- are both nonnegative functions and that

$$f = f^+ - f^-.$$

◁

Example 2.63 (Positive and Negative Parts of a Function) The positive and negative parts of a function are illustrated by Figure 2.7 showing the graph of the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \dfrac{x^3}{3} - \dfrac{x^2}{2} - 2x, & \text{if } -1.81 < x < 3.315, \\ 0, & \text{otherwise.} \end{cases}$$

The positive part f^+ takes on the value 0 if $x \le 0$ (see the red line on the horizontal axis), whereas negative part f^- takes on the value 0 if $x \ge 0$ (see the blue line on the horizontal axis). ◁

Remark 2.64 (Absolute Value Function) Furthermore, we consider the *absolute value function* $|f|\colon \Omega \to \overline{\mathbb{R}}$ defined by

$$|f|(\omega) := |f(\omega)| := \begin{cases} f(\omega), & \text{if } f(\omega) \ge 0 \\ -f(\omega), & \text{if } f(\omega) < 0. \end{cases}$$

Note that $|f| = f^+ + f^- = \max(f^+, f^-)$. Hence, in Figure 2.7, the absolute value function is represented by the red and blue lines *above* (if $-1.81 < x < 3.315$) or *on* (if $x < -1.81$ or $x > 3.315$) the horizontal axis. ◁

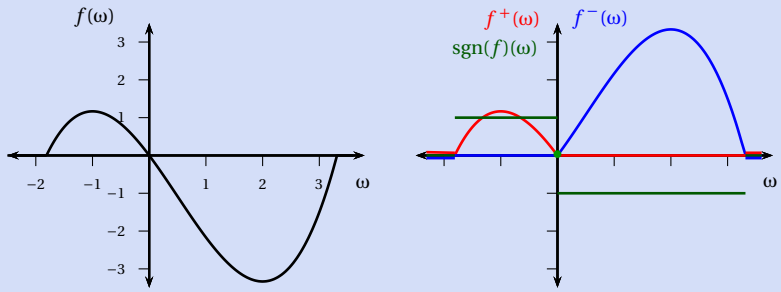


Figure 2.7. The positive and negative parts of a function and the sign function

Remark 2.65 (Sign Function) In Theorem 2.66, we also refer to $\operatorname{sgn}(f)\colon \Omega \to \mathbb{R}$, called the *sign function*, which is defined by

$$\operatorname{sgn}(f)(\omega) = \begin{cases} 1, & \text{if } f(\omega) > 0 \\ 0, & \text{if } f(\omega) = 0 \\ -1, & \text{if } f(\omega) < 0. \end{cases}$$

In Figure 2.7, the graph of this function is represented by the green lines above and below the horizontal axis, and by the green point. ◁

Theorem 2.66 (Positive and Negative Parts of a Function)
Let $(\Omega, \mathscr{A})$ be a measurable space. If $f\colon \Omega \to \overline{\mathbb{R}}$ is $(\mathscr{A}, \overline{\mathscr{B}})$-measurable, then the functions f^+, f^-, $|f|$, and $\operatorname{sgn}(f)$ are $(\mathscr{A}, \overline{\mathscr{B}})$-measurable as well.

For a proof see Klenke (2008, Corollary 1.89, p. 39). The positive part f^+, the negative part f^-, and the absolute value function $|f|$ of a function f plays an important role in integration theory (see ch. 3).

Another implication of Theorem 2.57 on the measurability of some sets that are often used is formulated in the following remark.

Remark 2.67 (Some Important Measurable Sets) Let $(\Omega, \mathscr{A})$ be a measurable space and let $f, g\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable functions. Then

(a) $\{\omega \in \Omega\colon f(\omega \geq g(\omega)\} \in \mathscr{A}$.
(b) $\{\omega \in \Omega\colon f(\omega > g(\omega)\} \in \mathscr{A}$.
(c) $\{\omega \in \Omega\colon f(\omega = g(\omega)\} \in \mathscr{A}$

(see Exercise 2-12). ◁

2.5 Equivalence of Two Mappings With Respect to a Measure

Now we study some properties of mappings $f: \Omega \to \Omega'$ involving a measure space $(\Omega, \mathscr{A}, \mu)$. In this case, we use the notation

$$f: (\Omega, \mathscr{A}, \mu) \to \Omega'$$

to express that $f: \Omega \to \Omega'$ is a mapping and that μ is a measure on the measurable space $(\Omega, \mathscr{A})$. If there is also a σ-algebra $\mathscr{A}'$ on Ω', then we use the notation

$$f: (\Omega, \mathscr{A}, \mu) \to (\Omega', \mathscr{A}')$$

to express that the mapping $f: \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable and that μ is a measure on the measurable space $(\Omega, \mathscr{A})$.

Remember, two mappings f and g are *identical* $(f = g)$ if and only if

$$\{\omega \in \Omega: f(\omega) \neq g(\omega)\} = \emptyset.$$

A less restrictive concept is their *equivalence with respect to the measure* μ.

Definition 2.68 (Equivalence of Two Mappings With Respect to a Measure) *Let $f, g: (\Omega, \mathscr{A}, \mu) \to \Omega'$ be mappings. Then f and g are called μ-equivalent, denoted by*

$$f \underset{\mu}{=} g,$$

if there is an $A \in \mathscr{A}$ with $\mu(A) = 0$ and, for all $\omega \in \Omega \setminus A$, $f(\omega) = g(\omega)$.

Because $\mu(\emptyset) = 0$, $f = g$ implies $f \underset{\mu}{=} g$.

Remark 2.69 (A Note on Notation) If f, g are denoted by $f: (\Omega, \mathscr{A}, \mu) \to \Omega'_f$ and $g: (\Omega, \mathscr{A}, \mu) \to \Omega'_g$, then we can choose $\Omega' = \Omega'_f \cup \Omega'_g$ and denote $f, g: (\Omega, \mathscr{A}, \mu) \to \Omega'$ (see Rem. 2.4). ◁

Remark 2.70 (An Alternative Notation) If $f \underset{\mu}{=} g$ we also say that $f = g$, μ-almost everywhere (μ-a.e.). Furthermore, we also write

$$f(\omega) = g(\omega), \quad \text{for } \mu\text{-almost all } \omega \in \Omega, \tag{2.32}$$

and use $f(\omega) \underset{\mu\text{-a.a.}}{=} g(\omega)$ as a shortcut. ◁

Remark 2.71 (Singleton With a Positive Value of a Measure) If $f \underset{\mu}{=} g$ or, equivalently, $f(\omega) \underset{\mu\text{-a.a.}}{=} g(\omega)$, and $\{\omega\} \in \mathscr{A}$ with $\mu(\{\omega^*\}) > 0$, then

$$f(\omega^*) = g(\omega^*).$$

◁

Remark 2.72 (μ-Equivalence, Restricted Functions, and Compositions) Let f, $g\colon (\Omega, \mathscr{A}, \mu) \to \Omega'$ be mappings.

(i) If $\Omega' = \mathbb{R}$, then

$$f \underset{\mu}{=} g \;\Rightarrow\; 1_A \cdot f \underset{\mu}{=} 1_A \cdot g, \quad \forall\, A \in \mathscr{A}. \tag{2.33}$$

(ii) If $h\colon \Omega' \to \Omega''$ is also a mapping, then

$$f \underset{\mu}{=} g \;\Rightarrow\; h \circ f \underset{\mu}{=} h \circ g \tag{2.34}$$

(see Exercise 2-14). ◁

Remark 2.73 (Equivalence Relation) If $\mathscr{M}$ is a set of mappings $(\Omega, \mathscr{A}, \mu) \to \Omega'$, then $\underset{\mu}{=}$ is an *equivalence relation* on $\mathscr{M}$ (see Exercise 2-15). In other words, if $f, g, h \in \mathscr{M}$, then the following propositions hold for $\underset{\mu}{=}$:

(i) $f \underset{\mu}{=} f$ (reflexivity).

(ii) $g \underset{\mu}{=} f$ if and only if $f \underset{\mu}{=} g$ (symmetry).

(iii) If $f \underset{\mu}{=} g$ and $g \underset{\mu}{=} h$, then $f \underset{\mu}{=} h$ (transitivity).

◁

Definition 2.74 (Equivalence Class With Respect to a Measure)
Let $\mathscr{M}$ be a set of mappings $(\Omega, \mathscr{A}, \mu) \to \Omega'$ and let $f \in \mathscr{M}$. Then

$$C(f) := \{g \in \mathscr{M}\colon g \underset{\mu}{=} f\}$$

is called the μ-equivalence class of f in $\mathscr{M}$ and f a representative of the class $C(f)$.

Remark 2.75 (A Partition of the Set $\mathscr{M}$) If $\mathscr{M}$ is a set of mappings $(\Omega, \mathscr{A}, \mu) \to \Omega'$, then the set $\{C(f)\colon f \in \mathscr{M}\}$ is a *partition* of $\mathscr{M}$, i. e.,

(a) $\forall\, f \in \mathscr{M} : C(f) \neq \emptyset$.

(b) $\forall\, f, g \in \mathscr{M} : C(f) = C(g)$ or $C(f) \cap C(g) = \emptyset$.

(c) $\bigcup_{f \in \mathscr{M}} C(f) = \mathscr{M}$.

(see Exercise 2-16). ◁

Remark 2.76 (Other Properties of μ-Equivalence)

(i) Let $f, g\colon (\Omega, \mathscr{A}, \mu) \to \Omega'$. If $\mu(\Omega) > 0$, then

$$\forall\, \alpha \in \Omega', \forall\, \beta \in \Omega'\colon\; f \underset{\mu}{=} \alpha \;\wedge\; g \underset{\mu}{=} \beta \;\wedge\; f \underset{\mu}{=} g \quad\Rightarrow\quad \alpha = \beta. \tag{2.35}$$

(ii) If $f, g, f^*, g^*\colon (\Omega, \mathscr{A}, \mu) \to \mathbb{R}$, then

$$\begin{aligned} f \underset{\mu}{=} f^* \wedge g \underset{\mu}{=} g^* \quad \Rightarrow \quad & f + g \underset{\mu}{=} f^* + g^*, \\ & f - g \underset{\mu}{=} f^* - g^*, \\ & f \cdot g \underset{\mu}{=} f^* \cdot g^*. \end{aligned} \tag{2.36}$$

Furthermore, suppose $\mu\big(\{\omega \in \Omega\colon g(\omega) = 0\}\big) = 0$, and define $\dfrac{f}{g}\colon \Omega \to \mathbb{R}$ by

$$\frac{f}{g}(\omega) := \begin{cases} \dfrac{f(\omega)}{g(\omega)}, & \text{if } g(\omega) \neq 0, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall \omega \in \Omega,$$

and let $\dfrac{f^*}{g^*}$ be defined analogously. Then

$$f \underset{\mu}{=} f^* \wedge g \underset{\mu}{=} g^* \;\Rightarrow\; \frac{f}{g} \underset{\mu}{=} \frac{f^*}{g^*}. \tag{2.37}$$

(iii) If $f_i, f_i^*\colon (\Omega, \mathscr{A}, \mu) \to \mathbb{R}$ and $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$, then

$$\big(\forall\, i = 1, \ldots, n\colon \; f_i \underset{\mu}{=} f_i^*\big) \;\Rightarrow\; \sum_{i=1}^{n} \alpha_i f_i \underset{\mu}{=} \sum_{i=1}^{n} \alpha_i f_i^*. \tag{2.38}$$

(iv) If $f_1, f_2, \ldots, f_1^*, f_2^*, \ldots : (\Omega, \mathscr{A}, \mu) \to \mathbb{R}$ and $\alpha_1, \alpha_2, \ldots \in \mathbb{R}$, then

$$\big(\forall\, i = 1, 2, \ldots\colon \; f_i \underset{\mu}{=} f_i^*\big) \;\Rightarrow\; \sum_{i=1}^{\infty} \alpha_i f_i \underset{\mu}{=} \sum_{i=1}^{\infty} \alpha_i f_i^*, \tag{2.39}$$

provided that the limits denoted by the infinite sums (see Box 0.1) exist.

For proofs see Exercise 2-17. ◁

Remark 2.77 (Order Relations With Respect to a Measure μ) For functions such as $f, g, h\colon (\Omega, \mathscr{A}, \mu) \to \mathbb{R}$, we also use the notation

$$f \underset{\mu}{\leq} g,$$

if there is an $A \in \mathscr{A}$ with $f(\omega) < g(\omega)$ for all $\omega \in \Omega \setminus A$ and $\mu(A) = 0$. The notation $f \underset{\mu}{>} g$, $f \underset{\mu}{\leqq} g$, and $f \underset{\mu}{\geqq} g$ is used correspondingly. Furthermore,

$$f \underset{\mu}{\leq} g \quad \text{and} \quad g \underset{\mu}{=} h \quad \Rightarrow \quad f \underset{\mu}{\leq} h. \tag{2.40}$$

The analog propositions hold for $\underset{\mu}{>}$, $\underset{\mu}{\leqq}$, and $\underset{\mu}{\geqq}$ (see Exercise 2-18). ◁

2.6 Image Measure

In the definition of a measurable mapping $f\colon (\Omega,\mathscr{A}) \to (\Omega',\mathscr{A}')$ we required: for all $A' \in \mathscr{A}'$: $f^{-1}(A') \in \mathscr{A}$. Because a measure μ assigns a value to *all* elements $A \in \mathscr{A}$, the measure μ also assigns a value to each $f^{-1}(A') := \{\omega \in \Omega\colon f(\omega) \in A'\}$. This is the reason for choosing the term *measurable mapping*: If μ is a measure on $\mathscr{A}$ and f is $(\mathscr{A},\mathscr{A}')$-measurable, then there is a value $\mu[f^{-1}(A')]$ for all inverse images $f^{-1}(A')$, $A' \in \mathscr{A}'$.

According to the following theorem, a measurable mapping $f\colon (\Omega,\mathscr{A},\mu) \to (\Omega',\mathscr{A}')$ induces a measure on the codomain space $(\Omega',\mathscr{A}')$.

Theorem 2.78 (Image Measure)
Let $f\colon (\Omega,\mathscr{A},\mu) \to (\Omega',\mathscr{A}')$ be a measurable mapping. Then the function $\mu_f\colon \mathscr{A}' \to \overline{\mathbb{R}}$ defined by

$$\mu_f(A') := \mu[f^{-1}(A')], \quad \forall\, A' \in \mathscr{A}', \tag{2.41}$$

is a measure on the measurable space $(\Omega',\mathscr{A}')$.

(Proof p. 74)

Definition 2.79 (Image Measure)
If $f\colon (\Omega,\mathscr{A},\mu) \to (\Omega',\mathscr{A}')$ is a measurable mapping, then $\mu_f\colon \mathscr{A}' \to \mathbb{R}$ defined by Equation (2.41) is called the image measure of μ under f.

Example 2.80 (Rectangles – continued) Now we consider a measure $\mu\colon \mathscr{A} \to \mathbb{R}$ restricted to $\mathscr{A}$, which is specified by

$$\mu(A) \;=\; (7-2)\cdot(5-2) \;=\; 15$$

and

$$\mu(\Omega) \;=\; (10-0)\cdot(6-0) \;=\; 60.$$

This specification determines the areas of all four sets in $\mathscr{A}$, because $\mu(A^c) = \mu(\Omega) - \mu(A) = 60 - 15 = 45$ and $\mu(\emptyset) = 0$. Hence, the measure space $(\Omega,\mathscr{A},\mu)$ is completely determined. Note that μ is the restriction of the Lebesgue measure λ_2 to the σ-algebra $\mathscr{A}$, i. e., $\mu(A) = \lambda_2(A)$, for all $A \in \mathscr{A}$.

In Example 2.2.1 we considered the mapping $f\colon \Omega \to \Omega' = \Omega$ defined by $f(x) = \frac{3}{4}x$. Furthermore, we considered the rectangle $B' =]4.5, 7.5] \times [0, 4.5]$ and the σ-algebra $\mathscr{A}' = \{\Omega', \emptyset, B', (B')^c\}$. If we specify $\mathscr{A}$ such that $f^{-1}(B') \in \mathscr{A}$, then f is $(\mathscr{A},\mathscr{A}')$-measurable. In this case, all inverse images $f^{-1}(A')$ of sets $A' \in \mathscr{A}'$ are elements of the σ-algebra $\mathscr{A}$. Therefore, the areas $\lambda_2[f^{-1}(A')]$ of these inverse images are defined by the measure λ_2 on $\mathscr{A}$ that assigns the area to *all* elements of $\mathscr{A}$. If we specify $\mathscr{A} = \{\Omega, \emptyset, f^{-1}(B'), f^{-1}[(B')^c]\}$, then

$$\lambda_{2_f}(B') \;=\; \lambda_2[f^{-1}(B')] \;=\; \lambda_2\big(]6,\ 10] \times [0,\ 6]\big) \;=\; (10-6)\cdot(6-0) \;=\; 24,$$

$$\lambda_{2_f}\big((B')^c\big) \;=\; \lambda_2\big(f^{-1}[(B')^c]\big) \;=\; \lambda_2\big([0,6]\times[0,6]\big) \;=\; (6-0)\cdot(6-0) \;=\; 36,$$

$\lambda_{2_f}(\Omega') = 60$, and $\lambda_{2_f}(\emptyset) = 0$. Then the function $\lambda_{2_f}\colon \mathscr{A}' \to \mathbb{R}$ defined by

$$\lambda_{2_f}(B') \;=\; \lambda_2[f^{-1}(B')], \quad \forall\, B' \in \mathscr{A}',$$

is again a measure, the *image measure* of λ_2 under f. Therefore, $(\Omega', \mathscr{A}', \lambda_{2_f})$ is a measure space.

Note that the image measure λ_{2_f} on the σ-algebra $\mathscr{A}'$ differs from the area measure on $\mathscr{A}'$. In fact, the area of B' is $(7.5-4.5)\cdot 4.5 = 13.5$ and the area of $(B')^c$ is $60 - 13.5 = 46.5$. ◁

Remark 2.81 (Cumulation of the Values $\mu(\{\omega\})$) If $\{\omega\} \in \mathscr{A}$, for all $\omega \in \Omega$, then

$$\mu_f(\{\omega'\}) \;=\; \sum_{\omega\colon f(\omega)=\omega'} \mu(\{\omega\}), \tag{2.42}$$

provided that the sum is over a finite or countable number of summands. The measure μ assigns to the singletons and other elements $A \in \mathscr{A}$ a nonnegative number $\mu(A)$, and f maps each element $\omega \in \Omega$ to an element ω' in Ω'. Thereby it translates the values $\mu(A)$ of the measure μ to their images $f(A)$. In particular, this applies to the singletons $\{\omega\}$. This is illustrated in the following example. ◁

Example 2.82 (Flipping Two Coins – continued) In this example,

$$P(\{\omega\}) \;=\; \frac{1}{4}, \quad \forall \omega \in \Omega, \tag{2.43}$$

uniquely defines a measure $P\colon \mathscr{P}(\Omega) \to \mathbb{R}$ and the measure space $\big(\Omega, \mathscr{P}(\Omega), P\big)$. The reason is that the singletons $\{\omega\}$ are pairwise disjoint and Rule (x) of Box 1.1 implies

$$P(A) \;=\; P\Big(\bigcup_{\omega\in A}\{\omega\}\Big) \;=\; \sum_{\omega\in A} P(\{\omega\}), \quad \forall\, A \in \mathscr{A}.$$

For instance, the set $A =$ *flipping one and only one head* is the union $A = \{(h,t)\} \cup \{(t,h)\} = \{(h,t),\ (t,h)\}$. Hence,

$$P(A) \;=\; \sum_{\omega\in A} P(\{\omega\}) \;=\; P\big(\{(h,t)\}\big) + P\big(\{(t,h)\}\big) \;=\; \frac{1}{4} + \frac{1}{4} \;=\; \frac{1}{2}.$$

Now consider Figure 2.4 and realize that each arrow translates the value $\mu(\{\omega\}) = \frac{1}{4}$ from left to right. According to Equation (2.42), this yields

$$P_X(\{0\}) \;=\; P[X^{-1}(\{0\})] \;=\; P[\{(t,t)\}] \;=\; \frac{1}{4},$$

$$P_X(\{1\}) \;=\; P[X^{-1}(\{1\})] \;=\; P[\{(t,h),(h,t)\}] \;=\; \frac{2}{4},$$

and

$$P_X(\{2\}) \;=\; P[X^{-1}(\{2\})] \;=\; P[\{(h,h)\}] \;=\; \frac{1}{4}.$$

◁

Example 2.83 (Image Measure Under a Step Function) If $f\colon (\Omega, \mathscr{A}, \mu) \to (\mathbb{R}, \mathscr{B})$ is measurable such that $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ with pairwise different $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $\alpha_i \neq 0$, and pairwise disjoint $A_i \in \mathscr{A}$, $i = 1, \ldots, n$, and if we define $A_{n+1} := \Omega \setminus \left(\bigcup_{i=1}^{n} A_i\right)$, and $\alpha_{n+1} := 0$, then the image measure is

$$\mu_f = \sum_{i=1}^{n+1} \mu(A_i) \cdot \delta_{\alpha_i} \tag{2.44}$$

(see Exercise 2-19). Equation (2.44) generalizes Equation (2.42): For all $\omega \in A_i$

$$f(\omega) = \alpha_i \cdot 1_{A_i}(\omega) = \alpha_i.$$

Hence, f translates the value $\mu(A_i)$ to $\alpha_i \in \mathbb{R}$ and μ_f assigns the value $\mu(A_i)$ to the singleton $\{\alpha_i\}$, $i = 1, \ldots, n+1$. ◁

Our next theorem deals with the image measures of μ-equivalent measurable mappings. As a random variable is a particular measurable mapping and the distribution of a random variable a particular image measure (see section 5.1), this theorem has important implications on all concepts that in some sense describe properties of distributions of random variables such as expectations, variances, covariances, etc.

Theorem 2.84 (μ-Equivalence Implies Equality of Image Measures)
If $f, g\colon (\Omega, \mathscr{A}, \mu) \to (\Omega', \mathscr{A}')$ are measurable mappings, then

$$f \underset{\mu}{=} g \quad \Rightarrow \quad \mu_f = \mu_g. \tag{2.45}$$

(Proof p. 75)

In the following theorem we present a necessary and sufficient condition for μ-equivalence of two compositions $g \circ f$ and $g^* \circ f$.

Theorem 2.85 (μ-Equivalence of Compositions)
If $f\colon (\Omega, \mathscr{A}, \mu) \to (\Omega', \mathscr{A}')$ and $g, g^*\colon (\Omega', \mathscr{A}') \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ are measurable mappings, then

$$g \underset{\mu_f}{=} g^* \quad \Leftrightarrow \quad g \circ f \underset{\mu}{=} g^* \circ f. \tag{2.46}$$

(Proof p. 75)

2.7 Proofs

Proof of Lemma 2.19

(a) If $f = \sum_i \alpha_i 1_{A_i}$, then for all $B \in \overline{\mathscr{B}}$,

$$f^{-1}(B) = \bigcup_{i:\alpha_i \in B} A_i \in \mathscr{A},$$

because $\mathscr{A}$ is closed with respect to finite and countable unions.

(b) Assume that there are no $\alpha_1, \alpha_2, \ldots \in \overline{\mathbb{R}}$ such that $f = \sum_i \alpha_i 1_{A_i}$. Then there are an i and elements $\omega_1, \omega_2 \in A_i$ with $f(\omega_1) \neq f(\omega_2)$. Applying Equation (2.5) yields

$$f^{-1}\big(\{f(\omega_1)\}\big) \cap f^{-1}\big(\{f(\omega_2)\}\big) = \emptyset.$$

Furthermore, because $\omega_j \in f^{-1}\big(\{f(\omega_j)\}\big)$, $j = 1,2$,

$$f^{-1}\big(\{f(\omega_1)\}\big) \cap A_i \neq \emptyset, \quad \text{and} \quad f^{-1}\big(\{f(\omega_2)\}\big) \cap A_i \neq \emptyset.$$

Therefore, we conclude: $f^{-1}\big(\{f(\omega_1)\}\big) \notin \mathscr{A}$ and $f^{-1}\big(\{f(\omega_2)\}\big) \notin \mathscr{A}$. Because $\{f(\omega_1)\}, \{f(\omega_2)\} \in \overline{\mathscr{B}}$, it follows that f is not $(\mathscr{A}, \overline{\mathscr{B}})$-measurable.

Proof of Lemma 2.29

(i) $\cap$-stability of $\mathscr{D}$. If $C_1, C_2 \in \mathscr{C}$, and $\omega_1', \omega_2' \in \Omega'$, then

$$\begin{aligned}
&[C_1 \cap f^{-1}(\{\omega_1'\})] \cap [C_2 \cap f^{-1}(\{\omega_2'\})] \\
&= (C_1 \cap C_2) \cap [f^{-1}(\{\omega_1'\}) \cap f^{-1}(\{\omega_2'\})] && [\cap \text{ is associative and commutative}] \\
&= \begin{cases} (C_1 \cap C_2) \cap f^{-1}(\{\omega_1'\}), & \text{if } \omega_1' = \omega_2' \\ \emptyset, & \text{if } \omega_1' \neq \omega_2' \end{cases} && [(2.5)]
\end{aligned}$$

is an element of $\mathscr{D}$, because $C_1 \cap C_2 \in \mathscr{C}$ and $\emptyset \in \mathscr{C}$, which follows from the definition of a σ-algebra .

(ii) Denote $\mathscr{A}' = \mathscr{P}(\Omega')$ and define $\sigma(\mathscr{D}) = \sigma(\mathscr{C}, f) := \sigma[\mathscr{C} \cup f^{-1}(\mathscr{A}')]$.

(a) $\sigma(\mathscr{D}) \subset \sigma[\mathscr{C} \cup f^{-1}(\mathscr{A}')]$. Obviously, $\{f^{-1}(\{\omega'\})\colon \omega' \in \Omega'\} \subset f^{-1}(\mathscr{A}')$. Therefore,

$$\begin{aligned}
\mathscr{D} &= \{C \cap f^{-1}(\{\omega'\})\colon \omega' \in \Omega', C \in \mathscr{C}\} \\
&\subset \{C \cap f^{-1}(A')\colon A' \in \mathscr{A}', C \in \mathscr{C}\} \\
&\subset \sigma[\mathscr{C} \cup f^{-1}(\mathscr{A}')]. && [\text{Rem. 1.2}]
\end{aligned}$$

Hence, according to Remark 1.23, $\sigma(\mathscr{D}) \subset \sigma[\mathscr{C} \cup f^{-1}(\mathscr{A}')]$.

(b) $\sigma[\mathscr{C} \cup f^{-1}(\mathscr{A}')] \subset \sigma(\mathscr{D})$. Because $\Omega \in \mathscr{C}$ and $\Omega \in f^{-1}(\mathscr{A}')$, all $C \in \mathscr{C}$ and all $f^{-1}(A')$ are elements of $\sigma(\mathscr{D})$ (see Def. 1.1, Ω' is finite or countable). Therefore, $\mathscr{C} \cup f^{-1}(\mathscr{A}') \subset \sigma(\mathscr{D})$. Proposition (1.11) then implies $\sigma[\mathscr{C} \cup f^{-1}(\mathscr{A}')] \subset \sigma(\mathscr{D})$.

Proof of Lemma 2.35

(i) We have to show that $\mathscr{C}_f'$ satisfies conditions (a) to (c) of Definition 1.1.

(a)

$$\Omega = f^{-1}(\Omega') \in \mathscr{C} \;\Rightarrow\; \Omega' \in \mathscr{C}_f'. \qquad [(2.16)]$$

(b)

$$\begin{aligned}
A' \in \mathscr{C}_f' &\;\Rightarrow\; f^{-1}(A') \in \mathscr{C} && [(2.16)] \\
&\;\Rightarrow\; f^{-1}(A')^c = f^{-1}[(A')^c] \in \mathscr{C} && [\text{Def. 1.1 (b), (2.4)}] \\
&\;\Rightarrow\; (A')^c \in \mathscr{C}_f'. && [(2.16)]
\end{aligned}$$

(c)

$$\begin{aligned} A_1', A_2', \ldots \in \mathscr{C}_f' &\Rightarrow f^{-1}(A_1'), f^{-1}(A_2'), \ldots \in \mathscr{C} && [(2.16)] \\ &\Rightarrow \bigcup_{i=1}^{\infty} f^{-1}(A_i') = f^{-1}\Big(\bigcup_{i=1}^{\infty} A_i'\Big) \in \mathscr{C} && [\text{Def. 1.1 (c), (2.6)}] \\ &\Rightarrow \bigcup_{i=1}^{\infty} A_i' \in \mathscr{C}_f'. && [(2.16)] \end{aligned}$$

(ii) For all $A' \in \mathscr{A}'$,

$$\begin{aligned} A' \in \mathscr{A}' &\Rightarrow f^{-1}(A') \in \mathscr{C} && [(\mathscr{C},\mathscr{A}')\text{-measurability of } f] \\ &\Rightarrow A' \in \mathscr{C}_f'. && [(2.16)] \end{aligned}$$

Hence, $(\mathscr{C},\mathscr{A}')$-measurability of f implies $\mathscr{A}' \subset \mathscr{C}_f'$.

Proof of Lemma 2.37

First, note that, for $A_i' \in \mathscr{A}_i'$, $i = 1,\ldots,n$,

$$\begin{aligned} f^{-1}(A_1' \times \ldots \times A_n') &= \{\omega \in \Omega\colon f(\omega) \in A_1' \times \ldots \times A_n'\} \\ &= \{\omega \in \Omega\colon \big(f_1(\omega),\ldots,f_n(\omega)\big) \in A_1' \times \ldots \times A_n'\} \\ &= \{\omega \in \Omega\colon f_1(\omega) \in A_1', \ldots, f_n(\omega) \in A_n'\} \\ &= \bigcap_{i=1}^{n} \{\omega \in \Omega\colon f_i(\omega) \in A_i'\} \\ &= \bigcap_{i=1}^{n} f_i^{-1}(A_i'). \end{aligned} \tag{2.47}$$

Hence,

$$\begin{aligned} \sigma(f) &= \big\{f^{-1}(A')\colon A' \in \bigotimes_{i=1}^{n} \mathscr{A}_i'\big\} && [\text{Def. 2.26}] \\ &= \sigma\big(\big\{f^{-1}(A_1' \times \ldots \times A_n')\colon A_i' \in \mathscr{A}_i', i = 1,\ldots,n\big\}\big) && [\text{Th. 2.20, Defs. 1.13, 1.31}] \\ &= \sigma\big(\big\{f_1^{-1}(A_1') \cap \ldots \cap f_n^{-1}(A_n')\colon A_i' \in \mathscr{A}_i', i = 1,\ldots,n\big\}\big) && [(2.47)] \\ &\supset \sigma\Big(\bigcup_{i=1}^{n}\big\{f_i^{-1}(A_i') \cap \bigcap_{j=1,j\neq i}^{n} f_j^{-1}(\Omega_j')\colon A_i' \in \mathscr{A}_i', i = 1,\ldots,n\big\}\Big) && [\text{Rem. 1.23}] \\ &= \sigma\Big(\bigcup_{i=1}^{n}\big\{f_i^{-1}(A_i')\colon A_i' \in \mathscr{A}_i', i = 1,\ldots,n\big\}\Big) && [f_j^{-1}(\Omega_j') = \Omega] \\ &= \sigma\Big(\bigcup_{i=1}^{n}\sigma(f_i)\Big). && [\text{Def. 2.26}] \end{aligned}$$

Furthermore,

$$\begin{aligned} &\big\{f_1^{-1}(A_1') \cap \ldots \cap f_n^{-1}(A_n') : A_i' \in \mathscr{A}_i', i = 1,\ldots,n\big\} \\ &\subset \sigma\left(\bigcup_{i=1}^{n}\big\{f_i^{-1}(A_i') : A_i' \in \mathscr{A}_i', i = 1,\ldots,n\big\}\right). \qquad [\text{Rem. 1.2, finite intersections}] \end{aligned}$$

Therefore,

$$\begin{aligned}
\sigma\Big(\bigcup_{i=1}^{n}\sigma(f_i)\Big) &= \sigma\Big(\bigcup_{i=1}^{n}\big\{f_i^{-1}(A_i')\colon A_i'\in\mathscr{A}_i',\, i=1,\dots,n\big\}\Big) && \text{[Def. 2.26]}\\
&\supset \sigma\Big(\big\{f_1^{-1}(A_1')\cap\dots\cap f_n^{-1}(A_n')\colon A_i'\in\mathscr{A}_i',\, i=1,\dots,n\big\}\Big) && \text{[Rem. 1.23]}\\
&= \sigma(f).
\end{aligned}$$

Hence, $\sigma(f)=\sigma\big(\bigcup_{i=1}^{n}\sigma(f_i)\big)$.

Proof of Theorem 2.38

(b) ⇒ (a): For all $i=1,\dots,n$: Let $A_i'\in\mathscr{A}_i'$. If f_i is measurable, then $f_i^{-1}(A_i')\in\mathscr{A}$. Hence,

$$\begin{aligned}
f^{-1}(A_1'\times\dots\times A_n') &= \{\omega\in\Omega\colon f(\omega)\in A_1'\times\dots\times A_n'\}\\
&= \bigcap_{i=1}^{n} f_i^{-1}(A_i')\in\mathscr{A}.
\end{aligned}$$

Because $\{A_1'\times\dots\times A_n' : A_i'\in\mathscr{A}_i', i=1,\dots,n\}$ is a generating system of $\bigotimes_{i=1}^{n}\mathscr{A}_i'$, Theorem 2.20 implies that f is measurable.

(a) ⇒ (b): If f is measurable, then for all $i=1,\dots,n$,

$$\begin{aligned}
f_i^{-1}(\mathscr{A}_i') &= \big\{f_i^{-1}(A_i')\colon A_i'\in\mathscr{A}_i'\big\}\\
&= \big\{f_i^{-1}(A_i')\cap\bigcap_{j=1,\,j\neq i}^{n} f_j^{-1}(\Omega_j')\colon A_i'\in\mathscr{A}_i'\big\}\\
&= \big\{f^{-1}(\Omega_1'\times\dots\times\Omega_{i-1}'\times A_i'\times\Omega_{i+1}'\times\dots\times\Omega_n')\colon A_i'\in\mathscr{A}_i'\big\} && \text{[(2.47)]}\\
&\subset f^{-1}\Big(\bigotimes_{i=1}^{n}\mathscr{A}_i'\Big)\subset\mathscr{A}.
\end{aligned}$$

Proof of Lemma 2.42

Consider

$$\pi = (\pi_1,\dots,\pi_n)\colon\Big(\bigtimes_{i=1}^{n}\Omega_i,\bigotimes_{i=1}^{n}\mathscr{A}_i\Big)\to\Big(\bigtimes_{i=1}^{n}\Omega_i,\bigotimes_{i=1}^{n}\mathscr{A}_i\Big).$$

Analogously to the proof of Lemma 2.37 and using Definition 1.31,

$$\begin{aligned}
\bigotimes_{i=1}^{n}\mathscr{A}_i &= \sigma(\{A_1\times\dots\times A_n\colon A_i\in\mathscr{A}_i,\, i=1,\dots,n\})\\
&= \sigma(\{\pi^{-1}(A_1\times\dots\times A_n)\colon A_i\in\mathscr{A}_i,\, i=1,\dots,n\}) && \text{[(2.20), (2.21)]}\\
&= \sigma(\pi) && \text{[Th. 2.20, (2.12)]}\\
&= \sigma(\pi_i,\, i=1,\dots,n). && \text{[Lem. 2.37, (2.18)]}
\end{aligned}$$

Proof of Corollary 2.44

If Ω' is finite or countable and we consider the measurable space $\big(\Omega', \mathscr{P}(\Omega')\big)$, then $\sigma(f) = \sigma\big[\big\{f^{-1}(\{\omega'\})\colon \omega' \in \Omega'\big\}\big]$ [see Lemma 2.29 with $\mathscr{C} = \{\Omega, \emptyset\}$]. Because $\big\{f^{-1}(\{\omega'\})\colon \omega' \in \Omega'\big\}$ is a finite or countable partition of Ω, this corollary is an immediate implication of Lemma 2.19.

Proof of Lemma 2.46

For all $\omega \in \Omega$ and all $\omega' \in \Omega'$,

$$g[f(\omega)] \cdot 1_{f=\omega'}(\omega) \;=\; g(\omega') \cdot 1_{f=\omega'}(\omega) \;=\; \begin{cases} 0, & \text{if } f(\omega) \neq \omega' \\ g(\omega'), & \text{if } f(\omega) = \omega'. \end{cases} \tag{2.48}$$

This equation is equivalent to

$$g(f) \cdot 1_{f=\omega'} \;=\; g(\omega') \cdot 1_{f=\omega'}. \tag{2.49}$$

Because the set $\{f^{-1}(\{\omega'\})\colon \omega' \in \Omega'\}$ is a finite or countable partition of Ω we can conclude: $1_\Omega = \sum_{\omega' \in \Omega'} 1_{f=\omega'}$. Therefore,

$$g(f) \;=\; g(f) \cdot 1_\Omega \;=\; \sum_{\omega' \in \Omega'} g(f) \cdot 1_{f=\omega'} \;=\; \sum_{\omega' \in \Omega'} g(\omega') \cdot 1_{f=\omega'},$$

and this implies Equation (2.27).

Proof of Theorem 2.49

If $f\colon (\Omega, \mathscr{A}) \to (\Omega', \mathscr{A}')$, $g\colon (\Omega', \mathscr{A}') \to (\Omega'', \mathscr{A}'')$ are measurable mappings, then, according to Corollary 2.28, $f^{-1}(\mathscr{A}') \subset \mathscr{A}$ and $g^{-1}(\mathscr{A}'') \subset \mathscr{A}'$. Hence, for all $A'' \in \mathscr{A}''$,

$$\begin{aligned} (g \circ f)^{-1}(A'') \;&=\; \{\omega \in \Omega\colon g[f(\omega)] \in A''\} \\ &=\; \{\omega \in \Omega\colon f(\omega) \in g^{-1}(A'')\} \\ &=\; f^{-1}[g^{-1}(A'')]. \end{aligned}$$

Furthermore,

$$\begin{aligned} f^{-1}[g^{-1}(\mathscr{A}'')] \;&\subset\; f^{-1}(\mathscr{A}') && [(2.15)] \\ &\subset\; \mathscr{A}. && [(2.15)] \end{aligned}$$

Proof of Theorem 2.78

We show that μ_f has the properties (a) to (c) required in Definition 1.43. For each property of μ_f we use the corresponding property of μ.

(a) $\mu_f(\emptyset) = \mu[f^{-1}(\emptyset)] = \mu(\emptyset) = 0$.

(b) For all $A' \in \mathscr{A}'$: $\mu_f(A') \;=\; \mu[f^{-1}(A')] \;\geq\; 0$.

(c) If $A'_1, A'_2, \ldots \in \mathscr{A}'$ are pairwise disjoint, then, according to Equation (2.5), for $i \neq j$,

$$f^{-1}(A'_i) \cap f^{-1}(A'_j) \;=\; f^{-1}(A'_i \cap A'_j) \;=\; f^{-1}(\emptyset),$$

i. e., the inverse images $f^{-1}(A'_1), f^{-1}(A'_2), \ldots$ are pairwise disjoint as well. Therefore,

$$\begin{aligned}
\mu_f\Big(\bigcup_{i=1}^{\infty} A'_i\Big) &= \mu\Big(f^{-1}\Big(\bigcup_{i=1}^{\infty} A'_i\Big)\Big) && [(2.41)]\\
&= \mu\Big(\bigcup_{i=1}^{\infty} f^{-1}(A'_i)\Big) && [(2.6)]\\
&= \sum_{i=1}^{\infty} \mu\big(f^{-1}(A'_i)\big) && [\text{Def. 1.43 (c)}]\\
&= \sum_{i=1}^{\infty} \mu_f(A'_i). && [(2.41)]
\end{aligned}$$

Proof of Theorem 2.84

If $f \underset{\mu}{=} g$, then there is a set A satisfying

$$f(\omega) = g(\omega), \quad \forall\, \omega \in \Omega \setminus A \text{ and } \mu(A) = 0.$$

Monotonicity of μ implies $\mu(\{\omega \in A\colon f(\omega) \in A'\}) = 0 = \mu(\{\omega \in A\colon g(\omega) \in A'\})$ for all $A' \in \mathscr{A}'$. Hence, using additivity of μ,

$$\begin{aligned}
\mu_f(A') &= \mu[f^{-1}(A')]\\
&= \mu(\{\omega \in \Omega \setminus A\colon f(\omega) \in A'\}) + \mu(\{\omega \in A\colon f(\omega) \in A'\})\\
&= \mu(\{\omega \in \Omega \setminus A\colon g(\omega) \in A'\}) + \mu(\{\omega \in A\colon g(\omega) \in A'\})\\
&= \mu[g^{-1}(A')] = \mu_g(A').
\end{aligned}$$

Proof of Theorem 2.85

For measurable functions $g, g^*\colon (\Omega', \mathscr{A}') \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ define $A' := \{\omega' \in \Omega'\colon g(\omega') \neq g^*(\omega')\}$. Note that $A' \in \mathscr{A}'$ [see Rem. 2.67 (c)]. Then

$$\begin{aligned}
f^{-1}(A') &= \big\{\omega \in \Omega\colon f(\omega) \in A'\big\}\\
&= \big\{\omega \in \Omega\colon g[f(\omega)] \neq g^*[f(\omega)]\big\}\\
&= \big\{\omega \in \Omega\colon (g \circ f)(\omega) \neq (g^* \circ f)(\omega)\big\}.
\end{aligned}$$

Hence, $g \underset{\mu_f}{=} g^* \Leftrightarrow \mu_f(A') = 0 \Leftrightarrow \mu[f^{-1}(A')] = 0 \Leftrightarrow g \circ f \underset{\mu}{=} g^* \circ f$.

2.8 Exercises

▷ **Exercise 2-1** Prove Equations (2.4) to (2.6).

▷ **Exercise 2-2** Consider Example 2.2.1 and compute the inverse images of the sets $\{(4.5, 0)\}$, $\{(7.5, 0)\}$, $\{(7.5, 4.5)\}$, and $\{(4.5, 4.5)\}$ under the function

$$f(x_1, x_2) = \frac{3}{4} \cdot (x_1, x_2) = \Big(\frac{3}{4} \cdot x_1, \frac{3}{4} \cdot x_2\Big).$$

▷ **Exercise 2-3** Consider Example 2.2.1 and specify the inverse images of the rectangles $[8,10]\times[0,2]$ and $[3,7.5]\times[0,3]$ under the function $f\colon \Omega \to \Omega'$ defined by $f(x_1,x_2) = \frac{3}{4}\cdot(x_1,x_2)$.

▷ **Exercise 2-4** Consider Example 2.2.1 and use Equation (2.4) to determine the inverse image $f^{-1}[(C')^c]$.

▷ **Exercise 2-5** Prove the proposition of Example 2.14.

▷ **Exercise 2-6** Prove the proposition of Example 2.15.

▷ **Exercise 2-7** Prove the proposition of Example 2.17.

▷ **Exercise 2-8** Prove the proposition of Remark 2.18.

▷ **Exercise 2-9** Prove the proposition formulated in Remark 2.33.

▷ **Exercise 2-10** Consider Example 2.40 and show that X, Y, and (X,Y) are measurable with respect to $\mathscr{A}$ whenever the two inverse images $X^{-1}(\{1\})$ and $Y^{-1}(\{1\})$ are elements of $\mathscr{A}$.

▷ **Exercise 2-11** In Example 2.55 we considered $\Omega = [0,4]$, $A_1 = [0,1]$, $A_2 =]1,2]$, $A_3 =]2,3]$, and $A_4 =]3,4]$. There, we also defined the functions f and h. Show that $\sigma(h) = \sigma(\{A_j \cup A_{j+1}\colon j \in J\})$ is a subset of $\sigma(f) = \sigma(\{A_i\colon i = 1,\dots,4\})$

▷ **Exercise 2-12** Proof the propositions of Remark 2-12.

▷ **Exercise 2-13** Prove the proposition of Example 2.56.

▷ **Exercise 2-14** Prove the propositions of Remark 2.72.

▷ **Exercise 2-15** Consider Remark 2.73 and show: If $\mathscr{M}$ is a set of mappings $(\Omega,\mathscr{A},\mu) \to \Omega'$, then $\underset{\mu}{=}$ is an equivalence relation on $\mathscr{M}$.

▷ **Exercise 2-16** Show that $\{C(f)\colon f \in \mathscr{M}\}$ is a partition of $\mathscr{M}$ (see Remark 2.75).

▷ **Exercise 2-17** Prove the propositions of Remark 2.76.

▷ **Exercise 2-18** Prove proposition (2.40).

▷ **Exercise 2-19** Prove the proposition of Example 2.83.

Solutions

▷ **Solution 2-1** Equation (2.4):

$$f^{-1}[(A')^c] = \{\omega\in\Omega\colon f(\omega)\in(A')^c\} = \{\omega\in\Omega\colon f(\omega)\notin A'\} = [f^{-1}(A')]^c.$$

Equation (2.5):

$$\begin{aligned} f^{-1}\Big(\bigcap_{i\in I} A_i'\Big) &= \{\omega\in\Omega\colon f(\omega)\in\bigcap_{i\in I}A_i'\} = \{\omega\in\Omega\colon f(\omega)\in A_i', \forall\, i\in I\} \\ &= \bigcap_{i\in I}\{\omega\in\Omega\colon f(\omega)\in A_i'\} = \bigcap_{i\in I} f^{-1}(A_i'). \end{aligned}$$

Equation (2.6):

$$f^{-1}\Big(\bigcup_{i\in I} A_i'\Big) = \{\omega\in\Omega\colon f(\omega)\in\bigcup_{i\in I} A_i'\} = \{\omega\in\Omega\colon \exists\, i\in I\colon f(\omega)\in A_i'\}$$
$$= \bigcup_{i\in I}\{\omega\in\Omega\colon f(\omega)\in A_i'\} = \bigcup_{i\in I} f^{-1}(A_i').$$

▷ **Solution 2-2** The inverse images are the sets $f^{-1}[\{(4.5,0)\}] = \{(6,0)\}$, $f^{-1}[\{(7.5,0)\}] = \{(10,0)\}$, $f^{-1}[\{(7.5,4.5)\}] = \{(10,6)\}$, and $f^{-1}[\{(4.5,4.5)\}] = \{(6,6)\}$.

▷ **Solution 2-3** $f^{-1}([8,10]\times[0,2]) = \emptyset$ and $f^{-1}([3,7.5]\times[0,3]) = [4,10]\times[0,4]$.

▷ **Solution 2-4** According to Equation (2.4), the inverse image of $(C')^c$ under f is

$$\begin{aligned} f^{-1}[(C')^c] &= [f^{-1}(C')]^c \\ &= \Omega \setminus (]6,10]\times[0,6]) \\ &= ([0,10]\times[0,6]) \setminus (]6,10]\times[0,6]) \\ &= [0,6]\times[0,6]. \end{aligned}$$

▷ **Solution 2-5** If $f\colon\Omega\to\Omega'$ is constant, then, according to Example 2.10, it is $(\mathscr{A},\mathscr{A}')$-measurable for $\mathscr{A} = \{\Omega,\emptyset\}$. Now, assume that f is not constant, i. e., $\exists\omega_1,\omega_2\in\Omega\colon f(\omega_1)\neq f(\omega_2)$. According to our assumptions,

$$\{f(\omega_1)\},\{f(\omega_2)\}\in\mathscr{A}'.$$

Furthermore, $\omega_i\in f^{-1}[f\{(\omega_i)\}]$, for $i=1,2$, i. e., the inverse images are nonempty sets. Now, $f(\omega_1)\neq f(\omega_2)$ implies

$$\{f(\omega_1)\}\cap\{f(\omega_2)\} = \emptyset,$$

and, using Equation (2.5),

$$f^{-1}\big(\{f(\omega_1)\}\big)\cap f^{-1}\big(\{f(\omega_2)\}\big) = f^{-1}\big(\{f(\omega_1)\}\cap\{f(\omega_2)\}\big) = f^{-1}(\emptyset) = \emptyset.$$

Hence, the inverse images are nonempty disjoint sets, and therefore none of them is in $\mathscr{A} = \{\Omega,\emptyset\}$. This implies that f is not $(\mathscr{A},\mathscr{A}')$-measurable if it is not constant.

▷ **Solution 2-6** We consider $\{A, A^c, \Omega, \emptyset\}$. If $f = \alpha_1 1_A + \alpha_2 1_{A^c}$, then for all $A'\in\mathscr{B}$,

$$f^{-1}(A') = \begin{cases} \emptyset, & \text{if } \alpha_1\notin A',\ \alpha_2\notin A' \\ A, & \text{if } \alpha_1\in A',\ \alpha_2\notin A' \\ A^c, & \text{if } \alpha_1\notin A',\ \alpha_2\in A' \\ \Omega, & \text{if } \alpha_1\in A',\ \alpha_2\in A'. \end{cases}$$

Hence, f is $(\mathscr{A},\mathscr{B})$-measurable. (Note that this also holds if $A=\emptyset$ or $A=\Omega$, and also if $\alpha_1=\alpha_2$.)

Now assume that f is an $(\mathscr{A},\mathscr{B})$-measurable function.

(a) If f takes one only one single value, say α, then

$$f = \alpha 1_\Omega = \alpha_1 1_A + \alpha_2 1_{A^c}, \quad \text{with } \alpha_1=\alpha_2=\alpha.$$

(b) If f takes on exactly two different values $\beta_1\neq\beta_2$, then $f^{-1}(\{\beta_1,\beta_2\}) = f^{-1}(\{\beta_1\})\cup f^{-1}(\{\beta_2\}) = \Omega$, and according to Equation (2.5), $f^{-1}(\{\beta_1\})\cap f^{-1}(\{\beta_2\}) = \emptyset$, and $f^{-1}(\{\beta_i\})\neq\emptyset$, for $i=1,2$. Hence, f is $(\mathscr{A},\mathscr{B})$-measurable if and only if

$$f^{-1}(\{\beta_1\}) = A \quad \text{or} \quad f^{-1}(\{\beta_1\}) = A^c \quad \text{and} \quad A, A^c \neq \emptyset.$$

This implies

$$f = \beta_1 1_A + \beta_2 1_{A^c} \quad \text{or} \quad f = \beta_2 1_A + \beta_1 1_{A^c},$$

respectively.

(c) If f takes on *three or more* pairwise different values, then, using the same kind of argument as in (a), we can conclude that there are at least three pairwise disjoint and nonempty inverse images under f, say $A_1, A_2, A_3 \subset \Omega$. Hence, in this case f is not $(\mathscr{A},\mathscr{B})$-measurable.

▷ **Solution 2-7** If $A_1, \ldots, A_n \in \mathscr{A}$ are pairwise disjoint and we define $A_{n+1} := \Omega \setminus \left(\bigcup_{i=1}^n A_i\right)$, and $\alpha_{n+1} := 0$, then

$$f = \sum_{i=1}^{n} \alpha_i 1_{A_i} = \sum_{i=1}^{n+1} \alpha_i 1_{A_i}.$$

Because $A_1, \ldots, A_{n+1}$ are pairwise disjoint and $\bigcup_{i=1}^{n+1} A_i = \Omega$, there is, for all $\omega \in \Omega$, exactly one $i \in \{1, \ldots, n+1\}$ such that $\omega \in A_i$, and therefore $f(\omega) = \alpha_i$. Hence, the codomain of f is $\{\alpha_1, \ldots, \alpha_{n+1}\}$. Vice versa, for all α_i, $i = 1, \ldots, n+1$, we obtain the inverse image

$$f^{-1}(\{\alpha_i\}) = \{\omega \in \Omega \colon f(\omega) = \alpha_i\} = \bigcup_{j \colon \alpha_i = \alpha_j} A_j \tag{2.50}$$

(If the $\alpha_1, \ldots, \alpha_{n+1}$ are pairwise different, then $f^{-1}(\{\alpha_i\}) = A_i$.) Now, for all $A' \subset \mathbb{R}$,

$$\begin{aligned}
f^{-1}(A') &= \{\omega \in \Omega \colon f(\omega) \in A'\} && \text{[Def. 2.1]} \\
&= \Big\{ \omega \in \Omega \colon f(\omega) \in \bigcup_{i \colon \alpha_i \in A'} \{\alpha_i\} \Big\} && \text{[codomain of } f \text{ is } \{\alpha_1, \ldots, \alpha_{n+1}\}] \\
&= f^{-1}\left(\bigcup_{i \colon \alpha_i \in A'} \{\alpha_i\} \right) && \text{[Def. 2.1]} \\
&= \bigcup_{i \colon \alpha_i \in A'} f^{-1}(\{\alpha_i\}) && \text{[(2.6)]} \\
&= \bigcup_{i \colon \alpha_i \in A'} A_i. && \text{[(2.50)]}
\end{aligned}$$

▷ **Solution 2-8** If $A_1, \ldots, A_n \in \mathscr{A}$ are pairwise disjoint and if we define $A_{n+1} := \Omega \setminus \left(\bigcup_{i=1}^n A_i\right)$, and $\alpha_{n+1} := 0$, then

$$\forall\, A' \in \mathscr{B} \colon\ f^{-1}(A') = \bigcup_{\substack{i=1,\ldots,n+1, \\ \alpha_i \in A'}} A_i \ \in \mathscr{A},$$

[see Eq. (2.10)].

If $A_1, \ldots, A_n \in \mathscr{A}$ are *not* pairwise disjoint, define the 2^n sets

$$B_j := A_1^{c_1(j)} \cap \ldots \cap A_n^{c_n(j)} \in \mathscr{A}, \quad j = 1, \ldots, 2^n,$$

with $\big(c_1(j), \ldots, c_n(j)\big) \in \{0,1\}^n$ and

$$A_i^0 := A_i, \qquad A_i^1 := A_i^c.$$

Note that some of the sets B_j can be empty. Then

$$f = \sum_{i=1}^{n} \alpha_i 1_{A_i} = \sum_{j=1}^{2^n} \beta_j 1_{B_j},$$

with $\beta_j = \sum_{i:\, c_i(j)=0} \alpha_i$. Because $B_1, \ldots, B_{2^n}$ are pairwise disjoint and $\bigcup_{j=1}^{2^n} B_j = \Omega$, the function f is $(\mathscr{A},\mathscr{B})$-measurable (see the first part of this solution).

▷ **Solution 2-9** If all values of f are elements of Ω', then

$$f^{-1}(B) \;=\; f^{-1}(\Omega' \cap B), \quad \forall B \in \mathscr{B}.$$

Therefore, $f^{-1}(\mathscr{B}) = f^{-1}(\mathscr{B}|_{\Omega'})$, where $\mathscr{B}|_{\Omega'}$ denotes the trace of $\mathscr{B}$ in Ω' (see Example 1.10). Note that $\mathscr{B}|_{\Omega'} = \mathscr{P}(\Omega')$ (see Exercise 1-13). Hence, $f^{-1}(\mathscr{B}) = f^{-1}(\mathscr{B}|_{\Omega'}) = f^{-1}[\mathscr{P}(\Omega')]$.

▷ **Solution 2-10** First of all note that X, Y are measurable with respect to $\mathscr{A}$ if and only if and (X,Y) is measurable with respect to $\mathscr{A}$ (see Theorem 2.38). Hence, it suffices to show that X is measurable with respect to $\mathscr{A}$ if $X^{-1}(\{1\}) \in \mathscr{A}$. Because $X\colon \Omega \to \mathbb{R}$ is an indicator, $X^{-1}(\{0\}) = X^{-1}(\{1\}^c) = \big(X^{-1}(\{1\})\big)^c$ [see Eq. (2.5)]. Hence, if $X^{-1}(\{1\}) \in \mathscr{A}$, then $\big(X^{-1}(\{1\})\big)^c = X^{-1}(\{0\}) \in \mathscr{A}$. Furthermore, for all $B \in \mathscr{B}$,

$$X^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B, \\ X^{-1}(\{1\}), & \text{if } 0 \notin B, 1 \in B \\ \Omega \setminus X^{-1}(\{1\}), & \text{if } 0 \in B, 1 \notin B \\ \Omega, & \text{if } \{0,1\} \subset B. \end{cases}$$

(The proof for Y is analog.)

▷ **Solution 2-11** Because the codomains of f and h are finite,

$$\begin{aligned} \sigma(f) \;&=\; \sigma\Big(f^{-1}\big(\mathscr{P}(\{1, 2.5, 2, 0.5\})\big)\Big) && \text{[Rem. 2.33]} \\ &=\; \sigma\Big(\big\{f^{-1}(\{1\}), f^{-1}(\{2.5\}), f^{-1}(\{2\}), f^{-1}(\{0.5\})\big\}\Big) && \text{[(2.12)]} \\ &=\; \sigma(\{A_1, A_2, A_3, A_4\}). \end{aligned}$$

Analogously,

$$\begin{aligned} \sigma(g) \;&=\; \sigma\Big(\big\{g^{-1}(\{1.5\}), g^{-1}(\{3\})\big\}\Big) \\ &=\; \sigma(\{A_1 \cup A_2,\, A_3 \cup A_4\}). \end{aligned}$$

Because

$$\{A_1 \cup A_2,\, A_3 \cup A_4\} \;\subset\; \sigma(\{A_1, A_2, A_3, A_4\}) \qquad \text{[Rem. 1.2]}$$

monotonicity of the generated σ-algebras (see Remark 1.23) implies $\sigma(g) \subset \sigma(f)$.

▷ **Solution 2-12** Denote $A_\infty := \{\omega \in \Omega\colon f(\omega) = \infty\}$, $A_{-\infty} := \{\omega \in \Omega\colon f(\omega) = -\infty\}$, $B_\infty := \{\omega \in \Omega\colon g(\omega) = \infty\}$, and $B_{-\infty} := \{\omega \in \Omega\colon g(\omega) = -\infty\}$. Because $\{\infty\}, \{-\infty\} \in \overline{\mathscr{B}}$, all four sets defined above are elements of $\mathscr{A}$. Furthermore,

$$A := \{\omega \in \Omega\colon -\infty < f(\omega) < \infty\} = f^{-1}(\mathbb{R}) \in \mathscr{A}$$

and

$$B := \{\omega \in \Omega\colon -\infty < g(\omega) < \infty\} = g^{-1}(\mathbb{R}) \in \mathscr{A}.$$

(a) Now

$$\begin{aligned} \{\omega \in \Omega\colon f(\omega \geq g(\omega)\} \;&=\; A_\infty \cup \{\omega \in A \cap B\colon f(\omega) \geq g(\omega)\} \cup B_{-\infty} \\ &=\; A_\infty \cup \{\omega \in A \cap B\colon f(\omega) - g(\omega) \geq 0\} \cup B_{-\infty} \\ &=\; A_\infty \cup \big[\mathbf{1}_{A\cap B} \cdot (f-g)\big]^{-1}\big([0,\infty[\big) \cup B_{-\infty} && \text{[Def. 2.1]} \\ &\in\; \mathscr{A}. && \text{[Th. 2.57]} \end{aligned}$$

(b) Analogously,

$$\begin{aligned}\{\omega\in\Omega\colon f(\omega>g(\omega)\} &= \big(A_\infty\cap(B\cup B_{-\infty}\big)\cup\{\omega\in A\cap B\colon f(\omega>g(\omega)\}\\ &= \big(A_\infty\cap(B\cup B_{-\infty}\big)\cup\big[1_{A\cap B}\cdot(f-g)\big]^{-1}\big(]0,\infty[\big) && \text{[Def. 2.1]}\\ &\in\mathscr{A}. && \text{[Th. 2.57]}\end{aligned}$$

(c) Finally,

$$\begin{aligned}\{\omega\in\Omega\colon f(\omega=g(\omega)\} &= \{\omega\in\Omega\colon f(\omega\geq g(\omega)\}\setminus\{\omega\in\Omega\colon f(\omega>g(\omega)\}\\ &\in\mathscr{A}. && \text{[Rem. 1.2]}\end{aligned}$$

▷ **Solution 2-13** For any real-valued measurable function $f\colon(\Omega,\mathscr{A})\to(\mathbb{R},\mathscr{B})$, Lemma 2.52 yields $\sigma(f^2)\subset\sigma(f)$, because $f^2=g(f)$ for $g\colon(\mathbb{R},\mathscr{B})\to(\mathbb{R},\mathscr{B})$ with the measurable function $g(x)=x^2$, for all $x\in\mathbb{R}$. [Note that g is a continuous function that is $\mathscr{B}$-measurable (see Klenke, 2008, Th. 1.88, p. 38)].

(i) We assume that f is nonnegative and measurable. Then $f^2(\omega)=x$ if and only if $f(\omega)=\sqrt{x}$, for all $x\geq 0$. Hence, for all $A\in\mathscr{A}$,

$$\begin{aligned}A\in\sigma(f) &\Rightarrow \exists B_1\in\mathscr{B}\colon A=f^{-1}(B_1)\\ &\Rightarrow \exists B_2\in\mathscr{B}\colon A=(f^2)^{-1}(B_2) && \text{[choose } B_2:=g^{-1}(B_1)\text{]}\\ &\Rightarrow A\in\sigma(f^2).\end{aligned}$$

This implies $\sigma(f)\subset\sigma(f^2)$.

(ii) Assume that there are $\omega_1,\omega_2\in\Omega$ with $f(\omega_1)<0<f(\omega_2)$ and $f^2(\omega_1)=f^2(\omega_2)$. Then $A:=f^{-1}(]-\infty,0[)$ implies $A\in f^{-1}(\mathscr{B})$, and $\omega_1\in A$ and $\omega_2\notin A$. Furthermore, for all $B\in\mathscr{B}$: $\{\omega_1,\omega_2\}\subset(f^2)^{-1}(B)$ if $f^2(\omega_1)\in B$ and $\{\omega_1,\omega_2\}\cap(f^2)^{-1}(B)=\emptyset$ if $f^2(\omega_1)\notin B$. Hence, $A\notin(f^2)^{-1}(\mathscr{B})$.

▷ **Solution 2-14** (i) If $f\underset{\mu}{=}g$, then there is a set $B\in\mathscr{A}$ with $\mu(B)=0$ and $f(\omega)=g(\omega)$ for all $\omega\in\Omega\setminus B$. Hence, $1_A(\omega)\cdot f(\omega)=1_A(\omega)\cdot g(\omega)$ for all $\omega\in\Omega\setminus B$. According to Definition 2.68, $1_Af\underset{\mu}{=}1_Ag$.

(ii) Note that $A_h:=\big\{\omega\in\Omega\colon h[f(\omega)]\neq h[g(\omega)]\big\}\subset\big\{\omega\in\Omega\colon f(\omega)\neq g(\omega)\big\}=:A$. Therefore, $\mu(A)=0$ implies $\mu(A_h)=0$ [see Box 1.1 (v)].

▷ **Solution 2-15** Reflexivity. $\mu\big(\{\omega\in\Omega\colon f(\omega)\neq f(\omega)\}\big)=\mu(\emptyset)=0$. Hence, $f\underset{\mu}{=}f$.

Symmetry. Assume that $f,g\in\mathscr{M}$ and $f\underset{\mu}{=}g$. Then

$$\mu\big(\{\omega\in\Omega\colon g(\omega)\neq f(\omega)\}\big)=\mu\big(\{\omega\in\Omega\colon f(\omega)\neq g(\omega)\}\big)=0.$$

Hence, $g\underset{\mu}{=}f$.

Transitivity. Assume that $f,g,h\in\mathscr{M}$, $f\underset{\mu}{=}g$, and $g\underset{\mu}{=}h$. Then transitivity of = and subadditivity of μ yield

$$\begin{aligned}\mu\big(\{\omega\in\Omega\colon f(\omega)\neq h(\omega)\}\big) &\leq \mu\big(\{\omega\in\Omega\colon f(\omega)\neq g(\omega)\}\cup\{\omega\in\Omega\colon g(\omega)\neq h(\omega)\}\big)\\ &\leq \mu\big(\{\omega\in\Omega\colon f(\omega)\neq g(\omega)\}\big)+\mu\big(\{\omega\in\Omega\colon g(\omega)\neq h(\omega)\}\big)\\ &= 0+0 \qquad [f\underset{\mu}{=}g,\ g\underset{\mu}{=}h]\\ &= 0.\end{aligned}$$

Therefore, $f\underset{\mu}{=}h$.

▷ **Solution 2-16** (a) $\forall f \in \mathcal{M} : f \underset{\mu}{=} f$ (reflexivity). This implies: $\forall f \in \mathcal{M} : f \in C(f)$ and therefore $\forall f \in \mathcal{M} : C(f) \neq \emptyset$.

(b) Let $f, g \in \mathcal{M}$. We consider two cases, $f \underset{\mu}{=} g$ and $\neg(f \underset{\mu}{=} g)$.
$f \underset{\mu}{=} g$. Transitivity implies: $\forall h \in \mathcal{M} : f \underset{\mu}{=} h$ if and only if $g \underset{\mu}{=} h$. Hence, $\forall h \in \mathcal{M}$: $h \in C(f)$ if and only if $h \in C(g)$. This means that $C(f) = C(g)$.
$\neg(f \underset{\mu}{=} g)$. We show $C(f) \cap C(g) = \emptyset$ by contraposition. Assume: $\exists h \in \mathcal{M} : h \in C(f) \cap C(g)$. Then $f \underset{\mu}{=} h$, $h \underset{\mu}{=} g$, and transitivity implies: $f \underset{\mu}{=} g$, which is a contradiction to $\neg(f \underset{\mu}{=} g)$.

(c) In part (a) we showed that $\forall f \in \mathcal{M} : f \in C(f)$. Therefore, $\forall f \in \mathcal{M} : f \in \bigcup_{f^* \in \mathcal{M}} C(f^*)$.

▷ **Solution 2-17** (i) Suppose that $\mu \neq 0, \alpha, \beta \in \Omega'$ and $f \underset{\mu}{=} \alpha \ \wedge\ g \underset{\mu}{=} \beta \ \wedge\ f \underset{\mu}{=} g$. If $\alpha \neq \beta$, then subadditivity and monotonicity of μ yield

$$
\begin{aligned}
0 \ &< \ \mu(\Omega) \\
&= \ \mu\big(\{\omega \in \Omega\colon f(\omega) = \alpha \wedge g(\omega) = \beta\}\big) \ + \ \mu\big(\{\omega \in \Omega\colon f(\omega) \neq \alpha \vee g(\omega) \neq \beta\}\big) && \text{[Box 1.1 (iv)]} \\
&= \ \mu\big(\{\omega \in \Omega\colon f(\omega) = \alpha \wedge g(\omega) = \beta\}\big) && [f \underset{\mu}{=} \alpha,\ g \underset{\mu}{=} \beta] \\
&\leq \ \mu\big(\{\omega \in \Omega\colon f(\omega) \neq g(\omega)\}\big) && [\alpha \neq \beta] \\
&= \ 0, && [f \underset{\mu}{=} g]
\end{aligned}
$$

which proves (i) by contraposition.

(ii) If $f \underset{\mu}{=} f^*$ and $g \underset{\mu}{=} g^*$ and $A_0 := \{\omega \in \Omega\colon f(\omega) \neq f^*(\omega) \vee g(\omega) \neq g^*(\omega)\}$, then

$$
\begin{aligned}
\mu(A_0) \ &= \ \mu\big(\{\omega \in \Omega\colon f(\omega) \neq f^*(\omega)\} \cup \{\omega \in \Omega\colon g(\omega) \neq g^*(\omega)\}\big) \\
&\leq \ \mu\big(\{\omega \in \Omega\colon f(\omega) \neq f^*(\omega)\}\big) \ + \ \mu\big(\{\omega \in \Omega\colon g(\omega) \neq g^*(\omega)\}\big) && \text{[Box 1.1 (vii), (v)]} \\
&= \ 0. && [f \underset{\mu}{=} f^*,\ g \underset{\mu}{=} g^*]
\end{aligned}
$$

Note that $\{\omega \in \Omega\colon f(\omega) + g(\omega) \neq f^*(\omega) + g^*(\omega)\} \subset A_0$, and that this also holds for the corresponding sets for the difference, product, and ratio. This implies Equations (2.36) and (2.37).

(iii), (iv) Denote $I := \{i = 1, \ldots, n\}$ for (iii) and $I := \mathbb{N}$ for (iv), respectively. Furthermore, define

$$
A_0 \ := \ \bigcup_{i \in I} \{\omega \in \Omega\colon f_i(\omega \neq f_i^*(\omega)\} \ = \ \{\omega \in \Omega\colon \exists\, i \in I\colon f_i(\omega \neq f_i^*(\omega)\}.
$$

Then

$$
\begin{aligned}
\mu(A_0) \ &\leq \ \sum_{i \in I} \{\omega \in \Omega\colon f_i(\omega \neq f_i^*(\omega)\} && \text{[Box 1.1 (xi)]} \\
&= \ 0, \quad \text{if } f_i \underset{\mu}{=} f_i^*,\ \forall\, i \in I.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mu\Big(\big\{\,\omega \in \Omega\colon \sum_{i \in I} \alpha_i f_i(\omega) \neq \sum_{i \in I} \alpha_i f_i^*(\omega)\big\}\Big) \ &\leq \ \mu(A_0) && \text{[Box 1.1 (v)]} \\
&= \ 0, \quad \text{if } f_i \underset{\mu}{=} f_i^*,\ \forall\, i \in I.
\end{aligned}
$$

▷ **Solution 2-18**

$$
\forall\, \omega \in \Omega\colon \ \big(f(\omega) < g(\omega) \ \wedge\ g(\omega) = h(\omega)\big) \quad \Rightarrow \quad f(\omega) < h(\omega),
$$

which, by contraposition, is equivalent to

$$\forall\, \omega \in \Omega\colon\ \ f(\omega) \geq h(\omega) \quad \Rightarrow \quad \big(f(\omega) \geq g(\omega)\ \vee\ g(\omega) \neq h(\omega)\big).$$

Therefore,

$$\big\{\, \omega \in \Omega\colon f(\omega) \geq h(\omega)\big\} \subset \big\{\, \omega \in \Omega\colon f(\omega) \geq g(\omega)\big\} \cup \big\{\, \omega \in \Omega\colon g(\omega) \neq h(\omega)\big\}.$$

Now $\big(f \underset{\mu}{\leq} g \wedge g \underset{\mu}{=} h\big)$ implies

$$\begin{aligned}
&\mu\big(\big\{\, \omega \in \Omega\colon f(\omega) \geq h(\omega)\big\}\big) \\
&\leq \mu\big(\big\{\, \omega \in \Omega\colon f(\omega) \geq g(\omega)\big\}\big) + \mu\big(\big\{\, \omega \in \Omega\colon g(\omega) \neq h(\omega)\big\}\big) && \text{[Box 1.1 (vii)]} \\
&= 0 + 0 = 0. && [f \underset{\mu}{\leq} g,\ g \underset{\mu}{=} h]
\end{aligned}$$

Because a measure is nonnegative, this implies $\mu\big(\big\{\, \omega \in \Omega\colon f(\omega) \geq h(\omega)\big\}\big) = 0$, which is equivalent to $f \underset{\mu}{\leq} h$.

▷ **Solution 2-19** For all $A' \in \mathscr{A}'$,

$$\begin{aligned}
\mu_f(A') &= \mu[f^{-1}(A')] && \text{[(2.41)]} \\
&= \mu\left(\bigcup_{\substack{i=1,\ldots,n+1,\\ \alpha_i \in A'}} A_i\right) && \text{[Def. 2.16]} \\
&= \sum_{\substack{i=1,\ldots,n+1,\\ \alpha_i \in A'}} \mu(A_i) && \text{[Def. 1.43, (c)]} \\
&= \sum_{i=1}^{n+1} \mu(A_i) \cdot \delta_{\alpha_i}(A'). && \text{[(1.38)]}
\end{aligned}$$

Chapter 3
Integral

In the preceding chapters we introduced the most important concepts of measure theory related to the concepts of a measure and a measurable mapping. In this chapter we introduce the *integral* of measurable functions. This concept is fundamental also for probability theory, because the expectation of a numerical random variable with respect to a probability measure is the *integral* of a measurable function with respect to a probability measure. In chapter 6 we shall see that this also applies to variances, covariances, and correlations. We start defining the integral of a measurable function with respect to a measure μ. Then we study the most important rules of computation and other properties of integrals, introduce the concept of a *measure with density*, and treat the relationship between the Riemann integral and the integral with respect to the Lebesgue measure. The next section is on *absolute continuity* and the *Radon-Nikodym Theorem*. Both issues are crucial for conditional expectations (see ch. 10). A section on the integral with respect to a product measure concludes this chapter.

3.1 Definition

At first we define the integral for *nonnegative step functions*, then we extend the integral to *nonnegative measurable functions*, and finally we introduce the integral for *measurable functions* that may take on negative or nonnegative values.

3.1.1 Integral of a Nonnegative Step Function

In this subsection we introduce the integral of a *nonnegative step function*, also called *nonnegative simple function* or *elementary function*.

Nonnegative Step Function

Definition 3.1 (Nonnegative Step Function and Normal Representation) *Let $(\Omega, \mathscr{A})$ be a measurable space. Then $f: \Omega \to \mathbb{R}$ is called a nonnegative step function, if there is a finite sequence $A_1, \ldots, A_n \in \mathscr{A}$ and a finite sequence $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $\alpha_i \geq 0$, $i = 1, \ldots, n$, such that*

$$f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}. \tag{3.1}$$

If $A_1, \ldots, A_n \in \mathscr{A}$ are pairwise disjoint, then $f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$ is called a normal representation of f.

Remark 3.2 (Step Functions Take on Finitely Many Values) A nonnegative step function $f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$ is a measurable function $f\colon (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ taking on only a finite number of *nonnegative* values. These values are not necessarily $\alpha_1, \ldots, \alpha_n$. However, note:

(i) If $\mathscr{E} = \{A_1, \ldots, A_n\}$ is a partition of Ω, then $\alpha_1, \ldots, \alpha_n$ are the values of f.

(ii) If $A_1, \ldots, A_n$ are pairwise disjoint but $\mathscr{E}$ is not a partition of Ω, i. e., $A_{n+1} := \Omega \setminus \bigcup_{i=1}^{n} A_i \neq \emptyset$, then

$$f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i} + 0 \cdot \mathbb{1}_{A_{n+1}}.$$

This implies: $f(\omega) = 0$, for all $\omega \in A_{n+1}$.

(iii) If $A_1, \ldots, A_n$ are pairwise disjoint and additionally $\alpha_1, \ldots, \alpha_n$ are pairwise different and not 0, then $A_i = f^{-1}(\{\alpha_i\})$, $i = 1, \ldots, n$.

(iv) If $A_1, \ldots, A_n$ are pairwise disjoint, then, for all $\alpha_i \neq 0$,

$$f^{-1}(\{\alpha_i\}) = \bigcup_{j\colon \alpha_j = \alpha_i} A_j.$$

Hence in this case, the inverse image of the set $\{\alpha_i\}$ under f is the union of all sets A_j, $j \in \{1, \ldots, n\}$, for which $\alpha_j = \alpha_i$.

◁

Remark 3.3 (Different Representations of a Nonnegative Step Function) Note that nonnegative step functions can have different representations and also different normal representations (see Example 3.7). ◁

Example 3.4 (Indicator Function) Let $(\Omega, \mathscr{A})$ be a measurable space and $A \in \mathscr{A}$. The indicator function $\mathbb{1}_A$, which has already been introduced in Example 2.12, can also be written $1 \cdot \mathbb{1}_A + 0 \cdot \mathbb{1}_{A^c}$. Hence, because $A \in \mathscr{A}$ and 1 is a real number, $\mathbb{1}_A$ is a nonnegative step function. Note that $1 \cdot \mathbb{1}_A$ is also a normal representation of a nonnegative step function. ◁

Example 3.5 (Two Nonnegative Step Functions) In Example 2.55 we already presented two nonnegative step functions f and h and an illustrating Figure (see Fig. 2.6). The representations of both functions are normal. ◁

Example 3.6 (Tossing a Dice) Consider the set $\Omega = \{\omega_1, \ldots, \omega_6\}$ of possible outcomes of tossing a dice. Furthermore, let $\mathscr{A} = \mathscr{P}(\Omega)$ be the power set of Ω, and define $X\colon \Omega \to \mathbb{R}$ by

$$X(\omega_i) = i, \quad \forall \omega_i \in \Omega.$$

Hence, $X(\omega_i)$ is the *number of dots*. Considering the elements $\{\omega_1\}, \ldots, \{\omega_6\}$ of $\mathscr{A}$, and

$$X = \sum_{i=1}^{6} i \cdot 1_{\{\omega_i\}}$$

shows that X has a normal representation of a nonnegative step function. (For a related example see Exercise 3-1.) ◁

Example 3.7 (Several Representations of Nonnegative Step Functions) Consider the measurable space $(\mathbb{R}, \mathscr{B})$ and the nonnegative function $f: \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 2, \text{ if } x \in [0,1[\\ 5, \text{ if } x \in [1,2] \\ 4, \text{ if } x \in]2,3] \\ 1, \text{ if } x \in]3,4] \\ 0, \text{ otherwise.} \end{cases}$$

This function can also be represented by

$$\begin{aligned} f &= 2 \cdot 1_{[0,1[} + 5 \cdot 1_{[1,2]} + 4 \cdot 1_{]2,3]} + 1 \cdot 1_{]3,4]} \\ &= 2 \cdot 1_{[0,.5]} + 2 \cdot 1_{].5,1]} + 5 \cdot 1_{[1,2]} + 4 \cdot 1_{]2,3]} + 1 \cdot 1_{]3,4]} \\ &= 2 \cdot 1_{[0,2]} + 3 \cdot 1_{[1,3]} + 1 \cdot 1_{[2,4]} \\ &= 1 \cdot 1_{[0,4]} + 1 \cdot 1_{[0,3]} + 2 \cdot 1_{[1,3]} + 1 \cdot 1_{[1,2]}. \end{aligned} \tag{3.2}$$

The first two representations are normal, the latter two are nonnormal representations of f. ◁

Remark 3.8 (Existence of a Normal Representation) For every nonnegative step function there exists a normal representation (see Exercise 3-2).

If $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ is a normal representation of a nonnegative step function, then there may be another sequence $C_1, \ldots, C_m$ of pairwise disjoint elements of $\mathscr{A}$ and another sequence $\gamma_1, \ldots, \gamma_m$ of nonnegative real numbers such that

$$f = \sum_{i=1}^{n} \alpha_i 1_{A_i} = \sum_{i=1}^{m} \gamma_i 1_{C_i}.$$

Both sum terms are normal representations. The first two representations of f in Equation (3.2) provide an example. ◁

Integral of a Nonnegative Step Function

The following uniqueness property holds for two normal representations of a nonnegative step function:

Lemma 3.9 (A Uniqueness Property)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space. If $f : \Omega \to \mathbb{R}$ is a nonnegative step function and $f = \sum_{i=1}^{n} \alpha_i 1_{A_i} = \sum_{i=1}^{m} \gamma_i 1_{C_i}$ are two normal representations, then

$$\sum_{i=1}^{n} \alpha_i \, \mu(A_i) \; = \; \sum_{i=1}^{m} \gamma_i \, \mu(C_i). \tag{3.3}$$

For a proof see Klenke (2008, Lemma 4.1, p. 85). Note, by convention $0 \cdot \infty = 0$.

According to this lemma, the number $\sum_{i=1}^{n} \alpha_i \, \mu(A_i)$ assigned to a nonnegative step function f does not depend on the specific normal representation of f (for an illustration see Exercise 3-3). This property allows us to define the *integral* of a nonnegative step function with respect to a measure μ as follows:

Definition 3.10 (Integral of a Nonnegative Step Function)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and let $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ be a normal representation of a nonnegative step function $f : \Omega \to \mathbb{R}$. Then the number

$$\int f \, d\mu \; = \sum_{i=1}^{n} \alpha_i \, \mu(A_i) \tag{3.4}$$

is called the integral of f (over Ω) with respect to μ.

Remark 3.11 (Integral of a Constant) Let $(\Omega, \mathscr{A}, \mu)$ be a measure space. If $f = \alpha$, $\alpha \in \mathbb{R}$, then Equation (3.4) immediately implies

$$\int \alpha \, d\mu \; = \; \alpha \cdot \mu(\Omega). \tag{3.5}$$

◁

Remark 3.12 (Integral Over a Subset of Ω) Let $A \in \mathscr{A}$. If $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ is a normal representation of a nonnegative step function, then the product $1_A \cdot f$ is a nonnegative step function as well and can be written

$$1_A \cdot f \; = \sum_{i=1}^{n} \alpha_i \, 1_{A \cap A_i}, \tag{3.6}$$

which is a normal representation of $1_A \cdot f$ (see Fig. 3.1 and Exercise 3-4). Hence, we may also consider the integral $\int 1_A \cdot f \, d\mu$ and define the *integral of f over a subset A* of Ω by

$$\int_A f \, d\mu \; = \int 1_A \cdot f \, d\mu. \tag{3.7}$$

◁

The following corollary is a special case of Equation (3.7) if $f = \alpha$, $\alpha \in \mathbb{R}$. (For proof see Exercise 3-5).

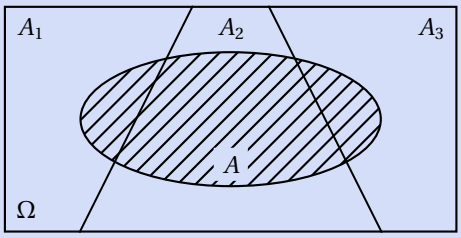


Figure 3.1. A partition and a subset of Ω

Corollary 3.13 (Constants)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and $\alpha \in \mathbb{R}$. If $A \in \mathscr{A}$, then

$$\int_A \alpha \, d\mu = \alpha \, \mu(A). \tag{3.8}$$

Examples

Example 3.14 (Indicator Function) Consider a measure space $(\Omega, \mathscr{A}, \mu)$ and the indicator 1_A of $A \in \mathscr{A}$. Then $1_A = 1 \cdot 1_A$ is a normal representation of 1_A. Therefore,

$$\int_A d\mu = \int 1_A \, d\mu = 1 \cdot \mu(A) = \mu(A). \tag{3.9}$$

◁

Example 3.15 (Nonnegative Step Function and Dirac Measure) Let $(\Omega, \mathscr{A})$ be a measurable space, and for $\omega \in \Omega$ let δ_ω denote the *Dirac measure* at ω (see Example 1.52). Furthermore, consider a normal representation $f = \sum_{i=1}^m \alpha_i 1_{A_i}$ of a nonnegative step function. Its integral with respect to the Dirac measure is

$$\int f \, d\delta_\omega = \sum_{i=1}^m \alpha_i \, \delta_\omega(A_i) = \sum_{i=1}^m \alpha_i \, 1_{A_i}(\omega) = f(\omega). \tag{3.10}$$

According to this equation, the integral of a nonnegative step function f with respect to the Dirac measure at ω is the value of f for the argument ω. Furthermore, if $f = 1_A$ is the indicator of $A \in \mathscr{A}$, then

$$\int 1_A \, d\delta_\omega = 1_A(\omega). \tag{3.11}$$

Hence, in this special case, the integral is the value of the indicator 1_A for the argument ω. ◁

Example 3.16 (Nonnegative Step Function and Counting Measure) Suppose $\Omega = \{1, \ldots, n\},\ n \in \mathbb{N}$. For the measurable space $\big(\Omega, \mathscr{P}(\Omega)\big)$, the counting measure $\mu_\#$ on the power set $\mathscr{P}(\Omega)$ is defined by

$$\mu_\#(A) = \sum_{\omega=1}^{n} 1_A(\omega), \quad \forall A \subset \Omega, \tag{3.12}$$

(see Example 1.54). Hence, $\mu_\#(A)$ is simply the number of elements, i. e., the cardinality of A. Now consider a nonnegative step function with normal representation $f = \sum_{i=1}^{m} \alpha_i 1_{A_i}$. According to Equations (3.4) and (3.12), its integral with respect to the counting measure is

$$\begin{aligned} \int f \, d\mu_\# &= \sum_{i=1}^{m} \alpha_i \, \mu_\#(A_i) = \sum_{i=1}^{m} \alpha_i \sum_{\omega=1}^{n} 1_{A_i}(\omega) = \sum_{\omega=1}^{n} \sum_{i=1}^{m} \alpha_i 1_{A_i}(\omega) \\ &= \sum_{\omega=1}^{n} f(\omega). \end{aligned} \tag{3.13}$$

Hence, the integral of a nonnegative step function f with respect to the counting measure $\mu_\#$ is the sum over all values of f (see also Exercise 3-6). Using Equations (1.40) and (3.10), this integral can also be written

$$\int f \, d\mu_\# = \int f \, d\Big(\sum_{\omega=1}^{n} \delta_\omega \Big) = \sum_{\omega=1}^{n} \int f \, d\delta_\omega. \tag{3.14}$$

◁

3.1.2 Integral of a Nonnegative Measurable Function

In this section we extend the concept of an integral to nonnegative measurable functions. Before introducing the definition we consider a theorem according to which every nonnegative measurable function can be represented as a limit of an increasing sequence of nonnegative step functions. We begin with an example.

Example 3.17 (Increasing Sequence of Nonnegative Step Functions) Consider the measurable space $(\mathbb{R}, \mathscr{B})$ and the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 - x^2, & \forall x \in [0,1], \\ 0, & \text{otherwise.} \end{cases} \tag{3.15}$$

Now we construct three functions $f_i : \mathbb{R} \to \mathbb{R},\ i = 1,2,3$, with $f_1 \le f_2 \le f_3 \le f$ that approximate f (see Fig. 3.2). Let us start with f_1 defined by

$$f_1(x) = \begin{cases} \alpha_1 = .50, & \text{if } x \in A_1 = \big[0, (1-.50)^{1/2}\big], \\ \alpha_2 = 0, & \text{if } x \in A_1^c, \end{cases}$$

where $\big[0, (1-.50)^{1/2}\big]$ denotes the closed interval between 0 and $(1-.50)^{1/2} \approx .707$. Because A_1 is an element of $\mathscr{B}$ and .50 is a nonnegative real number, $f_1 = \alpha_1 1_{A_1}$ is a nonnegative step function. Next consider f_2 defined by

$$f_2(x) = \begin{cases} \beta_1 = .75, & \text{if } x \in B_1 = \left[0, (1-.75)^{1/2}\right], \\ \beta_2 = .50, & \text{if } x \in B_2 = \left]\,(1-.75)^{1/2}, (1-.50)^{1/2}\right], \\ \beta_3 = .25, & \text{if } x \in B_3 = \left]\,(1-.50)^{1/2}, (1-.25)^{1/2}\right], \\ \beta_4 = 0, & \text{if } x \in (B_1 \cup B_2 \cup B_3)^c. \end{cases}$$

Because B_1, B_2, B_3 are elements of $\mathscr{B}$ and .75, .50, .25 are nonnegative real numbers, $f_2 = \sum_{i=1}^{3} \beta_i \mathbb{1}_{B_i}$ is a nonnegative step function. Finally, let f_3 be defined by

$$f_3(x) = \begin{cases} \gamma_1 = .875, & \text{if } x \in C_1 = \left[0, (1-.875)^{1/2}\right], \\ \gamma_2 = .750, & \text{if } x \in C_2 = \left]\,(1-.875)^{1/2}, (1-.750)^{1/2}\right], \\ \gamma_3 = .625, & \text{if } x \in C_3 = \left]\,(1-.750)^{1/2}, (1-.625)^{1/2}\right], \\ \gamma_4 = .500, & \text{if } x \in C_4 = \left]\,(1-.625)^{1/2}, (1-.500)^{1/2}\right], \\ \gamma_5 = .375, & \text{if } x \in C_5 = \left]\,(1-.500)^{1/2}, (1-.375)^{1/2}\right], \\ \gamma_6 = .250, & \text{if } x \in C_6 = \left]\,(1-.375)^{1/2}, (1-.250)^{1/2}\right], \\ \gamma_7 = .125, & \text{if } x \in C_7 = \left]\,(1-.250)^{1/2}, (1-.125)^{1/2}\right], \\ \gamma_8 = 0, & \text{if } x \in (C_1 \cup \ldots \cup C_7)^c. \end{cases}$$

Again, $C_1, \ldots, C_7$ is a sequence of elements of $\mathscr{B}$ and .875, .750, .625, .500, .375, .250, .125 is a sequence of nonnegative real numbers. Therefore, $f_3 = \sum_{i=1}^{7} \gamma_i \mathbb{1}_{C_i}$ is a nonnegative step function. The integral of the functions f_1 and f_2 are computed in Exercise 3-7. ◁

Convergence of an Increasing Sequence of Nonnegative Step Functions

Example 3.18 (Convergence) Figure 3.2 shows that $f_1(\omega) \le f_2(\omega) \le f_3(\omega) \le f(\omega)$ for all $\omega \in \Omega$. Hence, f_1, f_2, f_3 is a *finite increasing sequence* of nonnegative step functions. The interval $[0,1]$ on the vertical axis is partitioned and these partitions are refined step by step. In our example, we started with the partition $\{[0,.50[, [.50,1]\}$. Then we partitioned

$$[0,.50[\text{ to } \{[0,.25[,\ [.25,.50[\} \quad \text{and} \quad [.50,1] \text{ to } \{[.50,.75[,\ [.75,1]\}, \quad \text{etc.}$$

Following this idea, we can define functions $f_4, f_5, \ldots$ such that $f_1, f_2, \ldots$ is an infinite sequence of nonnegative step functions with $f_1(\omega) \le f_2(\omega) \le \ldots \le f(\omega)$, for all $\omega \in \Omega$, and $\lim_{n\to\infty} f_n(\omega) = f(\omega)$, for all $\omega \in \Omega$. According to the following theorem, this holds for *all* nonnegative measurable functions $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$. ◁

Theorem 3.19 (Approximation of Nonnegative Functions)
Let $(\Omega, \mathscr{A})$ be a measurable space and $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ a nonnegative measurable function. Then:

(i) There is a sequence $f_1, f_2, \ldots$ of nonnegative step functions such that

$$f_1(\omega) \le f_2(\omega) \le \ldots, \quad \forall\, \omega \in \Omega \tag{3.16}$$

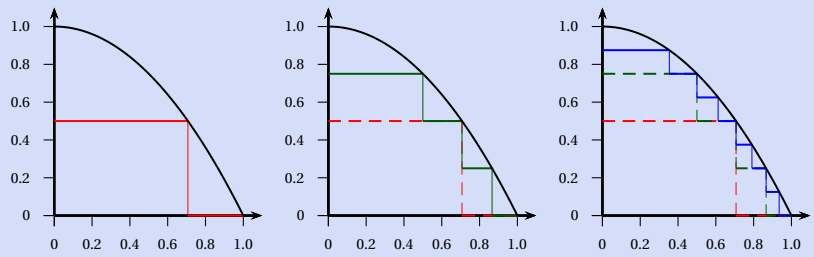


Figure 3.2. Increasing nonnegative step functions

and

$$\lim_{n\to\infty} f_n(\omega) = f(\omega), \quad \forall\, \omega \in \Omega. \tag{3.17}$$

(ii) There is a sequence of sets $A_1, A_2, \ldots \in \mathscr{A}$ and a sequence of nonnegative real numbers $\alpha_1, \alpha_2, \ldots$ such that

$$f = \sum_{i=1}^{\infty} \alpha_i 1_{A_i}. \tag{3.18}$$

For a proof see Klenke (2008, Theorem 1.96, p. 41).

Remark 3.20 (Infinite Sums) Equation (3.18) can be visualized by Figure 3.2. The function f_3 on the right-hand side of this figure is already close to f. Partitioning the intervals *on the vertical axis* again and again leads to better approximations of f. Note that the horizontal axis does not have to be a subset of $\mathbb{R}$; instead, it can be any nonempty set Ω.

Remember, the right-hand side of Equation (3.18) is just a symbol for the corresponding limit, i. e.,

$$\sum_{i=1}^{\infty} \alpha_i 1_{A_i} = \lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i 1_{A_i}. \tag{3.19}$$

Note that, for $\alpha_i \geq 0$, this limit always exists. ◁

Before turning to the definition of the integral of a nonnegative measurable function let us use the properties (3.16) and (3.17) to define the concepts *increasing sequence of nonnegative step functions* and *pointwise convergence*.

Definition 3.21 (Increasing Sequence of Nonnegative Step Functions) *Let $(\Omega, \mathscr{A})$ be a measurable space and $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ a nonnegative measurable function. A sequence $f_1, f_2, \ldots$ of nonnegative step functions satisfying*

(3.16) *is called* $increasing$*. If it also satisfies* (3.17)*, then we say that* $f_1, f_2, \ldots$ $converges\ pointwise\ to\ f$ *and denote it by* $f_n \uparrow f$*.*

Uniqueness of the Limits of an Integral

In Theorem 3.19 we have seen that every nonnegative measurable function f can be represented by the limit $\lim_{n\to\infty} f_n$ of an increasing sequence $f_1, f_2, \ldots$ of nonnegative step functions, i. e.,

$$f = \lim_{n\to\infty} f_n. \tag{3.20}$$

The definition of the integral of nonnegative step functions implies that the *integrals* of the functions f_n are increasing as well, i. e.,

$$f_n \le f_{n+1} \quad \Rightarrow \quad \int f_n \, d\mu \le \int f_{n+1} \, d\mu, \quad \forall n \in \mathbb{N},$$

[see Bauer, 2001, proposition (10.7), p. 55]. Hence, the sequence of the integrals either converges to a (finite) real number or diverges to $+\infty$.

In Figure 3.2 we presented the first three nonnegative step functions f_1, f_2, and f_3 of such an increasing sequence $f_1, f_2, \ldots$ that approximates the function $f : \mathbb{R} \to \mathbb{R}$ defined by Equation (3.15). Figure 3.3 visualizes the convergence of the integrals $\int f_n \, d\lambda$ with respect to the Lebesgue measure λ on $(\mathbb{R}, \mathscr{B})$ (see the shaded areas in Figure 3.3).

It should be noted, however, that there is not only one single increasing sequence of nonnegative step functions whose limit is f. This is illustrated in the following example.

Example 3.22 (Uniqueness) For an example, construct a sequence $g_1, g_2, \ldots$ analogously to the sequence $f_1, f_2, \ldots$ in Example 3.17, using other partitions of the interval $[0, 1]$ on the vertical axis, e. g.,

$$\big\{[0, .40[,\ [.40, 1]\big\} \quad \text{and} \quad \big\{[0, .20[,\ [.20, .40[,\ [.40, .80[,\ [.80, 1]\big\}, \quad \text{etc.}$$

Then $g_1, g_2, \ldots$ is a second increasing sequence that also approximates f. Figure 3.3 suggests that the specific choice of an increasing sequence of nonnegative step functions is irrelevant for the limit of their integrals. And in fact, according to the following theorem this does not only apply to our example and to the Lebesgue measure λ on $\mathscr{B}$, but to *any* nonnegative measurable function and *any* measure μ. ◁

Theorem 3.23 (Uniqueness of the Limits of Integrals)
If $f_1, f_2, \ldots$ *and* $g_1, g_2, \ldots$ *are two increasing sequences of nonnegative step functions* $f_n, g_n : (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$*, then* $\lim_{n\to\infty} f_n = \lim_{n\to\infty} g_n$ *implies*

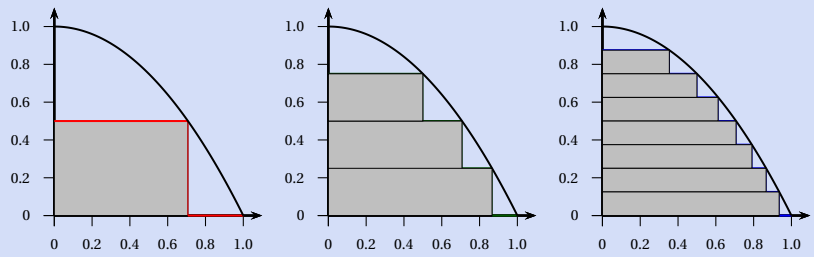


Figure 3.3. Integral of nonnegative step functions w.r.t. the Lebesgue measure

$$\lim_{n\to\infty} \int f_n \, d\mu = \lim_{n\to\infty} \int g_n \, d\mu. \tag{3.21}$$

For a proof see Bauer (2001, Corollary 11.2, p. 58).

According to this theorem, if we consider two increasing sequences of nonnegative step functions with identical limits, then we know that the limits of their integrals are identical.

Definition of the Integral of a Nonnegative Measurable Function

Based on the result of Theorem 3.23, we define the integral of any nonnegative measurable function.

Definition 3.24 (Integral of a Nonnegative Measurable Function) *Assume that $f\colon (\Omega, \mathcal{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ is a nonnegative measurable function and let $f = \lim_{n\to\infty} f_n$ be a representation of f as the limit of an increasing sequence $f_1, f_2, \ldots$ of nonnegative step functions. Then*

$$\int f \, d\mu := \lim_{n\to\infty} \int f_n \, d\mu \tag{3.22}$$

is called the integral of f (over Ω) with respect to μ.

Note that the integral of a nonnegative measurable function is either a nonnegative real number or $+\infty$.

Example 3.25 (Integral With Respect to a Dirac Measure) Suppose the assumptions of Definition 3.24 hold. Then, for $\omega \in \Omega$,

$$\begin{aligned} \int f \, d\delta_\omega &= \lim_{n\to\infty} \int f_n \, d\delta_\omega && [(3.22)] \\ &= \lim_{n\to\infty} f_n(\omega) && [(3.10)] \\ &= f(\omega). && [(3.20)] \end{aligned} \tag{3.23}$$

Hence, the integral of a nonnegative measurable function f with respect to the Dirac measure at ω is the value of f for ω. ◁

We conclude this section by the following lemma on monotonicity of the integrals of nonnegative measurable functions.

Lemma 3.26 (Monotonicity)
If $f, g\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ are nonnegative and measurable, then

$$f \le g \quad \Rightarrow \quad \int f \, d\mu \le \int g \, d\mu. \tag{3.24}$$

For a proof see Bauer (2001, p. 59, Eq. (11.8)).

Remark 3.27 (Bounds of the Integral of a Bounded Function) Let $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be nonnegative and measurable and $\alpha \in \mathbb{R}$. Then, for $g = \alpha$, Equations (3.24) and (3.5) imply

$$f \le \alpha \quad \Rightarrow \quad \int f \, d\mu \le \alpha \cdot \mu(\Omega), \tag{3.25}$$

and

$$f \ge \alpha \quad \Rightarrow \quad \int f \, d\mu \ge \alpha \cdot \mu(\Omega). \tag{3.26}$$

◁

3.1.3 Integral of a Measurable Function

Now we define the integral of a measurable function $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ using the positive part f^+ and the negative part f^- of f that have been introduced in section 2.4.2. According to Theorem 2.66, f^+ and f^- are both nonnegative measurable functions. Reading the following definition, remember the conventions: $\infty + \infty = \infty$, $-\infty - \infty = -\infty$, $x + \infty = \infty$, for all $x \in \mathbb{R}$, $x - \infty = -\infty$, for all $x \in \mathbb{R}$. Also note that $\infty - \infty$ is *not defined*, which has to be observed whenever integrals are not necessarily finite.

Definition 3.28 (Integral of a Measurable Function)
Let $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a measurable function. If $\int f^+ \, d\mu$ or $\int f^- \, d\mu$

are finite, then f is called quasi-integrable with respect to μ, or simply quasi-μ-integrable, and

$$\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu \tag{3.27}$$

is called the integral of f (over Ω) with respect to μ. If $\int f^+\,d\mu$ and $\int f^-\,d\mu$ are both finite, then f is called integrable with respect to μ, or simply μ-integrable.

Remark 3.29 (Integrability and Quasi-Integrability) Of course, every integrable measurable function is quasi-integrable and each nonnegative function is also quasi-integrable. Furthermore, assuming that a function $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is integrable or quasi-integrable includes the assumption that f is measurable. Finally, if f is μ-integrable, then

$$-\infty < \int f\,d\mu < +\infty,$$

i. e., the integral is *finite*, taking a value in $\mathbb{R}$. If f is quasi-μ-integrable, then the integral may also be *infinite*, i. e., it may also take on the values $+\infty$ or $-\infty$. ◁

Remark 3.30 (A Standard Method for Proofs) The integral of a quasi-integrable function has been defined in three steps, for *nonnegative measurable step functions*, for *nonnegative measurable functions*, and for *quasi-integrable functions*. Oftentimes, these steps are also followed in proofs of propositions involving integrals. That is, in a first step it is shown that the proposition holds for nonnegative measurable step functions. In a second step, using Equation (3.22), it is proven for nonnegative measurable functions, and finally, Equation (3.27) is applied to complete the proof for all quasi-integrable functions. An example is the proof of Theorem 3.36. Oftentimes, we only detail the first step, in particular, if the remaining two steps are straightforward. ◁

Example 3.31 (Integral With Respect to the Lebesgue Measure λ) Figure 3.4 displays the integral of a function $f\colon (\mathbb{R}, \mathscr{B}, \lambda) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ with respect to the Lebesgue measure λ. Because f^+ and f^- are both nonnegative (see Rem. 2.62), the integrals $\int f^+\,d\mu$ and $\int f^-\,d\mu$ are positive and identical to the white areas marked + and − in Figure 3.4. According to Equation (3.27), the integral of f is the difference between the area $\int f^+\,d\mu$ and the area $\int f^-\,d\mu$. ◁

Remark 3.32 (An Alternative Notation) An alternative notation for the integral of f is

$$\int f\,d\mu = \int f(\omega)\,\mu(d\omega) = \int_\Omega f(\omega)\,\mu(d\omega), \tag{3.28}$$

which explicitly uses the values $f(\omega)$ of f. This notation conveys the idea that the values $f(\omega)$ of f are weighted by the measure of $d\omega$. If $\Omega = \mathbb{R}$, then $d\omega$ symbolizes

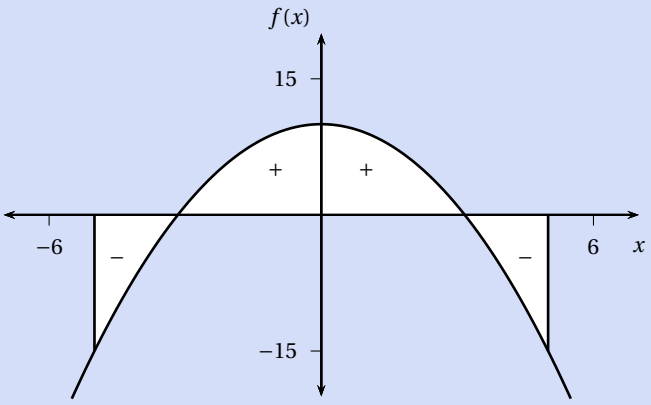


Figure 3.4. Lebesgue integral of a function from −5 to 5

the length of an infinitesimal interval between two elements in $\mathbb{R}$. If Ω is finite or countable, then $\mu(d\omega)$ symbolizes the value of μ for the singleton $\{\omega\}$ and the integral can be written as a sum (see Example 3.16). ◁

Lemma 3.33 (Integrability Carries Over to Restrictions of Functions)
(i) If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is quasi-μ-integrable and $A \in \mathscr{A}$, then $1_A f$ is quasi-μ-integrable.
(ii) If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is μ-integrable and $A \in \mathscr{A}$, then $1_A f$ is μ-integrable.

(Proof p. 111)

Remark 3.34 (Integral of $1_A f$) Lemma 3.33 (ii) means: If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is μ-integrable and $A \in \mathscr{A}$, then

$$\int f \, d\mu \text{ is finite} \quad \Rightarrow \quad \int 1_A f \, d\mu \text{ is finite.} \tag{3.29}$$

◁

If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is quasi-μ-integrable and $A \in \mathscr{A}$, then Lemma 3.33 implies that the integral $\int 1_A f \, d\mu$ is well-defined. Hence, we can now introduce the integral of f over a subset A of Ω as follows:

Definition 3.35 (Integral Over a Subset A of Ω)
If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is quasi-μ-integrable and $A \in \mathscr{A}$, then

$$\int_A f \, d\mu := \int 1_A \, f \, d\mu \tag{3.30}$$

is called the integral of f over A with respect to μ.

Because $1_\Omega \, f = f$, a special case of Equation (3.30) is

$$\int_\Omega f \, d\mu = \int 1_\Omega \, f \, d\mu = \int f \, d\mu. \tag{3.31}$$

3.2 Properties

In this section we consider some important properties and rules of computation for the integral of a measurable function $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$.

Theorem 3.36 (Linearity)
Consider the functions $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ and $g : (\Omega, \mathscr{A}, \mu) \to (\mathbb{R}, \mathscr{B})$.

(i) If f is quasi-μ-integrable and $\alpha \in \mathbb{R}$, then αf is quasi-μ-integrable and

$$\int \alpha f \, d\mu = \alpha \int f \, d\mu. \tag{3.32}$$

(ii) If f is quasi-μ-integrable and g is μ-integrable, then $f + g$ is quasi-μ-integrable, and

$$\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu. \tag{3.33}$$

(Proof p. 111)

Combining propositions (ii) and (ii) of Theorem 3.36 immediately yields the following corollary.

Corollary 3.37 (Linearity)
Consider the functions $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, $g\colon (\Omega, \mathscr{A}, \mu) \to (\mathbb{R}, \mathscr{B})$ and let $\alpha, \beta \in \mathbb{R}$. If f is quasi-μ-integrable and g is μ-integrable, then $\alpha f + \beta g$ is quasi-μ-integrable, and

$$\int (\alpha f + \beta g) \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu. \tag{3.34}$$

Linearity can also be used to prove the following corollary on the equivalence of integrability of a measurable function f and finiteness of the integral of the absolute value function $|f|$.

Corollary 3.38 (Integrability and Absolute Value Function)
The function $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is μ-integrable if and only if

$$\int |f| \, d\mu < \infty.$$

(Proof p. 115)

Example 3.39 (Integral Over the Union of Two Sets) If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is μ-integrable and $A, B \in \mathscr{A}$, then $1_{A\cup B} f$ is μ-integrable and

$$\int_{A\cup B} f \, d\mu = \int 1_{A\cup B} f \, d\mu = \int_A f \, d\mu + \int_B f \, d\mu - \int_{A\cap B} f \, d\mu. \tag{3.35}$$

If $A \cap B = \emptyset$ and f is quasi-μ-integrable, then

$$\int_{A\cup B} f \, d\mu = \int 1_{A\cup B} f \, d\mu = \int_A f \, d\mu + \int_B f \, d\mu. \tag{3.36}$$

(see Exercise 3-8). ◁

Lemma 3.40 (Measures That are Identical on a Sub-σ-Algebra)
Assume that $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is nonnegative or μ-integrable. Furthermore, let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, let f be $\mathscr{C}$-measurable, and suppose that $\nu(A) = \mu(A)$, for all $A \in \mathscr{C}$. Then $(\Omega, \mathscr{C}, \nu)$ is a measure space and

$$\int f \, d\nu = \int f \, d\mu. \tag{3.37}$$

(Proof p. 115)

Hence, the integral $\int f \, d\mu$ only depends on the values of μ on the σ-algebra $\sigma(f)$, the σ-algebra generated by f.

Lemma 3.41 (Integrable Functions are μ-a.e. Real-Valued)
Let $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable. If f is μ-integrable, then f is real-valued μ-almost everywhere. If f is quasi-μ-integrable, then

$$\mu\big(\{\omega \in \Omega\colon f(\omega) = \infty\}\big) > 0 \quad \Rightarrow \quad \int f \, d\mu = \infty, \tag{3.38}$$

$$\mu\big(\{\omega \in \Omega\colon f(\omega) = -\infty\}\big) > 0 \quad \Rightarrow \quad \int f \, d\mu = -\infty. \tag{3.39}$$

(Proof p. 115)

Remark 3.42 (Integrable Functions are Assumed to be Real-Valued) Contraposition of (3.38) and (3.39) yields: If $\int f\,d\mu$ is finite, then $f(\omega) \in \mathbb{R}$ (i. e., $-\infty < f(\omega) < \infty$), for μ-almost all $\omega \in \Omega$ (see Def. 2.68 and Remark 2.70). In this case, there is a *real-valued* measurable function $f^*\colon (\Omega, \mathscr{A}, \mu) \to (\mathbb{R}, \mathscr{B})$ with $f^* \underset{\mu}{=} f$. (For example, define $f^* := 1_A \cdot f + 1_{A^c} \cdot 0 = 1_A \cdot f$ for $A := \{\omega \in \Omega\colon f(\omega) \in \mathbb{R}\}$). Therefore, without substantial loss of generality, for simplicity, we often assume that a function is real-valued if it has a finite integral. ◁

3.2.1 Integral of μ-Equivalent Functions

The concept of equivalence of two measurable functions with respect to a measure has already been introduced in section 2.5. Now we treat the relationship of this concept to the integrals of two numerical functions.

Theorem 3.43 (A Condition Equivalent to $f \underset{\mu}{=} 0$)
If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a nonnegative measurable function, then

$$\int f\,d\mu = 0 \quad \Leftrightarrow \quad f \underset{\mu}{=} 0. \tag{3.40}$$

For a proof see Bauer (2001, Theorem 13.2, p. 71).

Lemma 3.44 (Integral of a Positive Function)
If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is quasi-μ-integrable and there is an $A \in \mathscr{A}$ such that $\mu(A) > 0$ and $f(\omega) > 0$, for all $\omega \in A$, then

$$\int 1_A \cdot f d\mu > 0. \tag{3.41}$$

(Proof p. 116)

If $(\Omega, \mathscr{A}, \mu)$ is a measure space, then a set $A \in \mathscr{A}$ with $\mu(A) = 0$ is called a *null set* with respect to μ. In the following lemma we consider the integral over such a null set (see Exercise 3-9).

Lemma 3.45 (Integral Over a Null Set)
Let $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable. If $A \in \mathscr{A}$, with $\mu(A) = 0$, then $1_A \cdot f$ is μ-integrable and

$$\int_A f\,d\mu = \int 1_A \cdot f\,d\mu = 0. \tag{3.42}$$

(Proof p. 116)

Remark 3.46 (Integration Over Null Sets Can be Neglected) The conjunction of Equations (3.36) and (3.42) implies: If f is quasi-μ-integrable and $A \in \mathscr{A}$ with $\mu(A) = 0$, then

$$\int f \, d\mu = \int_{\Omega} f \, d\mu = \int_{\Omega \setminus A} f \, d\mu + \int_{A} f \, d\mu = \int_{\Omega \setminus A} f \, d\mu. \tag{3.43}$$

◁

Lemma 3.47 (Integrals of μ-Equivalent Functions)
Let $f, g \colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be quasi-μ-integrable. Then

$$f \underset{\mu}{=} g \quad \Rightarrow \quad \int f \, d\mu = \int g \, d\mu. \tag{3.44}$$

(Proof p. 116)

The following theorem presents a condition that is necessary and sufficient for μ-equivalence of f and g.

Theorem 3.48 (Identity of Integrals of μ-Equivalent Functions)
If $f, g : (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ are μ-integrable, then

$$f \underset{\mu}{=} g \quad \Leftrightarrow \quad \int_A f \, d\mu = \int_A g \, d\mu, \quad \forall A \in \mathscr{A}. \tag{3.45}$$

(Proof p. 117)

In section 3.4 we shall see that, if f and g are μ-integrable and nonnegative, then it is sufficient to consider the integrals over all sets A in a $\cap$-stable generating system of $\mathscr{A}$ in order to show μ-equivalence of f and g (see Th. 3.68).

Remark 3.49 (A Counter-Example) Note that Equation (3.45) does not hold if f, g are nonnegative but not μ-integrable measurable functions. This is exemplified as follows: Consider $f, g : (\mathbb{R}, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, where $\mathscr{A} = \{\mathbb{R}, \emptyset\}$, $\mu(\mathbb{R}) = \infty$, $f = 1$, and $g = 2$. Then $\int_{\mathbb{R}} f \, d\mu = \int_{\mathbb{R}} g \, d\mu = \infty$ and $\int_{\emptyset} f \, d\mu = \int_{\emptyset} g \, d\mu = 0$. Hence, $\int_A f \, d\mu = \int_A g \, d\mu$, for all $A \in \mathscr{A}$, but f and g are *not equivalent* with respect to μ.
◁

Remark 3.50 (Some Special Cases) Theorem 3.48 implies: If $f : (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a measurable function, then $f \underset{\mu}{=} \alpha$, $\alpha \in \mathbb{R}$, is equivalent to

$$\int_A f \, d\mu = \int_A \alpha \, d\mu = \int 1_A \alpha \, d\mu = \alpha \cdot \int 1_A \, d\mu = \alpha \cdot \mu(A), \quad \forall A \in \mathscr{A}. \tag{3.46}$$

Furthermore, $f \underset{\mu}{=} 0$ is equivalent to

$$\int_A f \, d\mu = 0, \quad \forall A \in \mathscr{A}, \tag{3.47}$$

using the convention $0 \cdot \infty = 0$, if necessary.

An immediate implication of Equation (3.46) for $A = \Omega$ is

$$f \underset{\mu}{=} \alpha, \quad \alpha \in \mathbb{R} \quad \Rightarrow \quad \int f \, d\mu = \alpha \, \mu(\Omega). \tag{3.48}$$

For $\alpha = 0$, this yields

$$f \underset{\mu}{=} 0 \quad \Rightarrow \quad \int f \, d\mu = 0. \tag{3.49}$$

◁

Remark 3.51 (Almost Everywhere) The notion of μ-equivalence of f and g is an example of a property that holds for all $\omega \in \Omega \setminus A$ with $\mu(A) = 0$. We also say that such a property holds *μ-almost everywhere* (μ-a.e.). Another example is the property

$$f(\omega) \leq g(\omega), \quad \forall \omega \in \Omega \setminus A \text{ and } \mu(A) = 0,$$

which is denoted by $f \underset{\mu}{\leq} g$. ◁

The proposition of Lemma 3.47 analogously holds for the relations $\underset{\mu}{\leqq}$ and $\underset{\mu}{<}$. The following theorem generalizes Lemma 3.26.

Theorem 3.52 (Monotonicity)
Let $f, g : (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable functions.

(i) If f and g are quasi-μ-integrable, then

$$f \underset{\mu}{\leqq} g \quad \Rightarrow \quad \int f \, d\mu \leq \int g \, d\mu. \qquad (\textit{monotonicity}) \tag{3.50}$$

(ii) If $\mu(\Omega) > 0$ and f, g are μ-integrable, then

$$f \underset{\mu}{<} g \quad \Rightarrow \quad \int f \, d\mu < \int g \, d\mu. \qquad (\textit{strict monotonicity}) \tag{3.51}$$

(Proof p. 117)

3.2.2 Integral With Respect to a Weighted Sum of Measures

In Example 1.61 we already noted that a weighted sum of measures with nonnegative weights is again a measure. As a special case, if μ is a measure on $(\Omega, \mathscr{A})$ and α is a nonnegative number, then $\alpha \cdot \mu$ is a measure on $(\Omega, \mathscr{A})$ as well. Furthermore, if f is μ-integrable, then

$$\int f \, d(\alpha\mu) = \int \alpha f \, d\mu = \alpha \int f \, d\mu \tag{3.52}$$

(see Exercise 3-10). This is generalized in the following theorem.

Theorem 3.53 (Integral With Respect to a Weighted Sum of Measures) *If $f : (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is measurable and nonnegative, $\mu_1, \mu_2, \ldots$ are measures on $(\Omega, \mathscr{A})$, and $\alpha_1, \alpha_2, \ldots \in \mathbb{R}$ are nonnegative, then*

$$\int f \, d\Big(\sum_{i=1}^{\infty} \alpha_i \mu_i\Big) = \sum_{i=1}^{\infty} \alpha_i \int f \, d\mu_i. \tag{3.53}$$

For a proof, see Equation (3.52) and Bauer (2001, Example 3 on page 61).

If we consider a *finite* weighted sum of measures, the assumption that f is nonnegative can be replaced by integrability of f. In the following theorem we consider a weighted sum of two measures. In Remark 3.55 we extend the result to a finite weighted sum of measures.

Theorem 3.54 (Integral With Respect to a Weighted Sum of two Measures) *Let μ_1, μ_2 be measures on $(\Omega, \mathscr{A})$. If $f : (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is integrable with respect to μ_1 and μ_2, and $0 \le \alpha_1, \alpha_2 \in \mathbb{R}$, then f is integrable with respect to $\alpha_1 \mu_1 + \alpha_2 \mu_2$, and*

$$\int f \, d(\alpha_1 \mu_1 + \alpha_2 \mu_2) = \alpha_1 \int f \, d\mu_1 + \alpha_2 \int f \, d\mu_2. \tag{3.54}$$

For a proof, see Equation (3.52) and Bauer (2001, Example 5 on page 67).

Remark 3.55 (Integral With Respect to a Finite Weighted Sum of Measures) By induction, Theorem 3.54 yields, for nonnegative $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$\int f \, d\Big(\sum_{i=1}^{n} \alpha_i \mu_i\Big) = \sum_{i=1}^{n} \alpha_i \int f \, d\mu_i, \tag{3.55}$$

provided that f is integrable with respect to all measures $\mu_1, \ldots, \mu_n$. ◁

Example 3.56 (Integral With Respect to the Weighted Sum of Dirac Measures) Let $(\Omega, \mathscr{A})$ be a measurable space and, for $i \in \mathbb{N}$, let $\omega_i \in \Omega$, $\alpha_i \in \mathbb{R}$, $\alpha_i \ge 0$, and δ_{ω_i} denote the Dirac measure at ω_i. Then

$$\mu = \sum_{i=1}^{\infty} \alpha_i \delta_{\omega_i} \tag{3.56}$$

defined by $\mu(A) = \sum_{i=1}^{\infty} \alpha_i \delta_{\omega_i}(A)$, for all $A \in \mathscr{A}$, is a measure on $(\Omega, \mathscr{A})$ (see Example 1.61). For any nonnegative measurable function $f : \Omega \to \mathbb{R}$ we obtain

$$\begin{aligned} \int f \, d\mu &= \int f \, d \sum_{i=1}^{\infty} \alpha_i \delta_{\omega_i} && [(3.56)] \\ &= \sum_{i=1}^{\infty} \alpha_i \int f \, d\delta_{\omega_i} && [(3.53)] \\ &= \sum_{i=1}^{\infty} \alpha_i f(\omega_i). && [(3.23)] \end{aligned} \tag{3.57}$$

For $\mu(A) = \sum_{i=1}^{n} \alpha_i \delta_{\omega_i}(A)$, for all $A \in \mathscr{A}$, Equation (3.57) with $\alpha_i = 0$ for $i > n$, yields

$$\int f \, d\mu = \sum_{i=1}^{n} \alpha_i f(\omega_i). \tag{3.58}$$

Hence, the integral of a nonnegative measurable function f with respect to a finite or countable weighted sum of Dirac measures with nonnegative weights is a weighted sum of values of f. ◁

3.2.3 Integral With Respect to an Image Measure

The next theorem is relevant whenever we consider the integral of a composition $g \circ f$ of a mapping f with a numerical function g [see Eq. (2.25)] or the integral with respect to the image measure μ_f of μ under f [see Def. 2.79].

Theorem 3.57 (Transformation Theorem)
Let $f\colon (\Omega, \mathscr{A}, \mu) \to (\Omega', \mathscr{A}')$ and $g\colon (\Omega', \mathscr{A}') \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable.

(i) If g is nonnegative or integrable with respect to μ_f, then

$$\int g \, d\mu_f = \int g \circ f \, d\mu. \tag{3.59}$$

(ii) g is integrable with respect to μ_f if and only if $g \circ f$ is μ-integrable.

For a proof, see Bauer (2001, Corollary 19.2.1, p. 110).

If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a numerical measurable function and we replace g by the identity function $id\colon (\overline{\mathbb{R}}, \overline{\mathscr{B}}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, then Theorem 3.57 implies the following corollary.

Corollary 3.58 (An Implication of the Transformation Theorem)
If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is nonnegative or μ-integrable, then

$$\int id \, d\mu_f = \int_{\overline{\mathbb{R}}} id \, d\mu_f = \int f \, d\mu = \int_{\Omega} f \, d\mu. \tag{3.60}$$

Using the alternative notation of an integral introduced in Remark 3.32, Equation (3.59) can also be written

$$\int_{\bar{\mathbb{R}}} g(x)\, \mu_f(dx) = \int_{\Omega} g[f(\omega)]\, \mu(d\omega). \tag{3.61}$$

Correspondingly, Equation (3.60) can also be written

$$\int_{\bar{\mathbb{R}}} x\, \mu_f(dx) = \int_{\Omega} f(\omega)\, \mu(d\omega). \tag{3.62}$$

In Definition 3.10 we considered the case in which $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ is a nonnegative step function and defined its integral by $\int f\, d\mu = \sum_{i=1}^{n} \alpha_i\, \mu(A_i)$, presuming that $A_1, \ldots, A_n$ are pairwise disjoint. Now we consider a measurable function f with a finite number of values, which can be 0, positive, or negative.

Corollary 3.59 (Integral of a Function With a Finite Number of Values)
If $(\Omega, \mathscr{A}, \mu)$ is a measure space and $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ with pairwise different $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $\alpha_i \neq 0$, and pairwise disjoint $A_1, \ldots, A_n \in \mathscr{A}$, then f is μ-integrable if and only if $\mu(A_i) < \infty$ for all $i = 1, \ldots, n$. If f is μ-integrable, then

$$\int f\, d\mu = \sum_{i=1}^{n} \alpha_i\, \mu(A_i) = \sum_{i=1}^{n} \alpha_i\, \mu_f(\{\alpha_i\}). \tag{3.63}$$

(Proof p. 118)

3.2.4 Convergence Theorems

The next two theorems deal with convergence of integrals. In the first one, we assume that $f_1, f_2, \ldots$ is an increasing sequence of measurable functions that converge to f.

Theorem 3.60 (Monotone Convergence; B. Levi)
Let the functions $f_n \colon (\Omega, \mathscr{A}, \mu) \to (\bar{\mathbb{R}}, \bar{\mathscr{B}})$ be measurable, for all $n \in \mathbb{N}$.

(i) If the sequence $f_1, f_2, \ldots$ is increasing with $\lim_{n\to\infty} f_n = f$ and the functions f_n are nonnegative for all $n \in \mathbb{N}$ or μ-integrable for all $n \in \mathbb{N}$, then

$$\int f\, d\mu = \lim_{i\to\infty} \int f_n\, d\mu. \tag{3.64}$$

(ii) If the functions f_i are nonnegative for all $i \in \mathbb{N}$, then

$$\int \Big(\sum_{i=1}^{\infty} f_i \Big)\, d\mu = \sum_{i=1}^{\infty} \int f_i\, d\mu. \tag{3.65}$$

The integrals on both sides are finite or $+\infty$.

For a proof of (i), assuming nonnegativity, see Bauer (2001, Theorem 11.4, p. 59). For a proof of (i), assuming integrability, see Klenke (2008, Theorem 4.20, p. 93). For a proof of (ii), see Bauer (2001, Corollary 11.5, p. 60).

Note that, by definition of an 'infinite sum' (see Box 0.1), Equation (3.65) is equivalent to

$$\int \Big(\lim_{n\to\infty} \sum_{i=1}^{n} f_i \Big)\, d\mu = \lim_{n\to\infty} \sum_{i=1}^{n} \int f_i \, d\mu. \tag{3.66}$$

In the next theorem, we replace the assumption that $f_1, f_2, \ldots$ is increasing by the assumption that there is a μ-integrable function g dominating the absolute value functions of all f_n.

Theorem 3.61 (Dominated Convergence; Lebesgue Convergence Theorem) *If $g, f_n \colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, $n \in \mathbb{N}$, are μ-integrable and there is a measurable function $f \colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ with $\lim_{n\to\infty} f_n = f$, and $|f_n| \le g$ for all $n \in \mathbb{N}$, then*

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu, \tag{3.67}$$

and this integral is finite.

For a proof, see Bauer (2001, Theorem 15.6, p. 83).

3.3 Lebesgue and Riemann Integral

The Lebesgue measures λ_n on $(\mathbb{R}^n, \mathscr{B}_n)$, $n = 1, 2, 3$, represent *length*, *area*, and *volume*, respectively. As the examples illustrated by Figure 3.4 shows, the integral of the $(\mathscr{B}, \mathscr{B})$-measurable function $f \colon \mathbb{R} \to \mathbb{R}$ with respect to the Lebesgue measure $\lambda = \lambda_1$, i. e., the Lebesgue integral, yields the difference between the *areas* marked by + and the areas marked by −, respectively.

It is useful to know conditions under which the Lebesgue integral and the Riemann integrals are identical, because a lot of tools are available for Riemann integration (see, e. g., Ellis & Gulick, 2006). The following theorem is proved in Klenke (2008, Theorem 4.23, p. 96), who also provides a brief definition of the Riemann integral and Riemann integrability.

Theorem 3.62 (Lebesgue Integral and Riemann Integral)

Let λ denote the Lebesgue measure on $(\mathbb{R}, \mathscr{B})$ and let $[a, b]$, $a, b \in \mathbb{R}$, $a < b$, be a closed interval. If $f \colon [a, b] \to \mathbb{R}$ is Riemann integrable on $[a, b]$, then f is λ-integrable, and

$$\int_a^b f(x)\, dx = \int_{[a,b]} f \, d\lambda = \int_{[a,b]} f(x)\, \lambda(dx) = \int \mathbb{1}_{[a,b]} \cdot f \, d\lambda, \tag{3.68}$$

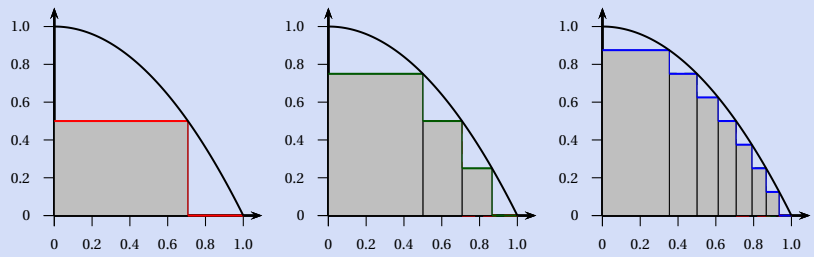


Figure 3.5. Illustrating the construction of the Riemann integral 3.22

where $\int_a^b f(x)\,dx$ *denotes the Riemann integral from* a *to* b.

Note that

$$\int_a^b f(x)\,dx \;=\; F(x)\Big|_a^b \;:=\; F(b)-F(a), \tag{3.69}$$

where F is an antiderivative of f.

Remark 3.63 (Lebesgue vs. Riemann Integral) If we want to define the integral of a measurable function $f\colon (\Omega,\mathscr{A}) \to (\mathbb{R},\mathscr{B})$, where the set Ω is not necessarily a subset of $\mathbb{R}$, then this means that the traditional Riemann integral cannot be used. The *Riemann integral* is constructed by partitioning the *domain* of f, the set $\mathbb{R}$ of real numbers into small intervals and adding the area of the rectangles on these intervals in order to approximate the area under the function $f\colon \mathbb{R} \to \mathbb{R}$ (see Fig. 3.5). If $\Omega \not\subset \mathbb{R}$, then this idea does not work any more. Instead, the *Lebesgue integral* is constructed by partitioning the *codomain* of f, which is the set $\mathbb{R}$ of real numbers into small intervals (see Fig. 3.4). This is also possible if $\Omega \not\subset \mathbb{R}$, and in this aspect, the Lebesgue integral is more general than the Riemann integral. ◁

Note, however, that even if the domain of f is a subset of the set of real numbers, there are functions for which the Lebesgue integral exists and the Riemann integral does not exist (see, e. g., Klenke, 2008, Example 4.24). Also note that there are functions that are Riemann integrable on a half-open or unbounded interval but not Lebesgue integrable (see, e. g., Klenke, 2008, Remark 4.25, p. 97).

Example 3.64 (Using the Riemann Integral) As a simple application of Theorem 3.62 consider the function f defined by $f(x) = 10 - x^2$ on a closed interval $[a,b]$. Because f is a continuous function, it is Riemann integrable. Hence, we can apply Equation (3.68). For $a = -5$ and $b = 5$, this equation yields

$$\int_{[-5,5]} f \, d\lambda = \int_{-5}^{5} f(x) \, dx = 16.\bar{6}$$

(see Exercise 3-11). This integral is the difference between the areas marked by + and the areas marked by − in Figure 3.4. ◁

3.4 Density

A density f can be interpreted as a weighting function of the values of the original measure μ. If we consider a measure μ on a measurable space $(\Omega, \mathscr{A})$ with $\mu(\{x\}) > 0$ for all $x \in \Omega$, then this means that the values $\mu(\{x\})$ of the singletons $\{x\}$ are multiplied by a nonnegative number $f(x)$. If we consider the Lebesgue measure λ on $\mathbb{R}$, then, intuitively speaking, any infinitesimal interval dx gets a weight $f(x)$. Using such a density, a new measure ν on $\mathscr{A}$ is introduced, where $\nu(A)$ is the integral of f over A with respect to μ. The most important examples are densities with respect to the Lebesgue measure (see Example 3.69).

Theorem 3.65 (Measure With Density)
Let $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a nonnegative measurable function. Then the function $\nu\colon \mathscr{A} \to \overline{\mathbb{R}}$ defined by

$$\nu(A) = \int_A f \, d\mu, \quad \forall A \in \mathscr{A}, \tag{3.70}$$

is a measure, called the measure with density f with respect to μ. It is denoted by $f \odot \mu$, i.e., $f \odot \mu := \nu$.

For a proof, see Bauer (2001, Theorem 17.1, p. 96).

The notation $f \odot \mu$ has been adopted from Elstrodt (2007, p. 127). Using this notation, Equation (3.70) can also be written

$$f \odot \mu(A) = \int_A f \, d\mu, \quad \forall A \in \mathscr{A}. \tag{3.71}$$

Definition 3.66 (Density)
Let ν be a measure on $(\Omega, \mathscr{A})$. If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a nonnegative measurable function satisfying Equation (3.70), then it is called a density of ν with respect to μ.

The following theorem generalizes Equation (3.52).

Theorem 3.67 (Integral With Respect to a Measure With Density)
Let $(\Omega, \mathscr{A}, \mu)$ be a measure space and $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ a nonnegative measurable function. Furthermore, let $f\odot\mu\colon \mathscr{A} \to \overline{\mathbb{R}}$ be the measure with density f with respect to μ and let $g\colon (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ be measurable.

(i) If g is nonnegative, then

$$\int g \, d\, f\odot\mu = \int g \cdot f \, d\,\mu. \tag{3.72}$$

(ii) g is integrable with respect to $f\odot\mu$ if and only if $g \cdot f$ is μ-integrable.
(iii) If g is integrable with respect to $f\odot\mu$, then Equation (3.72) holds.

For a proof, see Bauer (2001, Theorem 17.3, p. 96, 97).

In the following theorem we summarize some necessary and sufficient conditions for μ-equivalence of measurable functions on a measure space.

Theorem 3.68 (Necessary and Sufficient Condition of μ-Equivalence)
Let $f, g\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable functions, let $\mathscr{E} \subset \mathscr{A}$, and consider:

(a) $f \underset{\mu}{=} g$.
(b) $\int_A f \, d\,\mu = \int_A g \, d\,\mu, \quad \forall A \in \mathscr{A}$.
(c) $f\odot\mu = g\odot\mu$.
(d) $\int_A f \, d\,\mu = \int_A g \, d\,\mu, \quad \forall A \in \mathscr{E}$.

Then:

(i) (a) $\Rightarrow$ (b), if f and g are quasi-μ-integrable.
(ii) (a) $\Leftrightarrow$ (b), if f, g are μ-integrable.
(iii) (a) $\Leftrightarrow$ (c), if f, g are μ-integrable and nonnegative.
(iv) (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ (c) $\Leftrightarrow$ (d), if f, g are μ-integrable, nonnegative, and $\mathscr{E} \subset \mathscr{A}$ is $\cap$-stable with $\sigma(\mathscr{E}) = \mathscr{A}$.

(Proof p. 119)

Example 3.69 (A Density of the Normal Distribution) As a special case of Equation (3.70) consider

$$\nu(A) = \int_A f \, d\,\lambda, \quad \forall\, A \in \mathscr{A}, \tag{3.73}$$

with

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-x^2}{2}\right), \quad x \in \mathbb{R}\,. \tag{3.74}$$

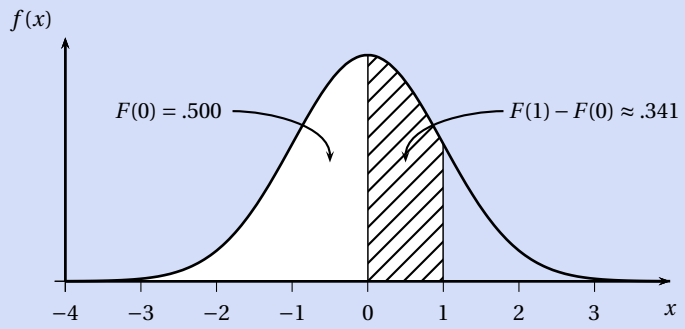


Figure 3.6. Integral of a density for two intervals

In this case, the measure $\nu = f \odot \lambda$ is a probability measure and it is called the *standard normal distribution*. For an interval $[a,b]$, Theorem 3.67 yields

$$\nu([a,b]) = \int 1_{[a,b]}\, d\nu = \int 1_{[a,b]}\, d\, f \odot \lambda = \int 1_{[a,b]} f\, d\lambda = \int_a^b f(x)\, dx, \quad (3.75)$$

because f is Riemann μ-integrable (see Th. 3.62). According to this equation, the value $\nu([a,b])$ of the interval $[a,b]$ can be represented as the area between the density and the x-axis above $[a,b]$. Figure 3.6 illustrates this fact for the interval $[0,1]$. In this figure,

$$F(\alpha) = \int_{-\infty}^{\alpha} f(x)\, dx, \quad \alpha \in \mathbb{R}, \quad (3.76)$$

denotes the corresponding distribution function (see Def. 5.77), which is a special antiderivative of f [see Eq. (3.69)]. ◁

3.5 Absolute Continuity and the Radon-Nikodym Theorem

Let μ and ν be measures on a measurable space $(\Omega, \mathscr{A})$. A necessary and sufficient condition for the existence of a density of ν with respect to μ is formulated in the Radon-Nikodym Theorem (see Th. 3.72), which is used not only for densities but also introducing conditional expectations (see ch. 10). The following definition prepares this theorem.

Definition 3.70 (Absolute Continuity)
Let μ and ν be measures on a measurable space $(\Omega, \mathscr{A})$.

(i) The measure ν is called absolutely continuous with respect to μ, denoted $\nu \underset{\mathscr{A}}{\ll} \mu$, if

$$\forall A \in \mathscr{A}: \quad \mu(A) = 0 \quad \Rightarrow \quad \nu(A) = 0. \tag{3.77}$$

(ii) The measures μ and ν are called null-set equivalent, denoted $\nu \underset{\mathscr{A}}{\approx} \mu$, if $\nu \underset{\mathscr{A}}{\ll} \mu$ and $\mu \underset{\mathscr{A}}{\ll} \nu$, i.e., if

$$\forall A \in \mathscr{A}: \quad \mu(A) = 0 \quad \Leftrightarrow \quad \nu(A) = 0. \tag{3.78}$$

If there is ambiguity about the measurable space, we use the terms absolutely continuous on $(\Omega, \mathscr{A})$ and null-set equivalent on $(\Omega, \mathscr{A})$.

Remark 3.71 (An Implication) If there is a density f of ν with respect to μ, then $\nu \underset{\mathscr{A}}{\ll} \mu$. This is a straightforward implication of Lemma 3.45 and (3.72) (see Exercise 3-12). ◁

Vice versa, if $\nu \underset{\mathscr{A}}{\ll} \mu$, then, according to the following theorem, there is a density f of ν with respect to μ, provided that μ and ν are σ-finite (see Definition 1.62).

Theorem 3.72 (Radon-Nikodym)
Let μ and ν be σ-finite measures on a measurable space $(\Omega, \mathscr{A})$.

(i) Then ν has a measurable density with respect to μ if and only if $\nu \underset{\mathscr{A}}{\ll} \mu$. This density is denoted by $\frac{d\nu}{d\mu}$ and called Radon-Nikodym derivative.
(ii) If $\nu \underset{\mathscr{A}}{\ll} \mu$, then $\frac{d\nu}{d\mu}$ is real-valued μ-almost everywhere.

For a proof, see Klenke (2008, Corollary 7.34, p. 157) or Bauer (2001, Theorem 17.10, p. 102 and Theorem 17.11, p. 104).

Remark 3.73 (μ-Equivalence of Densities) Theorems 3.72 and 3.68 (iii) imply that all densities of ν with respect to μ are pairwise μ-equivalent, provided that they exist. For σ-finite measures μ and ν, each density $\frac{d\nu}{d\mu}$ is also called a *Radon-Nikodym derivative* (of ν with respect to μ). ◁

Remark 3.74 (An Implication of the Radon-Nikodym Theorem) Theorem 3.72 implies for σ-finite measures ν and μ: If ν and μ are null-set equivalent, then $\frac{d\nu}{d\mu}$ and $\frac{d\mu}{d\nu}$ both exist. ◁

The Radon-Nikodym theorem is used to prove the existence of the conditional expectation [see the proof of Theorem 10.9 (Bauer, 1996, Theorem 15.1, p. 111)]. The following corollary immediately follows from Theorems 3.68 and 3.72.

Corollary 3.75 (An Implication of the Radon-Nikodym Theorem)
Let μ and ν be σ-finite measures on a measurable space $(\Omega, \mathscr{A})$, and suppose $\nu \underset{\mathscr{A}}{\ll} \mu$. Furthermore, let $g\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a measurable function.

(i) If g is nonnegative or ν-integrable, then

$$\int g \, d\nu = \int g \cdot \frac{d\nu}{d\mu} \, d\mu. \tag{3.79}$$

(ii) g is ν-integrable if and only if $g \cdot \frac{d\nu}{d\mu}$ is μ-integrable.

3.6 Integral With Respect to a Product Measure

The following theorem shows that integration with respect to a product measure can be decomposed into a two-fold iterated integration where the order of integration is arbitrary.

Theorem 3.76 (Fubini)
Let $(\Omega_i, \mathscr{A}_i, \mu_i)$, $i = 1,2$, be σ-finite measure spaces and let $f\colon \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ be $(\mathscr{A}_1 \otimes \mathscr{A}_2, \overline{\mathscr{B}})$-measurable. Furthermore, let $f_i\colon \Omega_i \to \overline{\mathbb{R}}$, $i = 1,2$, be defined by

$$f_1(\omega_1) := \int f(\omega_1, \omega_2) \, \mu_2(d\omega_2) \quad \text{and} \quad f_2(\omega_2) := \int f(\omega_1, \omega_2) \, \mu_1(d\omega_1).$$

If f is nonnegative or integrable with respect to the product measure $\mu_1 \otimes \mu_2$, then the functions f_i are $(\mathscr{A}_i, \overline{\mathscr{B}})$-measurable, $i = 1,2$. Furthermore,

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} f \, d(\mu_1 \otimes \mu_2) &= \int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) \, \mu_1 \otimes \mu_2 \, [d(\omega_1, \omega_2)] \\ &= \int_{\Omega_1} \Big(\int_{\Omega_2} f(\omega_1, \omega_2) \, \mu_2(d\omega_2) \Big) \mu_1(d\omega_1) \\ &= \int_{\Omega_2} \Big(\int_{\Omega_1} f(\omega_1, \omega_2) \, \mu_1(d\omega_1) \Big) \mu_2(d\omega_2). \end{aligned} \tag{3.80}$$

For a proof, see Klenke (2008, Th. 14.16, p. 276). If $f = 1_C$ for $C \in \mathscr{A}_1 \otimes \mathscr{A}_2$, then this theorem and (3.9) immediately yield the following corollary:

Corollary 3.77 (Indicators)
Let $(\Omega_i, \mathscr{A}_i, \mu_i)$, $i = 1,2$, be σ-finite measure spaces and let $f\colon \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ be $(\mathscr{A}_1 \otimes \mathscr{A}_2, \overline{\mathscr{B}})$-measurable. Furthermore, let $C \in \mathscr{A}_1 \otimes \mathscr{A}_2$ and define

$$\forall\, \omega_1 \in \Omega_1 \colon\ C_{\omega_1} := \{\omega_2 \in \Omega_2 \colon (\omega_1, \omega_2) \in C\}$$

and

$$\forall\, \omega_2 \in \Omega_2 \colon\ C_{\omega_2} := \{\omega_1 \in \Omega_1 \colon (\omega_1, \omega_2) \in C\}.$$

Then

$$\mu_1 \otimes \mu_2(C) \;=\; \int \mu_2(C_{\omega_1})\, \mu_1(d\omega_1) \;=\; \int \mu_1(C_{\omega_2})\, \mu_2(d\omega_2). \tag{3.81}$$

(Proof p. 119)

Remark 3.78 (A Special Case) Choosing $C = A_1 \times A_2$, Equation (3.81) yields

$$\mu_1 \otimes \mu_2(A_1 \times A_2) \;=\; \int \mu_2(A_2)\, \mu_1(d\omega_1) \;=\; \mu_2(A_2) \cdot \int \mu_1(d\omega_1) \;=\; \mu_1(A_1) \cdot \mu_2(A_2),$$

which is consistent with Equation (1.49). ◁

3.7 Proofs

Proof of Lemma 3.33

(i) If f is measurable and $A \in \mathscr{A}$, then 1_A is measurable as well (see Th. 2.57 and Example 2.12). Suppose that f is quasi-μ-integrable, i. e., suppose that $\int f^+ d\mu$ or $\int f^- d\mu$ are finite. Because

$$(1_A f)^+ \;=\; 1_A f^+ \quad \text{and} \quad (1_A f)^- \;=\; 1_A f^-$$

as well as

$$0 \le 1_A f^+ \le f^+ \quad \text{and} \quad 0 \le 1_A f^- \le f^-,$$

monotonicity of the integral of nonnegative measurable functions (Lemma 3.26) yields

$$0 \;\le \int (1_A f)^+ d\mu \;=\; \int 1_A f^+ d\mu \;\le \int f^+ d\mu$$

and

$$0 \;\le \int (1_A f)^- d\mu \;=\; \int 1_A f^- d\mu \;\le \int f^- d\mu,$$

which implies that $\int (1_A f)^+ d\mu$ or $\int (1_A f)^- d\mu$ is finite. Hence f is quasi-μ-integrable.

(ii) If f is μ-integrable, then $\int f^+ d\mu < \infty$ and $\int f^- d\mu < \infty$. Just like in the proof of (i) this implies $\int (1_A f)^+ d\mu < \infty$ and $\int (1_A f)^- d\mu < \infty$. Hence, $1_A f$ is μ-integrable.

Proof of Theorem 3.36

(i) Step 1: Let $\alpha \ge 0$, f be a nonnegative step function, and $f = \sum_{i=1}^n \alpha_i 1_{A_i}$ a normal representation (see Rem. 3.8). Then, according to (3.4),

$$\begin{aligned}
\int \alpha f \, d\mu &= \int \alpha \sum_{i=1}^{n} \alpha_i 1_{A_i} \, d\mu && \text{[Rem. 3.8]} \\
&= \int \sum_{i=1}^{n} (\alpha\alpha_i) 1_{A_i} \, d\mu \\
&= \sum_{i=1}^{n} (\alpha\alpha_i) \, \mu(A_i) && \text{[(3.4)]} \\
&= \alpha \sum_{i=1}^{n} \alpha_i \, \mu(A_i) \\
&= \alpha \int f \, d\mu. && \text{[(3.4)]}
\end{aligned}$$

Step 2: Let $\alpha \geq 0$, f be a nonnegative measurable function, and $f_1 \leq f_2 \leq \ldots$ an increasing sequence of nonnegative step functions with $\lim_{n\to\infty} f_n = f$ (see Th. 3.19). Then, according to Equation (3.22),

$$\begin{aligned}
\int \alpha f \, d\mu &= \int \alpha \lim_{n\to\infty} f_n \, d\mu \\
&= \int \lim_{n\to\infty} (\alpha f_n) \, d\mu \\
&= \lim_{n\to\infty} \int \alpha f_n \, d\mu && \text{[(3.22)]} \\
&= \lim_{n\to\infty} \alpha \int f_n \, d\mu && \text{[Step 1]} \\
&= \alpha \lim_{n\to\infty} \int f_n \, d\mu \\
&= \alpha \int f \, d\mu. && \text{[(3.22)]}
\end{aligned}$$

Step 3: Assume that $\alpha \geq 0$ and that f is quasi-μ-integrable. Because $\alpha f = \alpha(f^+ - f^-) = \alpha f^+ - \alpha f^-$,

$$\begin{aligned}
\int \alpha f \, d\mu &= \int \alpha f^+ \, d\mu - \int \alpha f^- \, d\mu && \text{[(3.27)]} \\
&= \alpha \int f^+ \, d\mu - \alpha \int f^- \, d\mu && \text{[Step 2]} \\
&= \alpha \left(\int f^+ \, d\mu - \int f^- \, d\mu \right) \\
&= \alpha \int f \, d\mu. && \text{[(3.27)]}
\end{aligned}$$

This proves Equation (3.32) for $\alpha \geq 0$. For $\alpha < 0$, note that

$$(\alpha f)^+ = -\alpha f^- \quad \text{and} \quad (\alpha f)^- = -\alpha f^+. \tag{3.82}$$

Therefore,

$$\int \alpha f \, d\mu = \int (\alpha f)^+ \, d\mu - \int (\alpha f)^- \, d\mu \qquad \text{[(3.27)]}$$

$$
\begin{aligned}
&= \int (-\alpha) f^- \, d\mu - \int (-\alpha) f^+ \, d\mu && \text{[(3.82)]} \\
&= -\alpha \int f^- \, d\mu - (-\alpha) \int f^+ \, d\mu && \text{[}-\alpha > 0\text{, first part of Step 3]} \\
&= \alpha \left(\int f^+ \, d\mu - \int f^- \, d\mu \right) \\
&= \alpha \int f \, d\mu. && \text{[(3.27)]}
\end{aligned}
$$

This shows that $\int \alpha f \, d\mu = \alpha \int f \, d\mu$ holds for all $\alpha \in \mathbb{R}$ all quasi-μ-integrable f, and therefore also for all integrable f. This also implies that αf is quasi-μ-integrable or μ-integrable if f is quasi-μ-integrable or μ-integrable, respectively.

(ii) Step 1: Let f and g be nonnegative step functions and let $f = \sum_{i=1}^n \alpha_i 1_{A_i}$, $g = \sum_{j=1}^m \beta_j 1_{B_j}$ be normal representations (see Rem. 3.8) with $\bigcup_{i=1}^n A_i = \Omega$ and $\bigcup_{j=1}^m B_j = \Omega$. (Note that these latter requirements can always be fulfilled using $A_n := \Omega \setminus \bigcup_{i=1}^{n-1} A_i$ and $\alpha_n := 0$, if $f = \sum_{i=1}^{n-1} \alpha_i 1_{A_i}$ is already a normal representation.) Then $f + g = \sum_{i=1}^n \alpha_i 1_{A_i} + \sum_{j=1}^m \beta_j 1_{B_j}$ is again a nonnegative step function (see Def. 3.9) and

$$
f + g = \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} 1_{C_{ij}}
$$

is a normal representation, where $C_{ij} = A_i \cap B_j$ and $\gamma_{ij} = \alpha_i + \beta_j$. Note that some of these sets C_{ij} may be empty. Now,

$$
\begin{aligned}
\int f + g \, d\mu &= \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} \, \mu(C_{ij}) && \text{[(3.4)]} \\
&= \sum_{i=1}^n \sum_{j=1}^m (\alpha_i + \beta_j) \, \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \sum_{j=1}^m \alpha_i \, \mu(A_i \cap B_j) + \sum_{i=1}^n \sum_{j=1}^m \beta_j \, \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \alpha_i \sum_{j=1}^m \mu(A_i \cap B_j) + \sum_{j=1}^m \beta_j \sum_{i=1}^n \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \alpha_i \, \mu(A_i) + \sum_{j=1}^m \beta_j \, \mu(B_j) && \text{[Rem. 1.47]} \\
&= \int f \, d\mu + \int g \, d\mu. && \text{[(3.4)]}
\end{aligned}
$$

Step 2: Let f, g be nonnegative measure functions and $f_1 \le f_2 \le \ldots$, $g_1 \le g_2 \le \ldots$ increasing sequences of nonnegative step functions with $\lim_{n\to\infty} f_n = f$ and $\lim_{n\to\infty} g_n = g$, respectively (see Th. 3.19). Then $f_1 + g_1 \le f_2 + g_2 \le \ldots$ is an increasing sequence of nonnegative step functions with $\lim_{n\to\infty}(f_n + g_n) = \lim_{n\to\infty} f_n + \lim_{n\to\infty} g_n = f + g$. Then

$$\begin{aligned}\int f+g\,d\mu &= \int \lim_{n\to\infty}(f_n+g_n)\,d\mu \\ &= \lim_{n\to\infty}\int (f_n+g_n)\,d\mu && [(3.22)] \\ &= \lim_{n\to\infty}\left(\int f_n\,d\mu+\int g_n\,d\mu\right) && [\text{Step 1}] \\ &= \lim_{n\to\infty}\int f_n\,d\mu+\lim_{n\to\infty}\int g_n\,d\mu \\ &= \int f\,d\mu+\int g\,d\mu. && [(3.22)]\end{aligned}$$

Step 3: Assume that f is quasi-μ-integrable and g is μ-integrable. Then

$$f+g = f^+-f^-+g^+-g^-,$$

$$f+g = (f+g)^+-(f+g)^-.$$

This implies

$$(f+g)^+-(f+g)^- = f^+-f^-+g^+-g^-,$$

which is equivalent to

$$(f+g)^++f^-+g^- = (f+g)^-+f^++g^+.$$

Applying the result of Step 2 yields

$$\begin{aligned}&\int (f+g)^+\,d\mu+\int f^-\,d\mu+\int g^-d\mu \\ =&\int (f+g)^-\,d\mu+\int f^+d\mu+\int g^+d\mu.\end{aligned} \tag{3.83}$$

If g is μ-integrable, then $\int g^+\,d\mu$ and $\int g^-\,d\mu$ are finite, and if f is quasi-μ-integrable, then at most one of $\int f^+\,d\mu$ and $\int f^-\,d\mu$ is infinite, the other one is finite. Furthermore, $(f+g)^+\le f^++g^+$ and $(f+g)^-\le f^-+g^-$. Hence, Lemma 3.26 implies

$$\int (f+g)^+\,d\mu\le\int f^++g^+\,d\mu=\int f^+\,d\mu+\int g^+\,d\mu \tag{3.84}$$

and

$$\int (f+g)^-\,d\mu\le\int f^-+g^-\,d\mu=\int f^-\,d\mu+\int g^-\,d\mu.$$

Therefore, at most one of the integrals $\int(f+g)^+\,d\mu$ and $\int(f+g)^-\,d\mu$ is infinite and this implies that $f+g$ is quasi-μ-integrable. If $\int(f+g)^+\,d\mu=\infty$, then

$$\int (f+g)\,d\mu=\int (f+g)^+\,d\mu-\int (f+g)^-\,d\mu=\infty,$$

and, according to (3.84), $\int f^+ \, d\mu = \infty$. This implies

$$\int f \, d\mu + \int g \, d\mu = \int f^+ d\mu - \int f^- d\mu + \int g d\mu = \infty.$$

Analogously, if $\int (f+g)^- \, d\mu = \infty$, then

$$\int (f+g) \, d\mu = -\infty = \int f \, d\mu + \int g \, d\mu.$$

If both, $\int (f+g)^+ \, d\mu$ and $\int (f+g)^- \, d\mu$ are finite, then (3.83) is equivalent to

$$\int (f+g)^+ \, d\mu - \int (f+g)^- \, d\mu = \int f^+ d\mu - \int f^- d\mu + \int g^+ d\mu - \int g^- d\mu,$$

which in turn is equivalent to

$$\int (f+g) \, d\mu = \int f \, d\mu + \int g \, d\mu.$$

Proof of Corollary 3.38

Because $|f| = f^+ + f^-$, this proposition immediately follows from the definition of integrability (see Def. 3.28) and linearity of the integral [see Eq. (3.34)].

Proof of Lemma 3.40

If f is $\mathscr{C}$-measurable, then f^+ and f^- are $\mathscr{C}$-measurable as well (see Th. 2.66). Furthermore, f^+ and f^- can be represented as limits of increasing sequences of nonnegative step functions on $(\Omega, \mathscr{C})$ [see Th. 3.19 (i)]. Hence, according to Equations (3.22) and (3.4), the values of the integrals $\int f^+ d\mu$ and $\int f^- d\mu$ only depend on the values $\mu(A)$, $A \in \mathscr{C}$. Therefore, if μ and ν are identical on $\mathscr{C}$, then $\int f \, d\mu = \int f \, d\nu$, for all $\mathscr{C}$-measurable functions that are nonnegative or μ-integrable.

Proof of Lemma 3.41

The proof is by contraposition. Define

$$A_+ := \{\omega \in \Omega \colon f(\omega) = \infty\} \quad \text{and} \quad A_- := \{\omega \in \Omega \colon f(\omega) = -\infty\}.$$

If $\mu(A_+ \cup A_-) > 0$, then $\mu(A_+) > 0$ or $\mu(A_-) > 0$. Now, if $\mu(A_+) > 0$, then define the increasing sequence $g_n \colon \Omega \to [0, \infty)$, $n \in \mathbb{N}$, by $g_n = n \cdot 1_{A_+}$. Because $1_{A_+} \cdot f^+ = \lim_{n \to \infty} g_n$,

$$\begin{aligned}
\int f^+ d\mu &= \int 1_{A_+} \cdot f^+ d\mu + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu && [(3.36)] \\
&= \lim_{n\to\infty} \int g_n \, d\mu + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu && [\text{Def. } 3.24] \\
&= \lim_{n\to\infty} \int n \cdot 1_{A_+} \, d\mu + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu \\
&= \lim_{n\to\infty} n \cdot \mu(A_+) + \int 1_{\Omega \setminus A_+} \cdot f^+ d\mu && [(3.4)] \\
&= \infty.
\end{aligned}$$

Analogously we can prove that $\int f^- d\mu = \infty$, if $\mu(A_-) > 0$. Hence, f is not μ-integrable.

Proof of Lemma 3.44

If $f(\omega) > 0$, for all $\omega \in A$, then $1_A \cdot f \colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a nonnegative measurable function (see Th. 2.57). Hence, $\int 1_A \cdot f \, d\mu \geq 0$ (see Defs. 3.24 and 3.10). Because $\mu\big(\{\omega \in \Omega \colon (1_A \cdot f(\omega) > 0\}\big) = \mu(A)$, the assumption $\mu(A) > 0$ implies that $1_A \cdot f \underset{\mu}{=} 0$ does *not hold*. Therefore, according to Equation (3.40), $\int 1_A \cdot f \, d\mu \neq 0$, and we can conclude: $\int 1_A \cdot f \, d\mu > 0$.

Proof of Lemma 3.45

If f is measurable and $A \in \mathscr{A}$ with $\mu(A) = 0$, then $1_A \cdot f$ is measurable (see Th. 2.57) and $1_A \cdot f \underset{\mu}{=} 0$ (see Exercise 3-9). This implies $(1_A \cdot f)^+ \underset{\mu}{=} 0$ and $(1_A \cdot f)^- \underset{\mu}{=} 0$. Now Equation (3.40) yields $\int (1_A \cdot f)^+ d\mu = 0$ and $\int (1_A \cdot f)^- d\mu = 0$. Hence, $\int 1_A \cdot f \, d\mu$ exists (see Def. 3.28) and Equation (3.27) implies

$$\int 1_A \cdot f \, d\mu = \int (1_A \cdot f)^+ d\mu - \int (1_A \cdot f)^- d\mu = 0.$$

Proof of Lemma 3.47

Define $A := \{\omega \in \Omega \colon f(\omega) \neq g(\omega)\}$. Then $f \underset{\mu}{=} g$ implies $\mu(A) = 0$. Hence,

$$\begin{aligned}
\int f \, d\mu &= \int_{\Omega \setminus A} f \, d\mu && [(3.43)] \\
&= \int_{\Omega \setminus A} g \, d\mu && [\text{Def. of } A] \\
&= \int g \, d\mu. && [(3.43)]
\end{aligned}$$

Proof of Theorem 3.48

(a)

$$
\begin{aligned}
& f, g \text{ are } \mu\text{-integrable and } f \underset{\mu}{=} g \\
\Rightarrow\ & \forall A \in \mathscr{A}\colon\ 1_A \cdot f \underset{\mu}{=} 1_A \cdot g \text{ and } 1_A \cdot f,\ 1_A \cdot g \text{ are } \mu\text{-integrable} && [(2.33), (3.29)] \\
\Rightarrow\ & \forall A \in \mathscr{A}\colon\ \int_A f\,d\mu = \int_A g\,d\mu. && [(3.44)]
\end{aligned}
$$

(b) If f, g are μ-integrable, then f, g are real-valued μ-a.e. (see Lemma 3.41). Hence, for

$$
B := \big\{\omega \in \Omega\colon f(\omega) \in \{-\infty, \infty\}\big\} \cup \big\{\omega \in \Omega\colon g(\omega) \in \{-\infty, \infty\}\big\},
$$

$\mu(B) = 0$. Now define

$$
A_> := \{\omega \in \Omega\colon f(\omega) > g(\omega)\} \quad \text{and} \quad A_< := \{\omega \in \Omega\colon f(\omega) < g(\omega)\}.
$$

Then

$$
\begin{aligned}
& \forall A \in \mathscr{A}\colon \int_A f\,d\mu = \int_A g\,d\mu \\
\Rightarrow\ & \forall A \in \mathscr{A}\colon \int_{A \cap B^c} f\,d\mu = \int_{A \cap B^c} g\,d\mu && [(3.43)] \\
\Rightarrow\ & \forall A \in \mathscr{A}\colon \int_{A \cap B^c} (f-g)\,d\mu = 0 && [f, g\ \mu\text{-integrable, } (3.34)] \\
\Rightarrow\ & \int_{A_> \cap B^c} (f-g)\,d\mu = 0 \quad \text{and} \quad \int_{A_< \cap B^c} (f-g)\,d\mu = 0 && [A_>, A_< \in \mathscr{A}\,] \\
\Rightarrow\ & \mu(A_> \cap B^c) = 0 \text{ and } \mu(A_< \cap B^c) = 0 && [(3.44)] \\
\Rightarrow\ & \mu(A_> \cup A_<) \\
& = \mu\big((A_> \cap B^c) \cup (A_> \cap B) \cup (A_< \cap B^c) \cup (A_< \cap B)\big) = 0 && [\text{Box 1.1 (ii)}] \\
\Rightarrow\ & f \underset{\mu}{=} g. && [\text{Def. 2.68}]
\end{aligned}
$$

Proof of Theorem 3.52

(i) Define $A := \{\omega \in \Omega\colon f(\omega) > g(\omega)\}$. Then $f \underset{\mu}{\leq} g$ implies $\mu(A) = 0$. Furthermore, define

$$
A_{-\infty} := \{\omega \in \Omega \setminus A\colon f(\omega) = -\infty\} \quad \text{and} \quad A_\infty := \{\omega \in \Omega \setminus A\colon f(\omega) = g(\omega) = \infty\}.
$$

If $\mu(A_{-\infty}) > 0$, then $\int f\,d\mu = -\infty$ [see (3.39)]. Therefore, $\int f\,d\mu \leq \int g\,d\mu$. If $\mu(A_\infty) > 0$, then $\int f\,d\mu = \int g\,d\mu = \infty$ [see (3.38)], and therefore $\int f\,d\mu \leq \int g\,d\mu$.

If $\mu(A_{-\infty}) = \mu(A_\infty) = 0$, then f, g are real-valued μ-a.e. Now define $B := (\Omega \setminus A) \setminus (A_\infty \cup A_{-\infty})$, i. e., $\mu(\Omega \setminus B) = 0$ and

$$\forall \omega \in B\colon\ f(\omega), g(\omega) \text{ are finite, } f(\omega) \le g(\omega).$$

If $\int (1_B f)^- d\mu = \infty$, then $\int_B f\, d\mu = -\infty$. Hence, $\int f\, d\mu \le \int g\, d\mu$. If $\int (1_B f)^+ d\mu = \infty$, then $1_B f \le 1_B g$ implies $(1_B f)^+ \le (1_B g)^+$ and $\int (1_B g)^+ d\mu = \infty$ [see Eq. (3.24)]. Therefore, $\int f\, d\mu = \int g\, d\mu = \infty$, which implies $\int f\, d\mu \le \int g\, d\mu$.

Now, if all $(1_B f)^+$, $(1_B f)^-$, $(1_B g)^+$, $(1_B g)^-$ are μ-integrable, then

$$\begin{aligned}
\int g\, d\mu &= \int_B g\, d\mu && [(3.43)] \\
&= \int_B (f + g - f)\, d\mu && [1_B f,\, 1_B g \text{ are real-valued}] \\
&= \int_B f\, d\mu + \int_B (g - f)\, d\mu && [(3.34)] \\
&\ge \int_B f\, d\mu && [1_B (g - f) \ge 0,\ (3.24)] \\
&= \int f\, d\mu. && [(3.43)]
\end{aligned}$$

(ii) Define

$$B := \big\{ \omega \in \Omega\colon f(\omega) \in \{-\infty, \infty\} \big\} \cup \big\{ \omega \in \Omega\colon g(\omega) \in \{-\infty, \infty\} \big\}.$$

If f, g are μ-integrable, then Lemma 3.41 implies $\mu(B) = 0$. Furthermore, define $A := \{\omega \in \Omega \setminus B\colon f(\omega) \ge g(\omega)\}$. Then $f \underset{\mu}{\le} g$ implies $\mu(A) = 0$. Hence, $\mu(A \cup B) = 0$ [see Box 1.1 (xi)]. Now,

$$\begin{aligned}
\int g\, d\mu &= \int_{\Omega \setminus (A \cup B)} g\, d\mu && [(3.43)] \\
&= \int_{\Omega \setminus (A \cup B)} (f + g - f)\, d\mu && [1_{\Omega \setminus (A \cup B)} f,\, 1_{\Omega \setminus (A \cup B)} g \text{ are real-valued}] \\
&= \int_{\Omega \setminus (A \cup B)} f\, d\mu + \int_{\Omega \setminus (A \cup B)} (g - f)\, d\mu && [(3.34)] \\
&> \int_{\Omega \setminus (A \cup B)} f\, d\mu. && [1_{\Omega \setminus (A \cup B)} (g - f) > 0,\ \mu\big(\Omega \setminus (A \cup B)\big) > 0,\ \text{Lem. } 3.44]
\end{aligned}$$

Proof of Corollary 3.59

Assume that $f\colon (\Omega, \mathscr{A}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is measurable with a finite number of *positive* values $\alpha_1, \ldots, \alpha_m > 0$ and a finite number of *negative* values $\alpha_{m+1}, \ldots, \alpha_n < 0$. By convention, if $m = n$, then $\sum_{i=m+1}^{n} \alpha_i\, 1_{A_i} = 0$, and if $m = 0$, then $\sum_{i=1}^{m} \alpha_i\, 1_{A_i} = 0$. Then

$$f = \sum_{i=1}^{n} \alpha_i\, 1_{A_i} = \sum_{i=1}^{m} \alpha_i\, 1_{A_i} + \sum_{i=m+1}^{n} \alpha_i\, 1_{A_i}$$

and $f^+ = \sum_{i=1}^{m} \alpha_i\, 1_{A_i}$ as well as $f^- = -\sum_{i=m+1}^{n} \alpha_i\, 1_{A_i} = \sum_{i=m+1}^{n} -\alpha_i\, 1_{A_i}$. Therefore,

$$\int f^+ \, d\mu = \sum_{i=1}^{m} \alpha_i \, \mu(A_i) \quad \text{and} \quad \int f^- \, d\mu = \sum_{i=m+1}^{n} -\alpha_i \, \mu(A_i),$$

and $\int f^+ \, d\mu$ as well as $\int f^- \, d\mu$ are finite if and only if $\mu(A_i) < \infty$, for all $i = 1, \ldots, n$. Now, $\mu(A_i) < \infty$, for all $i = 1, \ldots, n$, implies

$$\begin{aligned}
\int f \, d\mu &= \int f^+ \, d\mu - \int f^- \, d\mu && [\text{Def. 3.28}] \\
&= \sum_{i=1}^{m} \alpha_i \, \mu(A_i) + \sum_{i=m+1}^{n} \alpha_i \, \mu(A_i) \\
&= \sum_{i=1}^{n} \alpha_i \, \mu(A_i) \\
&= \sum_{i=1}^{n} \alpha_i \, \mu_f(\{\alpha_i\}) && [\text{Def. 2.79, (2.44)}].
\end{aligned}$$

Proof of Theorem 3.68

(i) This is the proposition of Lemma 3.47.

(ii) This proposition is Theorem 3.48.

(iii) If f, g are μ-integrable and nonnegative, then it suffices to show: (b) $\Leftrightarrow$ (c) [see (ii)]. Now

$$\begin{aligned}
&\int_A f \, d\mu = \int_A g \, d\mu, \quad \forall A \in \mathscr{A} \\
\Leftrightarrow\ &\int 1_A f \, d\mu = \int 1_A g \, d\mu, \quad \forall A \in \mathscr{A} && [\text{Def. 3.35}] \\
\Leftrightarrow\ &\int 1_A \, df \odot \mu = \int 1_A \, dg \odot \mu, \quad \forall A \in \mathscr{A} && [\text{Th. 3.67 (i)}] \\
\Leftrightarrow\ &f \odot \mu(A) = g \odot \mu(A), \quad \forall A \in \mathscr{A} && [(3.9)] \\
\Leftrightarrow\ &f \odot \mu = g \odot \mu.
\end{aligned}$$

(iv) If f, g are μ-integrable and nonnegative, then the equivalence of (a), (b), and (c) follows from (ii) and (iii). Hence, it suffices to show: (c) $\Leftrightarrow$ (d). Because μ-integrability of f and g implies that $f \odot \mu$ and $g \odot \mu$ are finite measures, applying Theorem 1.71 completes the proof.

Proof of Corollary 3.77

Note that

$$\forall (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 \colon 1_C(\omega_1, \omega_2) = 1_{C_{\omega_1}}(\omega_2) = 1_{C_{\omega_2}}(\omega_1). \tag{3.85}$$

Now,

$$\begin{aligned}
\mu_1 \otimes \mu_2(C) &= \int 1_C d(\mu_1 \otimes \mu_2) && [(3.9)]\\
&= \int\int 1_C(\omega_1,\omega_2)\, \mu_2(d\omega_2)\, \mu_1(d\omega_1) && [(3.80)]\\
&= \int\int 1_{C_{\omega_1}}(\omega_2)\, \mu_2(d\omega_2)\, \mu_1(d\omega_1) && [(3.85)]\\
&= \int \mu_2(C_{\omega_1})\, \mu_1(d\omega_1). && [(3.9)]
\end{aligned}$$

The proof of the second equation is analog.

3.8 Exercises

▷ **Exercise 3-1** Construct a representation of the identity function on $\Omega = \{1,\dots,n\}$ as a weighted sum of indicators of elements of $\mathscr{A}$, where $n \in \mathbb{N}$ and $(\Omega,\mathscr{A})$ with $\mathscr{A} = \mathscr{P}(\Omega)$.

▷ **Exercise 3-2** Prove that, for every nonnegative step function, there exists a normal representation (see Rem. 3.8).

▷ **Exercise 3-3** Consider the measure space $(\Omega,\mathscr{A},\lambda)$, where λ is the Lebesgue measure. Show that the number $\sum_{i=1}^n \alpha_i\, \lambda(A_i)$ assigned to the nonnegative step function f defined in Example 3.7 is identical for the four specified representations of f.

▷ **Exercise 3-4** Let $(\Omega,\mathscr{A})$ be a measurable space and let $A \in \mathscr{A}$. Show that if $f = \sum_{i=1}^n \alpha_i 1_{A_i}$ is a normal representation of a nonnegative step function, then the product $1_A f$ of the indicator 1_A and f is also a normal representation of a nonnegative step function, and $1_A f = \sum_{i=1}^n \alpha_i 1_{A\cap A_i}$.

▷ **Exercise 3-5** Prove Equation (3.8).

▷ **Exercise 3-6** Compute the integral of the identity mapping $id\colon \Omega \to \Omega$ with respect to the counting measure $\mu_\#$ on $\mathscr{P}(\Omega)$, where $\Omega = \{1,\dots,n\}$. Then look at it for $n = 5$.

▷ **Exercise 3-7** Compute the integrals $\int f_1\, d\lambda$ and $\int f_2\, d\lambda$ of the functions f_1 and f_2 defined in Example 3.17.

▷ **Exercise 3-8** Prove the propositions of Example 3.39.

▷ **Exercise 3-9** Show that $1_A f \underset{\mu}{=} 0$ if $\mu(A) = 0$.

▷ **Exercise 3-10** Prove Equation (3.52).

▷ **Exercise 3-11** Compute the integral of the function $f(x) = 10 - x^2$ considered in Example 3.64.

▷ **Exercise 3-12** Prove the proposition of Remark 3.71.

Solutions

▷ **Solution 3-1** The identity function $id\colon \Omega \to \Omega$ on $\Omega = \{1,\dots,n\}$ is defined by

$$id(i) \;=\; i, \quad \forall i \in \Omega.$$

According to Example 2.9 it is $(\mathscr{A},\mathscr{A}_0)$-measurable for all σ-algebras $\mathscr{A}_0 \subset \mathscr{A}$. Now consider the set $\{1,\dots,n\}$ of values of id and the partition $\big\{\{1\},\dots,\{n\}\big\}$ of Ω. Then

$$id \;=\; \sum_{i=1}^{n} i \cdot 1_{\{i\}} \;=\; 1\cdot 1_{\{1\}} + \ldots + n\cdot 1_{\{n\}}.$$

▷ **Solution 3-2** Let $f\colon (\Omega,\mathscr{A}) \to (\mathbb{R},\mathscr{B})$ be a nonnegative step function, with $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$. Define, for all nonempty $J \subset \{1,\dots,n\}$,

$$B_J \;:=\; \left(\bigcap_{i\in J} A_i\right) \cap \left(\bigcap_{i\notin J} A_i^c\right).$$

These are $2^n - 1$ sets, where several of them may be empty, and all are pairwise disjoint. Then

$$f \;=\; \sum_{J\colon B_J \neq \emptyset} \left(\sum_{i\in J} \alpha_i\right) \cdot 1_{B_J}$$

is a normal representation of f.

▷ **Solution 3-3** We compute the sum for all four representations $\sum_{i=1}^{n} \alpha_i 1_{A_i}$ of f. The first one is:

$$\begin{aligned}\sum_{i=1}^{4} \alpha_i\, \lambda(A_i) \;&=\; 2\cdot(1-0) \;+\; 5\cdot(2-1) \;+\; 4\cdot(3-2) \;+\; 1\cdot(4-3)\\ &=\; 2\cdot 1 \;+\; 5\cdot 1 \;+\; 4\cdot 1 \;+\; 1\cdot 1 \;=\; 12.\end{aligned}$$

The sum for the second representation of f is:

$$\begin{aligned}\sum_{i=1}^{5} \gamma_i\, \lambda(C_i) \;&=\; 2\cdot(.5-0) \;+\; 2\cdot(1-.5) \;+\; 5\cdot(2-1) \;+\; 4\cdot(3-2) \;+\; 1\cdot(4-3)\\ &=\; 2\cdot .5 \;+\; 2\cdot .5 \;+\; 5\cdot 1 \;+\; 4\cdot 1 \;+\; 1\cdot 1 \;=\; 12.\end{aligned}$$

The sum for the third representation of f is:

$$\begin{aligned}\sum_{i=1}^{3} \beta_i\, \lambda(B_i) \;&=\; 2\cdot(2-0) \;+\; 3\cdot(3-1) \;+\; 1\cdot(4-2)\\ &=\; 2\cdot 2 \;+\; 3\cdot 2 \;+\; 1\cdot 2 \;=\; 12.\end{aligned}$$

The sum for the fourth representation of f is:

$$\begin{aligned}\sum_{i=1}^{4} \delta_i\, \lambda(D_i) \;&=\; 1\cdot(4-0) \;+\; 1\cdot(3-0) \;+\; 2\cdot(3-1) \;+\; 1\cdot(2-1)\\ &=\; 1\cdot 4 \;+\; 1\cdot 3 \;+\; 2\cdot 2 \;+\; 1\cdot 1 \;=\; 12.\end{aligned}$$

Obviously, all four sums are identical.

▷ **Solution 3-4** Let $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$, where $A_1, \ldots, A_n \in \mathscr{A}$. This implies $A \cap A_1, \ldots, A \cap A_n \in \mathscr{A}$, for $A \in \mathscr{A}$. Therefore, and because of $1_A \cdot f = \sum_{i=1}^{n} \alpha_i 1_{A \cap A_i}$, the function $1_A \cdot f$ is a non-negative step function. If $f = \sum_{i=1}^{n} \alpha_i 1_{A_i}$ is a normal representation, then $A_i \cap A_j = \emptyset$ for $i \neq j$, which implies $(A \cap A_i) \cap (A \cap A_j) = A \cap (A_i \cap A_j) = A \cap \emptyset = \emptyset$, for $i \neq j$. Therefore, $1_A \cdot f = \sum_{i=1}^{n} \alpha_i 1_{A \cap A_i}$ is a normal representation as well.

▷ **Solution 3-5** $\int_A \alpha \, d\mu = \int \alpha 1_A \, d\mu = \alpha \mu(A)$ [see Eq. (3.4)].

▷ **Solution 3-6** Consider the elements $\{1\}, \ldots, \{n\}$ of $\mathscr{A} = \mathscr{P}(\Omega)$ and $id = \sum_{\omega=1}^{n} \omega \cdot 1_{\{\omega\}}$. According to Definition 3.10,

$$\begin{aligned} \int id \; d\mu_\# &= \sum_{\omega=1}^{n} \omega \cdot \mu_\#(\{\omega\}) \\ &= 1 \cdot \mu_\#(\{1\}) + 2 \cdot \mu_\#(\{2\}) + \ldots + n \cdot \mu_\#(\{n\}) \\ &= \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \end{aligned} \tag{3.86}$$

is the integral of id over Ω with respect to the measure $\mu_\#$. Hence, in this example, the integral $\int id \; d\mu_\#$ is the sum over all elements in Ω. For $n = 5$ this formula yields $\int id \; d\mu_\# = 15$.

▷ **Solution 3-7** In Example 3.17 we specified the measure space $(\mathbb{R}, \mathscr{B}, \lambda)$, where λ is the Lebesgue (or length) measure on $\mathscr{B}$. Remember, the Lebesgue measure satisfies

$$\lambda\big(]a,b]\big) = \lambda\big([a,b]\big) = b - a,$$

for $a < b$ [see Eq. (1.53)]. We also considered $f_1 = \alpha_1 1_{A_1}$ with $A_1 = [0, (1-.50)^{1/2}]$. Hence, $f_1 = .50 \cdot 1_{A_1}$. Therefore,

$$\int f_1 \, d\lambda = \alpha_1 \cdot \lambda(A_1) = .50 \cdot \lambda(A_1) = .50 \cdot (1-.50)^{1/2} = .50 \cdot .50^{1/2} \approx .3536.$$

This is the area shaded in the left part of Figure 3.3.

Similarly, in Example 3.17 we also considered $f_2 = \sum_{i=1}^{3} \beta_i 1_{B_i}$ with the three intervals

$$B_1 = [0, (1-.75)^{1/2}], \quad B_2 = \big](1-.75)^{1/2}, (1-.50)^{1/2}\big], \quad B_3 = \big](1-.50)^{1/2}, (1-.25)^{1/2}\big].$$

Again note that B_1, B_2, B_3 is a sequence of elements of $\mathscr{A}$. Furthermore, $f_2 = \sum_{i=1}^{3} \beta_i 1_{B_i}$ with $\beta_1 = .75$, $\beta_2 = .50$, and $\beta_3 = .25$. Hence, the integral of $f_2 = \sum_{i=1}^{3} \beta_i 1_{B_i}$ with respect to λ is

$$\begin{aligned} \int f_2 \, d\lambda &= \sum_{i=1}^{3} \beta_i \cdot \lambda(B_i) = .75 \cdot \lambda(B_1) + .50 \cdot \lambda(B_2) + .25 \cdot \lambda(B_3) \\ &= .75 \cdot [(1-.75)^{1/2}] + .50 \cdot [(1-.50)^{1/2} - (1-.75)^{1/2}] + .25 \cdot [(1-.25)^{1/2} - (1-.50)^{1/2}] \\ &\approx .75 \cdot .50 + .50 \cdot .2071 + .25 \cdot .1589 \approx 0.3750 + .1036 + .0397 = .5183. \end{aligned}$$

This is the area shaded in the middle part of Figure 3.3. The integral of $f_3 = \sum_{i=1}^{7} \gamma_i 1_{C_i}$ can be computed correspondingly. It is the area shaded in the right part of Figure 3.3.

▷ **Solution 3-8** If $f\colon (\Omega, \mathscr{A}, \mu) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is μ-integrable and $A, B \subset \mathscr{A}$, then $1_{A \cup B} f$ is μ-integrable (see Lemma 3.33) and

$$\begin{aligned}
\int_{A\cup B} f\,d\mu &= \int 1_{A\cup B} f\,d\mu && [(3.30)]\\
&= \int (1_A + 1_B - 1_{A\cap B})\, f\,d\mu && [(1.33)]\\
&= \int (1_A f + 1_B f - 1_{A\cap B} f)\,d\mu\\
&= \int 1_A f\,d\mu + \int 1_B f\,d\mu - \int 1_{A\cap B} f\,d\mu && [(3.34)]\\
&= \int_A f\,d\mu + \int_B f\,d\mu - \int_{A\cap B} f\,d\mu. && [(3.30)]
\end{aligned}$$

If $A\cap B=\emptyset$ and $f\colon(\Omega,\mathscr{A},\mu)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ is quasi-μ-integrable, then $1_{A\cup B}\cdot f = 1_A\cdot f + 1_B\cdot f$, and these three functions are quasi-μ-integrable (see Lem. 3.33). If $\int f^+ d\mu$ is finite, then $\int 1_{A\cup B}\cdot f^+ d\mu$, $\int 1_A\cdot f^+ d\mu$, and $\int 1_B\cdot f^+ d\mu$ are finite as well (see Lem. 3.33). If $\int f^+ d\mu$ is infinite, then quasi-μ-integrability of f implies that $\int f^- d\mu$ is finite and, according to Lemma 3.33, also the integrals $\int 1_{A\cup B}\cdot f^- d\mu$, $\int 1_A\cdot f^- d\mu$, and $\int 1_B\cdot f^- d\mu$. Hence, in both cases

$$\begin{aligned}
\int 1_{A\cup B}\cdot f\,d\mu &= \int 1_{A\cup B}\cdot f^+ d\mu - \int 1_{A\cup B}\cdot f^- d\mu && [(3.27)]\\
&= \int 1_A\cdot f^+ d\mu + \int 1_B\cdot f^+ d\mu - \left(\int 1_A\cdot f^- d\mu + \int 1_B\cdot f^- d\mu\right) && [(3.34)]\\
&= \left(\int 1_A\cdot f^+ d\mu - \int 1_A\cdot f^- d\mu\right) + \left(\int 1_B\cdot f^+ d\mu - \int 1_B\cdot f^- d\mu\right)\\
&= \int 1_A\cdot f\,d\mu + \int 1_B\cdot f\,d\mu. && [(3.27)]
\end{aligned}$$

▷ **Solution 3-9**

$$1_A(\omega)\cdot f(\omega) = \begin{cases} 0, & \text{if } \omega\notin A\\ f(\omega), & \text{if } \omega\in A.\end{cases}$$

Therefore, $\{\omega\in\Omega\colon 1_A(\omega)\cdot f(\omega)\neq 0\}\subset A$. Hence, $\mu\big(\{\omega\in\Omega\colon 1_A(\omega)\cdot f(\omega)\neq 0\}\big)\le\mu(A)=0$.

▷ **Solution 3-10** Let $\alpha\ge 0$ and remember that the measure $\alpha\mu$ on $(\Omega,\mathscr{A})$ is defined by $(\alpha\mu)(A))=\alpha\mu(A)$ for all $A\in\mathscr{A}$. The proof is conducted in three steps: (a) for a nonnegative step function, (b) for a nonnegative numerical measurable function, and (c) for an μ-integrable numerical function (see Rem. 3.30).

(a) If f is a nonnegative step function and $f=\sum_{i=1}^n \alpha_i 1_{A_i}$ a normal representation, then

$$\int f\,d(\alpha\mu) = \sum_{i=1}^n \alpha_i\,(\alpha\mu)(A_i) = \alpha\sum_{i=1}^n \alpha_i\,\mu(A_i) = \alpha\int f\,d\mu.$$

(b) If f is a nonnegative numerical measurable function and f_i, $i\in\mathbb{N}$, is an increasing sequence of nonnegative step functions with $\lim_{i\to\infty} f_i = f$, then

$$\begin{aligned}
\int f\,d(\alpha\mu) &= \lim_{i\to\infty}\int f_i\,d(\mu\alpha)\\
&= \lim_{i\to\infty}\alpha\int f_i\,d\mu && [(a)]\\
&= \alpha\lim_{i\to\infty}\int f_i\,d\mu = \alpha\int f\,d\mu.
\end{aligned}$$

(c) If f an μ-integrable numerical function, then

$$\begin{aligned}\int f\, d(\alpha\mu) &= \int f^+\, d(\alpha\mu) - \int f^-\, d(\alpha\mu)\\ &= \alpha\int f^+\, d\mu - \alpha\int f^-\, d\mu \qquad [(b)]\\ &= \alpha\int f\, d\mu.\end{aligned}$$

▷ **Solution 3-11** Because the derivative $g'(x)$ of a function $g(x) = \alpha+\beta x+\gamma x^n$, $\alpha,\beta,\gamma\in\mathbb{R}$, $n\in\mathbb{N}$, is $g'(x) = \beta+\gamma n x^{n-1}$, the indefinite integral of $f(x)$ is

$$\int f(x)\, dx = F(x) = 10x - \frac{x^3}{3} + c, \quad c\in\mathbb{R},$$

and therefore,

$$\int_a^b f(x)\, dx = F(x)\Big|_a^b := F(b)-F(a) = \left(10x-\frac{x^3}{3}\right)\Big|_a^b.$$

For $a=-5$ and $b=5$, this equation yields

$$\int_{-5}^{5} f(x)\, dx = \left(10x-\frac{x^3}{3}\right)\Big|_{-5}^{5} = \left(50-\frac{125}{3}\right) - \left(-50+\frac{125}{3}\right) = 100 - \frac{250}{3} = 16.\bar{6}.$$

Hence,

$$\int_{[-5,5]} f\, d\lambda = \int_{-5}^{5} f(x)\, dx = 16.\bar{6}.$$

▷ **Solution 3-12** Let $A\in\mathscr{A}$ with $\mu(A)=0$ and let f be a density f of ν with respect to μ, i. e., $\nu = f\odot\mu$. Then

$$\begin{aligned}\nu(A) &= \int 1_A d\nu = \int 1_A d\, f\odot\mu \qquad [(3.8),\ \nu = f\odot\mu]\\ &= \int 1_A\cdot f\, d\mu = 0. \qquad [(3.72),\ (3.42)]\end{aligned}$$

Hence, $\nu \underset{\mathscr{A}}{\ll} \mu$.

Part II
Probability, Random Variable and its Distribution

Chapter 4
Probability Measure

In chapter 1 we introduced the concept of a *measure*, treated various examples of measures, and some of their properties. In this chapter, we turn to a special class of examples, called *probability measures*. We start with the definition of a probability measure, then turn to conditional probabilities and the most important theorems related to conditional probability: the *multiplication rule*, the *theorem of total probability*, and *Bayes' Theorem*. Furthermore, we introduce the concept of a *conditional-probability measure*. Next, we define *independence of events* and of *independence of sets of events* with respect to a probability measure. A section on *conditional independence given an event* concludes this chapter.

4.1 Probability Measure and Probability Space

Now we introduce the concept of a *probability measure* as defined by Kolmogorov (1933/1977) (for the English version of this book see Kolmogorov, 1956). As we shall see, a probability measures is a special finite measure that is standardized.

4.1.1 Definition

Definition 4.1 (Probability Measure)
Let $(\Omega, \mathscr{A})$ *be a measurable space. Then the function* $P\colon \mathscr{A} \to [0,1]$ *is called a* probability measure on $(\Omega, \mathscr{A})$, *if the following conditions hold:*

(a) $P(\Omega) = 1$ *(standardization).*
(b) $P(A) \geq 0,\ \forall\, A \in \mathscr{A}$ *(nonnegativity).*
(c) $A_1, A_2, \ldots \in \mathscr{A}$ *are pairwise disjoint* $\Rightarrow P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$ *(σ-additivity).*

Remark 4.2 (Probability and Probability Space) Let P be a probability measure on $(\Omega, \mathscr{A})$. Then the triple $(\Omega, \mathscr{A}, P)$ is called a *probability space* and a value $P(A)$ of P is called the *probability* of A. ◁

Remark 4.3 (Elementary Event and Event) Let $(\Omega, \mathscr{A}, P)$ be a probability space. Then $A \in \mathscr{A}$ is called an *event*, and a singleton $\{\omega\}$, $\omega \in \Omega$, is called an *elemen-*

tary event, if $\{\omega\} \in \mathscr{A}$. Note the distinction between an *outcome* $\omega \in \Omega$ and an *elementary event* $\{\omega\} \in \mathscr{A}$ (see Exercise 4-1). Also note that the term *event* is only used in the context of a probability space $(\Omega, \mathscr{A}, P)$. Otherwise, $A \in \mathscr{A}$ is called a *measurable set*. ◁

Remark 4.4 (No Time Order Between Events) The intuitive concept of an event often implies that events are ordered with respect to time. That is, one event is prior, simultaneous, or posterior to another event. In contrast, events as defined in probability theory are not necessarily ordered with respect to time. ◁

Remark 4.5 (A Priori Perspective) In probability theory we consider random experiments from the *a priori perspective*. Hence, outcomes of a random experiment and events are considered *before* they happen. Only then it makes sense to talk about the probability of an event. Even if an event already happened, we do *as if* it did not happen when we talk about its probability. ◁

4.1.2 Properties of a Probability Measure

Comparing conditions (a) to (c) of the definition of a probability measure to the conditions defining a measure (see Def. 1.43) shows that (b) and (c) are identical; only condition (a) differs. However, $P(\Omega) = 1$ implies $P(\emptyset) = 0$, because σ-additivity of P yields $P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$. Therefore, $P(\emptyset) = P(\Omega) - P(\Omega) = 0$. This proves the following corollary:

Corollary 4.6 (A Probability Measure is a Measure)
A measure P on $(\Omega, \mathscr{A})$ is a probability measure on $\mathscr{A}$ if and only if $P(\Omega) = 1$.

A direct implication of this corollary is that all rules of computation for a measure (see Box 1.1) also hold for a probability measure. For convenience, these rules are explicitly formulated for probability measures in Box 4.1 using the additional property $P(\Omega) = 1$.

Remark 4.7 (Distribution vs. Probability Measure) A probability measure on $(\Omega, \mathscr{A})$ is also called a *distribution on* $(\Omega, \mathscr{A})$. Although this term is preferably used in the context of a random variable (see Def. 5.3), the term 'distribution' is well-defined without referring to a random variable. ◁

4.1.3 Examples

Example 4.8 (Continuous Uniform Distribution) Let $\mathscr{B}_2$ denote the Borel σ-algebra on $\mathbb{R}^2$ and consider a probability space $(\Omega, \mathscr{A}, P)$, where $\Omega \in \mathscr{B}_2$, $\mathscr{A} = \mathscr{B}_2|_\Omega := \{\Omega \cap A \colon A \in \mathscr{B}_2\}$ is the trace of $\mathscr{B}_2$ in Ω (see Example 1.10). Furthermore, let λ_2 denote the Lebesgue measure on $(\mathbb{R}^2, \mathscr{B}_2)$, assume $0 < \lambda_2(\Omega) < \infty$, and define

$$P(A) = \frac{\lambda_2(A)}{\lambda_2(\Omega)}, \quad \forall\, A \in \mathscr{A}. \tag{4.1}$$

Box 4.1 Rules of Computation for Probabilities

Let $(\Omega, \mathscr{A}, P)$ be a probability space.

If $A_1, A_2, \ldots \in \mathscr{A}$ are pairwise disjoint, then

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i) \qquad (\sigma\text{-}additivity) \qquad \text{(i)}$$

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i), \quad \forall n \in \mathbb{N}. \qquad (finite\ additivity) \qquad \text{(ii)}$$

If $A, B \in \mathscr{A}$, then:

$$P(A) = P(A \cap B) + P(A \setminus B) \qquad \text{(iii)}$$

$$P(A^c) = 1 - P(A) \qquad \text{(iv)}$$

$$P(A) \leq P(B), \quad \text{if } A \subset B \qquad (monotonicity) \qquad \text{(v)}$$

$$P(A \setminus B) = P(A) - P(A \cap B) \qquad \text{(vi)}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \text{(vii)}$$

$$P(A) = 1 \;\Rightarrow\; P(A \cap B) = P(B) \qquad \text{(viii)}$$

$$P(A) = 0 \;\Rightarrow\; P(A \cup B) = P(B). \qquad \text{(ix)}$$

Let $A \in \mathscr{A}$ and let $\Omega_0 \subset \Omega$ be finite or countable with $P(\Omega_0) = 1$.
If, for all $\omega \in \Omega_0$, $\{\omega\} \in \mathscr{A}$, then

$$P(A) = \sum_{\omega \in A \cap \Omega_0} P(\{\omega\}). \qquad \text{(x)}$$

If $A_1, A_2, \ldots \in \mathscr{A}$, then

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} P(A_i). \qquad (\sigma\text{-}subadditivity) \qquad \text{(xi)}$$

Then P is the *continuous uniform distribution* over Ω. The relative size of the set $A \in \mathscr{A}$ represents the probability $P(A)$, and Figure 4.1 can be used to illustrate some of its properties, e. g., Rules (iii) to (ix) of Box 4.1. This example will be generalized in section 8.2.1. ◁

Example 4.9 (Joe and Ann With Random Assignment – continued) In the example presented in Table 4.1, the probability measure P on $\mathscr{A} = \mathscr{P}(\Omega)$ is specified by the probabilities of the eight elementary events $\{\omega\} \in \Omega$, where

$$\Omega = \big\{(Joe, no, -), (Joe, no, +), \ldots, (Ann, yes, +)\big\}$$

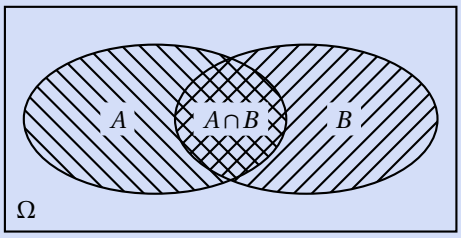


Figure 4.1. An Example in which probabilities are proportional to areas

Table 4.1. Joe and Ann With Random Assignment: Probability Measures

Elements of Ω			Probability measures			
Unit	Treatment	Success	$P(\{\omega\})$	$P^B(\{\omega\})$	$P^A(\{\omega\})$	$P^{A^c}(\{\omega\})$
(*Joe*,	*no*,	−)	.09	0	.18	0
(*Joe*,	*no*,	+)	.21	0	.42	0
(*Joe*,	*yes*,	−)	.04	.10	.08	0
(*Joe*,	*yes*,	+)	.16	.40	.32	0
(*Ann*,	*no*,	−)	.24	0	0	.48
(*Ann*,	*no*,	+)	.06	0	0	.12
(*Ann*,	*yes*,	−)	.12	.30	0	.24
(*Ann*,	*yes*,	+)	.08	.20	0	.16

Note. P, P^B, P^A, and P^{A^c} are probability measures on $(\Omega, \mathscr{A})$.

(see Exercise 4-2). Except for the empty set, which has probability $P(\emptyset) = 0$, all $2^8 = 256$ elements of $\mathscr{A}$ are either one of the eight elementary events $\{(Joe, no, -)\}$, $\{(Joe, no, +)\}, \ldots, \{(Ann, yes, +)\}$ or a union of some of these elementary events. Note that elementary events are always *disjoint*, i. e., $\{\omega_i\} \cap \{\omega_j\} = \emptyset$, if $\omega_i \neq \omega_j$. Therefore, the probabilities of their unions can easily be computed using finite additivity of the probability measure [see Rule (ii) of Box 4.1]. In order to illustrate this point, consider the event that *Joe is drawn*,

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\},$$

and the event that *the drawn person is successful*,

$$C = \{(Joe, no, +), (Joe, yes, +)\}, (Ann, no, +), (Ann, yes, +)\}.$$

The event A has the probability

$$\begin{aligned} P(A) &= P[\{(Joe, no, -)\}] + P[\{(Joe, no, +)\}] + P[\{(Joe, yes, -)\}] + P[\{(Joe, yes, +)\}] \\ &= .09 + .21 + .04 + .16 = .50. \end{aligned}$$

Similarly, the event C has the probability

$$\begin{aligned} P(C) &= P[\{(Joe, no, +)\}] + P[\{(Joe, yes, +)\}] + P[\{(Ann, no, +)\}] + P[\{(Ann, yes, +)\}] \\ &= .21 + .16 + .06 + .08 = .51, \end{aligned}$$

and the event *Joe is drawn and is successful*, $A \cap C = \{(Joe, no, +), (Joe, yes, +)\}$, has the probability

$$P(A \cap C) = P[\{(Joe, no, +)\}] + P[\{(Joe, yes, +)\}] = .21 + .16 = .37.$$

The probability measures specified in the last three columns of Table 4.1 will be treated in Examples 4.29 and 4.30. ◁

Example 4.10 (Finite Mixture of Probability Measures) In Example 1.61 we already noted that the weighted sum of measures on a measurable space $(\Omega, \mathscr{A})$ is again a measure on $(\Omega, \mathscr{A})$. With an additional assumption this also applies to probability measures. More precisely, if $P_1, \ldots, P_n$ are probability measures on $(\Omega, \mathscr{A})$, $\alpha_i \geq 0$, $i = 1, \ldots, n$, and we additionally assume $\sum_{i=1}^n \alpha_i = 1$, then $\sum_{i=1}^n \alpha_i P_i$ is again a probability measure on $(\Omega, \mathscr{A})$. It is called a *finite mixture* of $P_1, \ldots, P_n$. Such a finite mixture of probability measures will be illustrated by Example 4.30 using conditional-probability measures. ◁

Remark 4.11 (Other Examples) Another example with finite Ω is the *binomial distribution* (see Def. 8.7). Note that, even if Ω is infinite and countable, there are probability measures on $\big(\Omega, \mathscr{P}(\Omega)\big)$ with $P(\{\omega_i\}) > 0$, for all $\omega_i \in \Omega$, and

$$P(\Omega) = \sum_{\omega_i \in \Omega} P(\{\omega_i\}) = \sum_{i=1}^{\infty} P(\{\omega_i\}) = 1.$$

Examples in case are the *Poisson distribution* (see Def. 8.14) and the *geometric distribution* (see Def. 8.20). In all three examples, the probability measure P on $\big(\Omega, \mathscr{P}(\Omega)\big)$ is uniquely defined, if the probabilities $P(\{\omega_i\})$ are determined for all $\omega_i \in \Omega$ [see Box 4.1 (x) for $\Omega_0 = \Omega$]. ◁

4.2 Conditional Probability

Conditional probabilities can be used to describe *dependencies* between two events $A, B \in \mathscr{A}$ with respect to a probability measure P on $\mathscr{A}$. In section 4.2.6 we will also use this concept in order to introduce the concept of a *conditional-probability measure*.

4.2.1 Definition

Definition 4.12 (Conditional Probability)
Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $A, B \in \mathscr{A}$, and let $P(B) > 0$. Then

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)} \tag{4.2}$$

is called the conditional probability of A given B with respect to P.

Remark 4.13 (Continuous Uniform Distribution – continued) In Example 4.8 we defined the continuous uniform distribution on $(\Omega, \mathscr{A})$ by Equation (4.1). Using the area of the ellipses presented in Figure 4.1 (p. 130), the conditional probability $P(A \mid B)$ corresponds to the area of the intersection $A \cap B$ divided by the area of B. ◁

Example 4.14 (Flipping a Coin two Times) Consider the random experiment of flipping a coin two times. The conditional probability $P(B \mid A)$ that we flip *heads* in the second flip (B) given that we flip *heads* in the first flip (A) is $1/2$, which is equal to the *unconditional* probability $P(B)$ of flipping *heads* in the second flip. In such a case the two events A and B are *independent* (see section 4.3). Note that the conditional probability $P(A \mid B)$ that we flip *heads* in the first flip (A) given that we flip *heads* in the second flip (B) is also equal to the unconditional probability $P(A)$ of flipping *heads* in the first flip. This shows that we may condition on events that occur later in time and that a conditional probability does not necessarily a causal dependency.

As another example consider the event *flipping at least one heads* (A) and the event *no heads are flipped in the first flip* (B). In this case

$$P(A \mid B) = \frac{1}{2} \neq P(A) = \frac{3}{4},$$

and the two events are not independent (see section 4.3). ◁

Example 4.15 (Joe and Ann With Random Assignment – continued) Consider again Table 4.1 (p. 130), define $\Omega_U := \{Joe, Ann\}$ and $\Omega_X := \{yes, no\}$, and let

$$C = \Omega_U \times \Omega_X \times \{+\} = \{(Joe, no, +), (Joe, yes, +)\}, (Ann, no, +), (Ann, yes, +)\}$$

be the event that the *drawn person is successful.* Furthermore, let

$$B := \Omega_U \times \{yes\} \times \Omega_Y = \{(Joe, yes, -), (Joe, yes, +)\}, (Ann, yes, -), (Ann, yes, +)\}$$

denote the event that the *drawn person is treated.* Then Equation (4.2) yields:

$$P(C \mid B) = \frac{P(C \cap B)}{P(B)} = \frac{P(\Omega_U \times \{yes\} \times \{+\})}{P(\Omega_U \times \{yes\} \times \Omega_Y)} = \frac{.16 + .08}{.04 + .16 + .12 + .08} = .60.$$

Conditioning on the event B^c that the *drawn person is not treated* yields

$$P(C \mid B^c) = \frac{P(C \cap B^c)}{P(B^c)} = \frac{P(\Omega_U \times \{no\} \times \{+\})}{P(\Omega_U \times \{no\} \times \Omega_Y)} = \frac{.21 + .06}{.09 + .21 + .24 + .06} = .45.$$

In this example, the difference $P(C \mid B) - P(C \mid B^c) = .60 - .45$ can be used to evaluate the effect of the treatment. This will be substantiated in more detail in Example 4.31. ◁

4.2.2 Multiplication Rule

Now we treat some theorems involving conditional probabilities. The first one shows how the probability $P(A_1 \cap \ldots \cap A_n)$ can be factorized into a product of an unconditional probability and conditional probabilities.

Remark 4.16 (Multiplication Rule for Two and for Three Events) For two events A_1 and A_2, the multiplication rule is

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 \mid A_1), \tag{4.3}$$

provided that $P(A_1) > 0$. This equation directly follows from the definition of the conditional probability $P(A_2 \mid A_1)$. For three events A_1, A_2, and A_3, the multiplication rule is

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2), \tag{4.4}$$

provided that $P(A_1 \cap A_2) > 0$. This equation follows from the definition of the conditional probability

$$P(A_3 \mid A_1 \cap A_2) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)}, \tag{4.5}$$

inserting Equation (4.3) for $P(A_1 \cap A_2)$, and solving the resulting equation for $P(A_1 \cap A_2 \cap A_3)$. ◁

For n events $A_1, \ldots, A_n$, the multiplication rule is formulated in the following theorem.

Theorem 4.17 (Multiplication Rule)
Let $(\Omega, \mathscr{A}, P)$ be a probability space and $A_1, \ldots, A_n \in \mathscr{A}$, where $2 \leq n \in \mathbb{N}$. If $P(\bigcap_{i=1}^{n-1} A_i) > 0$, then

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \cdot \prod_{j=2}^{n} P\left(A_j \,\middle|\, \bigcap_{i=1}^{j-1} A_i\right). \tag{4.6}$$

(Proof p. 146)

Table 4.2. Joe and Ann With Self-Selection

Outcomes ω				Observables		
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y
(*Joe*, *no*, −)			.144	*Joe*	0	0
(*Joe*, *no*, +)			.336	*Joe*	0	1
(*Joe*, *yes*, −)			.004	*Joe*	1	0
(*Joe*, *yes*, +)			.016	*Joe*	1	1
(*Ann*, *no*, −)			.096	*Ann*	0	0
(*Ann*, *no*, +)			.024	*Ann*	0	1
(*Ann*, *yes*, −)			.228	*Ann*	1	0
(*Ann*, *yes*, +)			.152	*Ann*	1	1

Note. The probabilities of the elementary events are fictive

4.2.3 Examples

Example 4.18 (Joe and Ann With Self-Selection) Now we study a new example with Joe and Ann in order to illustrate the concepts introduced above. In this example, we fixed new probabilities of the elementary events (see Table 4.2).

Let

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\}$$

denote the event that Joe is drawn,

$$B = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}$$

the event that the drawn person is treated, and

$$C = \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\}$$

the event that there is success, irrespective of the drawn person and treatment received. Then

$$A \cap B \cap C = \{(Joe, yes, +)\}$$

is the event that Joe is drawn, receives the treatment, and is successful. According to Equation (4.4), the probability of this event can be computed by

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \cdot P(B \mid A) \cdot P(C \mid A \cap B) \\ &= (.144 + .336 + .004 + .016) \cdot \frac{.004 + .016}{.144 + .336 + .004 + .016} \cdot \frac{.016}{.004 + .016} \\ &= .50 \cdot .04 \cdot .80 = .016 \end{aligned}$$

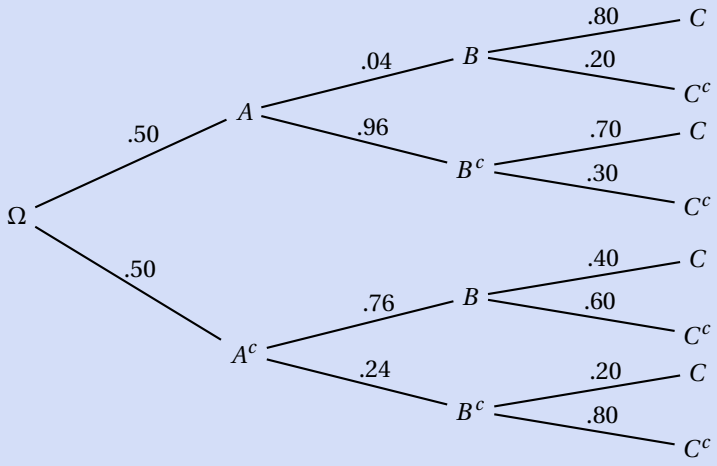


Figure 4.2. Probability tree illustrating the multiplication rule

(see Exercise 4-4). Of course, Equation (4.4) can also be applied to the other seven sets $A \cap B \cap C^c$ to $A^c \cap B^c \cap C^c$ in Figure 4.2. In this example, $P(A \cap B \cap C) = P\big(\{(Joe, yes, +)\}\big) = .016$ is already known (see Table 4.2). This is not the case in the following example. ◁

Example 4.19 (Drawing Three Balls) Consider drawing three balls without replacement from an urn containing two white balls and four black balls. Furthermore, let us consider the three events A_i to *draw a black ball at time* i, where $i = 1,2,3$. According to Theorem 4.17, the probability of drawing three black balls is

$$P(A_1 \cap A_2 \cap A_3) \;=\; P(A_1) \cdot P(A_2 \,|\, A_1) \cdot P(A_3 \,|\, A_1 \cap A_2),$$

where $P(A_1) = 4/6$, $P(A_2 \,|\, A_1) = 3/5$, and $P(A_3 \,|\, A_1 \cap A_2) = 2/4$. Hence,

$$P(A_1 \cap A_2 \cap A_3) \;=\; \frac{4}{6} \cdot \frac{3}{5} \cdot \frac{2}{4} \;=\; \frac{24}{120} \;=\; \frac{1}{5}.$$

◁

4.2.4 Theorem of Total Probability

In our next theorem, called the *theorem of total probability*, we show how the probability of an event $B \subset A_1 \cup \ldots \cup A_n$ can additively be decomposed into the products $P(B \,|\, A_i) \cdot P(A_i)$ of conditional and unconditional probabilities. In this theorem we assume that the events $A_1, \ldots, A_n$ are *pairwise disjoint*, i. e., we assume $A_i \cap A_j = \emptyset$, for all $i, j = 1, \ldots, n$, with $i \neq j$.

Theorem 4.20 (Theorem of Total Probability)
Let $(\Omega, \mathscr{A}, P)$ *be a probability space and* $B \in \mathscr{A}$.

(i) If

(a) $A_1, \ldots, A_n \in \mathscr{A}$ *are pairwise disjoint,*
(b) $B \subset \bigcup_{i=1}^{n} A_i$,

then

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i). \tag{4.7}$$

(ii) If (a) and (b) of (i) hold as well as

(c) $P(A_i) > 0, \quad \forall\, i = 1, \ldots, n,$

then

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i). \tag{4.8}$$

(iii) If

(a) $A_1, A_2, \ldots \in \mathscr{A}$ *are pairwise disjoint,*
(b) $B \subset \bigcup_{i=1}^{\infty} A_i$,

then

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i). \tag{4.9}$$

(iv) If (a) and (b) of (iii) hold as well as

(c) $P(A_i) > 0, \quad \forall\, i = 1, 2, \ldots,$

then

$$P(B) = \sum_{i=1}^{\infty} P(B \mid A_i) \cdot P(A_i). \tag{4.10}$$

(Proof p. 147)

Equation (4.7) can be illustrated by Figure 4.3. If we assume that $(\Omega, \mathscr{A}, P)$ is the probability space specified in Example 4.8, then the figure visualizes that $P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3)$. The crucial points are:

(a) If the events $A_1, \ldots, A_n$ are pairwise disjoint, then $B \cap A_1, \ldots, B \cap A_n$ are pairwise disjoint as well.

(b) The probability measure P is additive.

Equation (4.8) then follows using the factorization of each of the joint probabilities $P(B \cap A_i)$ into the product $P(B \mid A_i) \cdot P(A_i),\ i = 1, \ldots, n$, (see Th. 4.17).

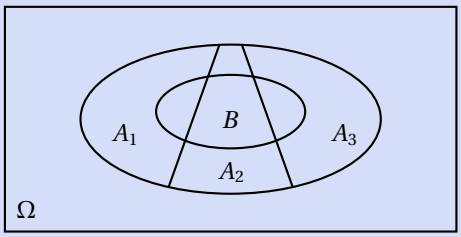


Figure 4.3. Venn diagram illustrating a partition of a set

4.2.5 Bayes' Theorem

Our next theorem, called *Bayes' theorem*, reveals how the conditional probabilities $P(B \mid A_i)$ are related to the conditional probabilities $P(A_i \mid B)$. Using the definitions of the conditional probabilities $P(A_i \mid B)$ and $P(B \mid A_i)$ yields

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)}. \tag{4.11}$$

Inserting Equation (4.8) for $P(B)$ then proves the following theorem.

Theorem 4.21 (Bayes' Theorem)
Let $(\Omega, \mathscr{A}, P)$ be a probability space, $B \in \mathscr{A}$, and $P(B) > 0$. Under the assumptions (a) to (c) of Theorem 4.20 (i) and (ii),

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) \cdot P(A_j)}, \quad \forall\, i = 1, \ldots, n. \tag{4.12}$$

Analogously, under the assumptions (a) to (c) of Theorem 4.20 (iii) and (iv),

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j) \cdot P(A_j)}, \quad \forall\, i \in \mathbb{N}. \tag{4.13}$$

Example 4.22 (Joe and Ann With Random Assignment – continued) Let

$$A = \{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\}$$

denote the event that Joe is drawn,

$$A^c = \{(Ann, no, -), (Ann, no, +), (Ann, yes, -), (Ann, yes, +)\}$$

the event that Ann is drawn, and

$$B = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -), (Ann, yes, +)\}$$

the event that the *drawn person is treated*. Then

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c)}$$
$$= \frac{.40 \cdot .50}{.40 \cdot .50 + .40 \cdot .50} = .50$$

is the conditional probability that *Joe is drawn* given that the *drawn person is treated* (see Table 4.1, p. 130). The corresponding probability that *Ann is drawn* given that the *drawn person is treated* is identical in this example, i. e., $P(A^c \mid B) = .50$. Hence, given treatment, each person has the same probability to be drawn. This is the *sampling perspective* of a randomized experiment supplementing the *assignment perspective*, according to which the treatment probability is the same for each person, i. e., $P(B \mid A) = P(B \mid A^c) = .40$ (see again Table 4.1). ◁

4.2.6 Conditional-Probability Measure

Just like probabilities, conditional probabilities of events $A \in \mathscr{A}$ given B are values of a probability measure.

Theorem 4.23 (Conditional-Probability Measure)
Let $(\Omega, \mathscr{A}, P)$ be a probability space. If $B \in \mathscr{A}$ and $P(B) > 0$, then the function $P^B \colon \mathscr{A} \to [0,1]$ defined by

$$P^B(A) = P(A \mid B), \quad \forall A \in \mathscr{A}, \tag{4.14}$$

is a probability measure on $(\Omega, \mathscr{A})$.

(Proof p. 147)

According to this theorem, for each $B \in \mathscr{A}$ with $P(B) > 0$, the triple $(\Omega, \mathscr{A}, P^B)$ is a probability space.

Definition 4.24 (Conditional-Probability Measure)
Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $B \in \mathscr{A}$, and let $P(B) > 0$. Then the function P^B defined by (4.14) is called the B-conditional-probability measure on $(\Omega, \mathscr{A})$.

In the following lemma we consider the relationship between conditional probabilities with respect to the measures P^B and P.

Lemma 4.25 (Conditional Probabilities With Respect to P^B)
Let $(\Omega, \mathscr{A}, P)$ be a probability space. If $A, B, C \in \mathscr{A}$ and $P(B \cap C) > 0$, then

$$P^B(A|C) \;=\; P(A|B \cap C). \tag{4.15}$$

(Proof p. 147)

Remark 4.26 (Total Conditional Probability) Suppose $A, B, C \in \mathscr{A}$, $P(B \cap C) > 0$, and $P(B \cap C^c) > 0$. This implies $P(B) > 0$ and $P^B(C) = P(C|B) = P(C \cap B)/P(B) > 0$. Applying Equation (4.8) to the measure P^B then yields

$$P^B(A) \;=\; P^B(A|C) \cdot P^B(C) \;+\; P^B(A|C^c) \cdot P^B(C^c), \tag{4.16}$$

and Equations (4.14) and (4.15) imply

$$P(A|B) \;=\; P(A|B \cap C) \cdot P(C|B) \;+\; P(A|B \cap C^c) \cdot P(C^c|B). \tag{4.17}$$

◁

According to the following lemma, P^B is *absolutely continuous with respect to* P. This is denoted by $P^B \underset{\mathscr{A}}{\ll} P$, and according to Definition 3.70 (i), it means

$$\forall\, A \in \mathscr{A}\colon\; P(A) = 0 \quad \Rightarrow \quad P^B(A) = 0. \tag{4.18}$$

In contrast, $P \underset{\mathscr{A}}{\ll} P^B$ does *not* always hold.

Lemma 4.27 (Absolute Continuity of the Conditional-Probability Measure)
Let $(\Omega, \mathscr{A}, P)$ be a probability space, $B \in \mathscr{A}$, and $P(B) > 0$. Then

(i) $P^B \underset{\mathscr{A}}{\ll} P$, *and*

(ii) $P^B = \left(\dfrac{1}{P(B)} \cdot 1_B\right) \odot P.$

(Proof p. 148)

Remark 4.28 (P^B is a Measure With Density) Proposition (ii) of Lemma 4.27 implies that P^B is a measure with density $1_B/P(B)$ with respect to P. The following equations show how $P^B(A)$ can be written as an integral in various ways:

$$\begin{aligned}
\forall\, A \in \mathscr{A}\colon\; P^B(A) \;&=\; \int_A dP^B && \text{[(3.8)]} \\
&=\; \int 1_A \, d\left(\frac{1}{P(B)} \cdot 1_B \odot P\right) && \text{[(3.30), Lem. 4.27 (ii)]} \\
&=\; \int 1_A \cdot 1_B \, \frac{1}{P(B)} \, dP && \text{[(3.72)]} \\
&=\; \frac{1}{P(B)} \cdot \int 1_{A \cap B} \, dP\,. && \text{[(1.32), (3.32)]}
\end{aligned} \tag{4.19}$$

Note that, according to Theorem 3.72 (i) and Remark 3.73, the density $1_B/P(B)$ can be written as a Radon-Nikodym derivate of P^B with respect to P, i. e.,

$$\frac{1}{P(B)} \cdot 1_B = \frac{dP^B}{dP}. \tag{4.20}$$

◁

Example 4.29 (Joe and Ann With Random Assignment – continued) Consider the example presented in Table 4.1 (p. 130). We specify the B-conditional-probability measure $P^B\colon \mathscr{A} \to [0,1]$ for the event that the *drawn person is treated*, i. e., for

$$B = \big\{(Joe, yes, -),\ (Joe, yes, +),\ (Ann, yes, -),\ (Ann, yes, +)\big\}.$$

For the first two elementary events, $P^B(\{(Joe, no, +)\}) = P^B(\{(Joe, no, -)\}) = 0$, because the intersections $\{(Joe, no, -)\} \cap B$ and $\{(Joe, no, +)\} \cap B$ are empty. For the next two elementary events, the B-conditional probabilities are

$$P^B\big(\{(Joe, yes, -)\}\big) = \frac{P\big(\{(Joe, yes, -)\} \cap B\big)}{P(B)} = \frac{.04}{.40} = .10$$

and

$$P^B\big(\{(Joe, yes, +)\}\big) = \frac{P\big(\{(Joe, yes, +)\} \cap B\big)}{P(B)} = \frac{.16}{.40} = .40.$$

For the next two elementary events, $P^B\big(\{(Ann, no, -)\}\big) = P^B\big(\{(Ann, no, +)\}\big) = 0$, because the intersections $\{(Ann, no, -)\} \cap B$ and $\{(Ann, no, +)\} \cap B$ are again empty. Finally, for the last two elementary events, the B-conditional probabilities are

$$P^B\big(\{(Ann, yes, -)\}\big) = \frac{P\big(\{(Ann, yes, -)\} \cap B\big)}{P(B)} = \frac{.12}{.40} = .30$$

and

$$P^B\big(\{(Ann, yes, +)\}\big) = \frac{P\big(\{(Ann, yes, +)\} \cap B\big)}{P(B)} = \frac{.08}{.40} = .20.$$

Except for $\emptyset$, all other events are unions of these elementary events. Because the elementary events are *disjoint*, the probabilities of their unions can easily be computed using finite additivity of the probability measure [see Rule (ii) of Box 4.1 and Exercise 4-5]. ◁

Example 4.30 (Joe and Ann With Random Assignment – continued) Two other conditional-probability measures on $(\Omega, \mathscr{A})$ are P^A and P^{A^c}, where A is the event

$$A = \{(Joe, no, -),\ (Joe, no, +),\ (Joe, yes, -),\ (Joe, yes, +)\}$$

that Joe is sampled and A^c the event

$$A^c = \{(Ann, no, -),\ (Ann, no, +),\ (Ann, yes, -),\ (Ann, yes, +)\}$$

that Ann is sampled. The values of these conditional-probability measures are presented in the last two columns of Table 4.1. These measures can also be used to illustrate a *mixture of two probability measures*. As is easily seen

$$P \;=\; .50 \cdot P^{A} \;+\; .50 \cdot P^{A^c},$$

i. e., the measure P is a mixture of the two conditional-probability measures P^A and P^{A^c} (see Examples 4.10 and 1.61). ◁

Example 4.31 (Joe and Ann With Random Assignment– continued) In Example 4.15 we computed the two conditional probabilities $P(C\,|\,B) = .60$ and $P(C\,|\,B^c) = .45$ of success given treatment and no treatment, respectively. These are conditional probabilities with respect to the measure P. Let us now consider the *individual treatment effects* of Joe and of Ann. These individual effects can be computed using the P^A- and P^{A^c}-conditional-probability measures, respectively. For Joe the individual treatment effect is

$$\begin{aligned} P^{A}(C\,|\,B) \;-\; P^{A}(C\,|\,B^c) \;&=\; \frac{P^{A}(\Omega_U \times \{yes\} \times \{+\})}{P^{A}(\Omega_U \times \{yes\} \times \Omega_Y)} \;-\; \frac{P^{A}(\Omega_U \times \{no\} \times \{+\})}{P^{A}(\Omega_U \times \{no\} \times \Omega_Y)} \\ &=\; \frac{.32+0}{.08+.32+0+0} \;-\; \frac{.32+0}{.18+.42+0+0} \\ &=\; .80 \;-\; .70 \;=\; .10, \end{aligned}$$

and for Ann it is

$$\begin{aligned} P^{A^c}(C\,|\,B) \;-\; P^{A^c}(C\,|\,B^c) \;&=\; \frac{P^{A^c}(\Omega_U \times \{yes\} \times \{+\})}{P^{A^c}(\Omega_U \times \{yes\} \times \Omega_Y)} \;-\; \frac{P^{A^c}(\Omega_U \times \{no\} \times \{+\})}{P^{A^c}(\Omega_U \times \{no\} \times \Omega_Y)} \\ &=\; \frac{.16+0}{.24+.16+0+0} \;-\; \frac{.12+0}{.48+.12+0+0} \\ &=\; .40 \;-\; .20 \;=\; .20. \end{aligned}$$

Hence, the treatment effect $P(C\,|\,B) - P(C\,|\,B^c) = .15$ (see Example 4.15), is just the weighted average $.50 \cdot .10 + .50 \cdot .20 = .15$ of the two individual treatment effects, where the weights are .50 for Joe and for Ann (see Example 4.30). Note that this property does not always hold. Table 4.2 (p. 134) displays an example with Joe and Ann, in which this property does *not* hold. In that example, all individual treatment effects are *positive*, whereas the difference $P(C\,|\,B) - P(C\,|\,B^c)$ is *negative*. Hence, in that example, the difference $P(C\,|\,B) - P(C\,|\,B^c)$ *can not* be used to evaluate the treatment effect. ◁

4.3 Independence

4.3.1 Independence of Events

Independence of two events A and B means that the conditional and unconditional probabilities are the same, i. e., $P(A\,|\,B) = P(A)$ and $P(B\,|\,A) = P(B)$. This

definition presupposes that $P(A), P(B) > 0$, because otherwise the two conditional probabilities are not defined. The following definition does not rest on this requirement and extends the concept of independence to more than two events.

Definition 4.32 (Independence of Events)
Let $(\Omega, \mathscr{A}, P)$ be a probability space.

(i) Two events $A, B \in \mathscr{A}$ are called P-independent, denoted $A \perp\!\!\!\perp_P B$, if

$$P(A \cap B) = P(A) \cdot P(B). \tag{4.21}$$

(ii) Let I be a nonempty set and let $A_i \in \mathscr{A}$, $i \in I$. Then $(A_i, i \in I)$ is called a family of P-independent events, denoted $\perp\!\!\!\perp_P (A_i, i \in I)$, if

$$P\Big(\bigcap_{i \in I_0} A_i\Big) = \prod_{i \in I_0} P(A_i), \quad \forall \text{ finite } I_0 \subset I. \tag{4.22}$$

Remark 4.33 (Pairwise and Triple-Wise Independence) For n events $A_1, \ldots, A_n$, P-independence will also be denoted by

$$\perp\!\!\!\perp_P A_1, \ldots, A_n.$$

For three events, for instance, it means that

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j), \quad i \neq j, \quad i, j = 1,2,3, \tag{4.23}$$

(*pairwise P-independence*) and

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) \tag{4.24}$$

(*triple-wise P-independence*) hold.

Note that pairwise P-independence of more than two events does not imply P-independence of these events. Furthermore, triple-wise P-independence, for instance, does not imply pairwise P-independence. For more propositions on independence of events see Box 4.2 (p. 145). ◁

Remark 4.34 (Independence of any event A with Ω and $\emptyset$) For any probability space $(\Omega, \mathscr{A}, P)$,

$$\forall A \in \mathscr{A}: \; A \perp\!\!\!\perp_P \Omega \quad \text{and} \quad A \perp\!\!\!\perp_P \emptyset. \tag{4.25}$$

(see Exercise 4-6). ◁

4.3.2 Independence of Set Systems

Now we extend the concept of P-independence to *set systems*, i. e., to sets of events, and illustrate independence by an example.

Definition 4.35 (Family of Independent Set Systems)
Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathcal{E}_i \subset \mathcal{A}$, $i \in I$. Then $(\mathcal{E}_i, i \in I)$ is called a family of P-independent set systems, denoted $\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I)$, if $\perp\!\!\!\perp_P (A_i, i \in I)$ holds for all families $(A_i, i \in I)$ with $A_i \in \mathcal{E}_i$, $i \in I$. If $I = \{1,2\}$ we also use the notation $\mathcal{E}_1 \perp\!\!\!\perp_P \mathcal{E}_2$ instead of $\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I)$.

Remark 4.36 (Independence of an Event and a Set System) Let $(\Omega, \mathcal{A}, P)$ be a probability space. An event $A \in \mathcal{A}$ and a set system $\mathcal{E} \subset \mathcal{A}$ are called *P-independent*, denoted $A \perp\!\!\!\perp_P \mathcal{E}$, if $\{A\} \perp\!\!\!\perp_P \mathcal{E}$. ◁

Remark 4.37 (Independence of σ-Algebras) Note that σ-algebras are special set systems referred to in the definition above. Hence, a family $(\mathcal{A}_i, i \in I)$ of sub-σ-algebras of $\mathcal{A}$ can be P-independent as well. This fact will be used introducing the concept of *P-independence of random variables* (see section 5.4). ◁

Example 4.38 (Joe and Ann With Random Assignment – continued) Suppose $A = \{Joe\} \times \Omega_X \times \Omega_Y$ denotes the event that Joe is sampled and $B = \Omega_U \times \{yes\} \times \Omega_Y$ the event that the person sampled is treated. Then A and B are independent, because

$$P(A \cap B) \;=\; P(\{Joe\} \times \{yes\} \times \Omega_Y) \;=\; .04 + .16 \;=\; .20$$

and

$$\begin{aligned} P(A) \cdot P(B) \;&=\; P(\{Joe\} \times \Omega_X \times \Omega_Y) \;\cdot\; P(\Omega_U \times \{yes\} \times \Omega_Y) \\ &=\; (.09 + .21 + .04 + .16) \;\cdot\; (.04 + .16 + .04 + .16) \\ &=\; .50 \;\cdot\; .40 \;=\; .20. \end{aligned}$$

Hence, $P(A \cap B) = P(A) \cdot P(B)$. This implies that the σ-algebras $\{A, A^c, \Omega, \emptyset\}$ and $\{B, B^c, \Omega, \emptyset\}$ are independent as well [see Solution 4-8 (iii)]. In fact, this is a special case of the following theorem, because the set systems $\mathcal{E}_1 := \{A\}$ and $\mathcal{E}_2 := \{B\}$ are ∩-stable (see Def. 1.36) and $\sigma(\mathcal{E}_1) = \{A, A^c, \Omega, \emptyset\}$ and $\sigma(\mathcal{E}_2) = \{B, B^c, \Omega, \emptyset\}$ are the σ-algebras generated by $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively (see Def. 1.13 and Example 1.17). ◁

According to the following theorem, it is sufficient to check P-independence of a family of ∩-stable generating system in order to check P-independence of a family of σ-algebras. In this theorem, $\big(\sigma(\mathcal{E}_i), i \in I\big)$ denotes the family of σ-algebras generated by the set systems $\mathcal{E}_i$, $i \in I$.

Theorem 4.39 (∩-Stable Set Systems and Independence)
If $(\Omega, \mathcal{A}, P)$ is a probability space and $\mathcal{E}_i \subset \mathcal{A}$, $i \in I$, are ∩-stable, then

$$\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I) \quad \Rightarrow \quad \perp\!\!\!\perp_P \big(\sigma(\mathcal{E}_i), i \in I\big). \tag{4.26}$$

For a proof see Georgii (2008, Theorem 3.19, p. 65).

4.4 Conditional Independence Given an Event

Now we extent the concept of independence of events and of sets of events introducing *conditional independence of events and of sets of events given an event.*

4.4.1 Conditional Independence of Events Given an Event

Definition 4.40 (Conditional Independence of Two Events)
Let $(\Omega, \mathscr{A}, P)$ *be a probability space,* $A, B, C \in \mathscr{A}$*, and* $P(B) > 0$*. Then the events* A *and* C *are called* B*-conditionally* P*-independent, denoted* $A \perp\!\!\!\perp_P C \mid B$*, if*

$$P(A \cap C \mid B) \;=\; P(A \mid B) \cdot P(C \mid B). \tag{4.27}$$

Remark 4.41 (A Condition Equivalent to Conditional Independence) Suppose $P(B \cap C) > 0$. Then Equation (4.27) is equivalent to

$$P(A \mid B \cap C) \;=\; P(A \mid B) \tag{4.28}$$

[see Box 4.2 (xii)]. Exchanging A and C immediately yields: If $P(A \cap B) > 0$ then Equation (4.27) is equivalent to

$$P(C \mid A \cap B) \;=\; P(C \mid B). \tag{4.29}$$

◁

Remark 4.42 (Independence of Events With Respect to P^B) Using the conditional-probability measure P^B defined by (4.14) we can rewrite Equation (4.27):

$$P^B(A \cap C) \;=\; P^B(A) \cdot P^B(C). \tag{4.30}$$

This equation shows that B-conditional P-independence of A and C is equivalent to P^B*-independence of* A *and* C, which will also be denoted by $A \perp\!\!\!\perp_{P^B} C$. ◁

Remark 4.43 (Independence and Conditional Independence) Assume that $B \in \mathscr{A}$ with $P(B) > 0$. Then P-independence of A and C neither implies nor is implied by B-conditional P-independence of A and C (see Exercise 4-7). However, P-independence of A, B, and C *does* imply B-conditional P-independence of A and C [see Box 4.2 (x)]. For more propositions on conditional independence of events see Box 4.2 (p. 145). ◁

4.4.2 Conditional Independence of Set Systems Given an Event

Now we extend the concept of conditional P-independence to *set systems.* In Remark 4.42 we already noted that B-conditional P-independence of two events A and C is equivalent to P^B-independence of A and C. Correspondingly, B-conditional P-independence of a family $(\mathscr{E}_i, i \in I)$ of events is equivalent to P^B-independence of $(\mathscr{E}_i, i \in I)$.

Box 4.2 Independence and Conditional Independence of Events

Let $(\Omega, \mathscr{A}, P)$ be a probability space and $A, B, C \in \mathscr{A}$. Then:

$$A \perp\!\!\!\perp_P B \;:\Leftrightarrow\; P(A \cap B) = P(A) \cdot P(B) \tag{i}$$

$$A \perp\!\!\!\perp_P B \;\Leftrightarrow\; A^c \perp\!\!\!\perp_P B \tag{ii}$$

$$A \perp\!\!\!\perp_P B \;\Leftrightarrow\; \sigma(\{A\}) \perp\!\!\!\perp_P \sigma(\{B\}). \tag{iii}$$

$$\perp\!\!\!\perp_P A, B, C \;:\Leftrightarrow\; \begin{aligned} &P(A \cap B) = P(A) \cdot P(B), \\ &P(A \cap C) = P(A) \cdot P(C), \\ &P(B \cap C) = P(B) \cdot P(C), \\ &P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C). \end{aligned} \tag{iv}$$

$$\perp\!\!\!\perp_P A, B, C \;\Rightarrow\; A \perp\!\!\!\perp_P B,\ A \perp\!\!\!\perp_P C,\ B \perp\!\!\!\perp_P C. \tag{v}$$

If $P(B) > 0$, then

$$A \perp\!\!\!\perp_P B \;\Leftrightarrow\; P(A \mid B) = P(A) \tag{vi}$$

$$A \perp\!\!\!\perp_P C \mid B \;:\Leftrightarrow\; P(A \cap C \mid B) = P(A \mid B) \cdot P(C \mid B) \tag{vii}$$

$$A \perp\!\!\!\perp_P C \mid B \;\Leftrightarrow\; A \perp\!\!\!\perp_{P^B} C \tag{viii}$$

$$A \perp\!\!\!\perp_P C \mid B \;\Leftrightarrow\; A \perp\!\!\!\perp_P C^c \mid B \tag{ix}$$

$$\perp\!\!\!\perp_P A, B, C \;\Rightarrow\; A \perp\!\!\!\perp_P C \mid B. \tag{x}$$

If $P(B), P(B^c) > 0$, then

$$A \perp\!\!\!\perp_P B \;\Leftrightarrow\; P(A \mid B) = P(A \mid B^c). \tag{xi}$$

If $P(B \cap C) > 0$, then

$$A \perp\!\!\!\perp_P C \mid B \;\Leftrightarrow\; P(A \mid B \cap C) = P(A \mid B). \tag{xii}$$

If $P(B \cap C^c) > 0$, then

$$A \perp\!\!\!\perp_P C \mid B \;\Leftrightarrow\; P(A \mid B \cap C^c) = P(A \mid B). \tag{xiii}$$

If $P(B \cap C),\ P(B \cap C^c) > 0$, then

$$A \perp\!\!\!\perp_P C \mid B \;\Leftrightarrow\; P(A \mid B \cap C) = P(A \mid B \cap C^c) \tag{xiv}$$

$$B \perp\!\!\!\perp_P C \;\Rightarrow\; P(A \mid B) = P(A \mid B \cap C) \cdot P(C) + P(A \mid B \cap C^c) \cdot P(C^c) \tag{xv}$$

$$A \perp\!\!\!\perp_P C \mid B \;\Rightarrow\; P(A \mid B) = P(A \mid B \cap C) \cdot P(C) + P(A \mid B \cap C^c) \cdot P(C^c). \tag{xvi}$$

Definition 4.44 (Family of Conditionally Independent Set Systems)
Let $(\Omega, \mathcal{A}, P)$ be a probability space, $B \in \mathcal{A}$ with $P(B) > 0$, and $\mathcal{E}_i \subset \mathcal{A}$, $i \in I$. Then $(\mathcal{E}_i, i \in I)$ is called a family of B-conditionally P-independent set systems, denoted $\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I) \,|\, B$, if $\perp\!\!\!\perp_{P^B} (\mathcal{E}_i, i \in I)$.

Remark 4.45 (Conditional Independence of σ-Algebras) Again, σ-algebras can be such set systems referred to in the definition above. Hence, a family $(\mathcal{A}_i, i \in I)$ of sub-σ-algebras of $\mathcal{A}$ can be B-conditionally P-independent as well. ◁

Remark 4.46 (Independence of Set Systems With Respect to P^B) According to Theorem 4.39, under the assumptions of Definition 4.44, ∩-stability of the set systems $\mathcal{E}_i$, $i \in I$, implies

$$\perp\!\!\!\perp_{P^B} (\mathcal{E}_i, i \in I) \quad \Rightarrow \quad \perp\!\!\!\perp_{P^B} \big(\sigma(\mathcal{E}_i), i \in I\big). \tag{4.31}$$

◁

Together with Definition 4.44, this remark immediately implies the following corollary.

Corollary 4.47 (∩-Stable Set Systems and Conditional Independence)
If $(\Omega, \mathcal{A}, P)$ is a probability space, $B \in \mathcal{A}$ with $P(B) > 0$, and $(\mathcal{E}_i, i \in I)$ is a family of ∩-stable set systems $\mathcal{E}_i \subset \mathcal{A}$, then

$$\perp\!\!\!\perp_P (\mathcal{E}_i, i \in I) \,|\, B \quad \Rightarrow \quad \perp\!\!\!\perp_P \big(\sigma(\mathcal{E}_i), i \in I\big) \,|\, B. \tag{4.32}$$

4.5 Proofs

Proof of Theorem 4.17

In Remark 4.16 we have already shown that Equation (4.6) holds for $n = 2$. Hence, for an induction over n it suffices to show that (4.6) holds for $A_1, \ldots, A_n$ if it holds for $A_1, \ldots, A_{n-1}$. Note that $P(\bigcap_{i=1}^{n-1} A_i) > 0$ implies $P(\bigcap_{i=1}^{j-1} A_i) > 0$ for $2 \le j \le n$. Hence,

$$
\begin{aligned}
P\Big(\bigcap_{i=1}^{n} A_i\Big) &= P\Big(\bigcap_{i=1}^{n-1} A_i \cap A_n\Big) \\
&= P\Big(\bigcap_{i=1}^{n-1} A_i\Big) \cdot P\Big(A_n \,\Big|\, \bigcap_{i=1}^{n-1} A_i\Big) && \text{[(4.6) for } n=2\text{]} \\
&= P(A_1) \cdot \Big[\prod_{j=2}^{n-1} P\Big(A_j \,\Big|\, \bigcap_{i=1}^{j-1} A_i\Big)\Big] \cdot P\Big(A_n \,\Big|\, \bigcap_{i=1}^{n-1} A_i\Big) && \text{[(4.6) for } n-1\text{]} \\
&= P(A_1) \cdot \prod_{j=2}^{n} P\Big(A_j \,\Big|\, \bigcap_{i=1}^{j-1} A_i\Big).
\end{aligned}
$$

Proof of Theorem 4.20

(i) This equation immediately follows from (1.28).

(ii) If $P(A_i) > 0$, then $P(B \cap A_i) = P(B|A_i) \cdot P(A_i)$ [see Eq. (4.2)]. Hence, (4.8) immediately implies

$$P(B) = \sum_{i=1}^{n} P(B \,|\, A_i) \cdot P(A_i).$$

(iii) This proposition immediately follows from (1.29).

(iv) This proposition immediately follows from (iii) inserting $P(B \cap A_i) = P(B|A_i) \cdot P(A_i)$ [see Eq. (4.2)].

Proof of Theorem 4.23

4.1 (a)

$$
\begin{aligned}
P^B(\Omega) &= \frac{P(B \cap \Omega)}{P(B)} && \text{[(4.2)]} \\
&= \frac{P(B)}{P(B)} && [B \subset \Omega] \\
&= 1.
\end{aligned}
$$

4.1 (b) We assume $P(B) > 0$. Therefore, $P(A \cap B) \geq 0$, for all $A \in \mathscr{A}$, implies that $P^B(A) = P(A \cap B)/P(B) \geq 0$, for all $A \in \mathscr{A}$.

4.1 (c) If $A_1, A_2, \ldots$ are pairwise disjoint, then $A_1 \cap B$, $A_2 \cap B, \ldots$ are pairwise disjoint. Therefore,

$$
\begin{aligned}
P^B\Big(\bigcup_{i=1}^{\infty} A_i\Big) &= \frac{P\big(\bigcup_{i=1}^{\infty} A_i \cap B\big)}{P(B)} && \text{[(4.2)]} \\
&= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} && \text{[Def. 4.1 (c)]} \\
&= \sum_{i=1}^{\infty} P^B(A_i). && \text{[(4.2)]}
\end{aligned}
$$

Proof of Lemma 4.25

$$\begin{aligned} P(A\,|\,B\cap C) &= \frac{P(A\cap B\cap C)}{P(B\cap C)} && [(4.2)]\\ &= \frac{P(A\cap C\,|\,B)\cdot P(B)}{P(C\,|\,B)\cdot P(B)} && [(4.2)]\\ &= \frac{P^B(A\cap C)}{P^B(C)} && [(4.14)]\\ &= P^B(A\,|\,C). && [(4.2)] \end{aligned}$$

Proof of Lemma 4.27

(i) For all $A\in\mathscr{A}$,

$$\begin{aligned} P(A) = 0 \quad &\Rightarrow \quad P(A\cap B) = 0 && [\text{Box 4.1 (v)}]\\ &\Rightarrow \quad \frac{P(A\cap B)}{P(B)} = 0 && [P(B)>0]\\ &\Rightarrow \quad P(A\,|\,B) = 0 && [(4.2)]\\ &\Rightarrow \quad P^B(A) = 0. && [(4.14)] \end{aligned}$$

Hence, $P^B \underset{\mathscr{A}}{\ll} P$ [see Eq. (3.77)].

(ii) For all $A\in\mathscr{A}$,

$$\begin{aligned} P^B(A) &= \frac{P(A\cap B)}{P(B)} && [(4.14), (4.2)]\\ &= \frac{1}{P(B)}\int 1_{A\cap B}\,dP && [(3.9)]\\ &= \frac{1}{P(B)}\int 1_A\cdot 1_B\,dP && [(1.32)]\\ &= \int_A \frac{1}{P(B)}\cdot 1_B\,dP. && [(3.30), (3.32)] \end{aligned}$$

According to Theorem 3.65, this means $P^B = \left(\frac{1}{P(B)}\cdot 1_B\right)\odot P$.

4.6 Exercises

▷ **Exercise 4-1** Consider flipping a coin n times and the event A_1 = *flipping heads at the first flip*. Specify the set of possible outcomes of this random experiment and the set A_1 as a subset of Ω. How many elements has Ω? How many elements has the event A_1?

▷ **Exercise 4-2** Define the set Ω of possible outcomes in the example displayed in Table 4.1 (p. 130) by a threefold Cartesian product.

▷ **Exercise 4-3** Draw the interval $[0,1]$, cut it in two halves, cut the right-hand piece in two halves, cut the remaining part in two halves, etc. In this way you can visualize the sequence $1/2, 1/4, 1/8, \ldots$ by lengths of intervals. This sequence can also be written: $1/2^i, i\in\mathbb{N}$. Note that all terms $1/2^i$ of this sequence are positive, i. e., $1/2^i > 0$ for all $i\in\mathbb{N}$. Determine

$$\sum_{i=1}^{\infty}\frac{1}{2^i} = \lim_{n\to\infty}\sum_{i=1}^{n}\frac{1}{2^i}.$$

▷ **Exercise 4-4** Compute the probabilities $P(A)$ and $P(C \mid A \cap B)$ of the events defined in Example 4.18 .

▷ **Exercise 4-5** Compute the B-conditional probability of $\{(Ann, yes, -),\ (Ann, yes, +)\}$, the event that *Ann is sampled and treated*. Use the results already obtained in Example 4.29.

▷ **Exercise 4-6** Prove the proposition of Remark 4.34.

▷ **Exercise 4-7** Show by examples that, for $B \in \mathscr{A}$ with $P(B) > 0$, P-independence of A and C neither implies nor is implied by B-conditional P-independence of A and C.

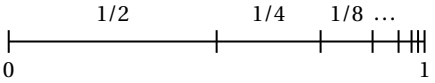
▷ **Exercise 4-8** Prove the propositions of Box 4.2.

Solutions

▷ **Solution 4-1** The set of possible outcomes is $\Omega = \{h, t\}^n = \{h, t\} \times \ldots, \times \{h, t\}$ (n-times). The event *flipping heads at time* 1 is $A_1 = \{h\} \times \{h, t\}^{n-1}$. The set Ω has 2^n elements and A_1 has $2^n/2 = 2^{n-1}$ elements.

▷ **Solution 4-2** If $\Omega_U = \{Joe, Ann\}$, $\Omega_X = \{yes, no\}$, and $\Omega_Y = \{+, -\}$, then $\Omega = \Omega_U \times \Omega_X \times \Omega_Y$. This set consists of all eight triples (a, b, c), for which $a \in \Omega_U$, $b \in \Omega_X$, and $c \in \Omega_Y$.

▷ **Solution 4-3** The picture of this interval is



and this illustrates that $\displaystyle\sum_{i=1}^{\infty} \frac{1}{2^i} = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{2^i} = 1$.

▷ **Solution 4-4** Because the four events $\{(Joe, no, -)\}, \ldots, \{(Joe, yes, +)\}$ are pairwise disjoint, we can simply add their probabilities. Hence, $P(A) = .144 + .336 + .004 + .016 = .5$ (see the second column of Table 4.2). In order to compute $P(C \mid A \cap B)$ note that $A \cap B = \{(Joe, yes, -), (Joe, yes, +)\}$ is the event that *Joe is drawn and treated*. Again, because the two elementary events involved are disjoint, $P(A \cap B) = .004 + .016 = 0.2$. Furthermore, $A \cap B \cap C = \{(Joe, yes, +)\}$ is the event that *Joe is drawn, treated, and successful*. Its probability is $P(A \cap B \cap C) = P\big(\{(Joe, yes, +)\}\big) = .016$. Hence,

$$P(C \mid A \cap B) \;=\; \frac{P(A \cap B \cap C)}{P(A \cap B)} \;=\; \frac{.016}{.004 + .016} \;=\; .8\,.$$

▷ **Solution 4-5** In Example 4.29 we already computed the two B-conditional probabilities $P^B\big(\{(Ann, yes, -)\}\big) = .30$ and $P^B\big(\{(Ann, yes, +)\}\big) = .20$. Because these elementary events are *disjoint*, the probabilities of their union can easily be computed using the additivity property of the probability measure P^B. Hence, $P^B\big(\{(Ann, yes, -),\ (Ann, yes, +)\}\big) = .30 + .20 = .50$.

▷ **Solution 4-6** Let $(\Omega, \mathscr{A}, P)$ be a probability space. Then for all $A \in \mathscr{A}$:
$P(\Omega \cap A) = P(A) = 1 \cdot P(A) = P(\Omega) \cdot P(A)$ and $P(\emptyset \cap A) = P(\emptyset) = 0 = 0 \cdot P(A) = P(\emptyset) \cdot P(A)$.

▷ **Solution 4-7** Consider Example 2.2.2, let $A = \{(h,t),(h,h)\}$ denote the event to flip heads with the first coin, $B = \{(t,t),(h,h)\}$ the event to flip tails or to flip heads with both coins, and $C = \{(t,h),(h,h)\}$ the event to flip heads with the second coin. All three events have the same probability $P(A) = P(B) = P(C) = .5$. Now,

$$P(A \cap C) \;=\; P(\{(h,h)\}) \;=\; .25 \;=\; .5 \cdot .5 \;=\; P(A) \cdot P(C)$$

and

$$P(B \cap C) \;=\; P(\{(h,h)\}) \;=\; .25 \;=\; .5 \cdot .5 \;=\; P(B) \cdot P(C).$$

Hence, A and C as well as B and C are P-independent, which implies $P(A \,|\, B) = .5$. and $P(C \,|\, B) = .5$. However,

$$\begin{aligned} P(A \cap C \,|\, B) \;&=\; \frac{P(A \cap C \cap B)}{P(B)} \;=\; \frac{.25}{.5} \;=\; .5 \\ &\neq\; .25 \;=\; .5 \cdot .5 \;=\; P(A \,|\, B) \cdot P(C \,|\, B), \end{aligned}$$

which shows that A and C are not B-conditionally P-independent.

Now we present an example in which A and C are B-conditionally P-independent but not (unconditionally) P-independent. Consider flipping three coins. This random experiment is represented by the probability space $(\Omega, \mathscr{A}, P)$, where $\Omega = \{h,t\}^3$, $\mathscr{A} = \mathscr{P}(\Omega)$, and $P\colon \mathscr{A} \to [0,1]$, satisfying $P(\{\omega\}) = .125$ for all $\omega \in \Omega$. Furthermore, let $A = \{(t,t,t),(t,t,h)\}$ denote the event to flip tails with the first two coins, $B = \{(t,t,t),(t,t,h),(t,h,t),(t,h,h)\}$ the event to flip tails with the first coin, and $C = \{(t,t,h),(t,h,h)\}$ the event to flip tails with the first coin and heads with the third coin. The two events A and C have the same probability $P(A) = P(C) = .25$ and $P(B) = .5$. Because

$$P(A \cap C) \;=\; P(\{(t,t,h)\}) \;=\; .125 \;\neq\; .25 \cdot .25 \;=\; P(A) \cdot P(C),$$

A and C are *not P-independent.* Further, $P(A \cap B) = P(\{(t,t,t),(t,t,h)\}) = .25$, $P(C \cap B) = P(\{(t,h,t),(t,h,h)\}) = .25$, and

$$\begin{aligned} P(A \cap C \,|\, B) \;&=\; \frac{P(A \cap B \cap C)}{P(B)} \;=\; \frac{.125}{.5} \;=\; .25 \\ &=\; .5 \cdot .5 \;=\; \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B \cap C)}{P(B)} \;=\; P(A \,|\, B) \cdot P(C \,|\, B). \end{aligned}$$

This shows that A and C are B-conditionally P-independent.

▷ **Solution 4-8** (i) This is the definition of $A \underset{P}{\perp\!\!\!\perp} B$.

(ii)
$$\begin{aligned} P(A^c \cap B) \;&=\; P\big(B \setminus A\big) \\ &=\; P(B) - P(A \cap B) && \text{[Box 4.1 (vi)]} \\ &=\; P(B) - P(A) \cdot P(B) && [A \underset{P}{\perp\!\!\!\perp} B] \\ &=\; [1 - P(A)] \cdot P(B) \\ &=\; P(A^c) \cdot P(B) && \text{[Box 4.1 (iv)]}, \end{aligned}$$

which is $A^c \underset{P}{\perp\!\!\!\perp} B$.

(iii) We have to show that $A \underset{P}{\perp\!\!\!\perp} B$ implies

$$P(A_i \cap B_j) \;=\; P(A_i) \cdot P(B_j), \quad \forall A_i \in \{A, A^c, \Omega, \emptyset\} \text{ and } \forall B_j \in \{B, B^c, \Omega, \emptyset\}.$$

Whenever A_i or B_j is Ω or $\emptyset$ this equation holds [see (4.25)]. Furthermore, $P(A\cap B) = P(A)\cdot P(B)$ is equivalent to our premise $A \underset{P}{\perp\!\!\!\perp} B$, and $P(A^c\cap B) = P(A^c)\cdot P(B)$ is proposition (ii). The corresponding argument holds for $P(A\cap B^c) = P(A)\cdot P(B^c)$ and $P(A^c\cap B^c) = P(A^c)\cdot P(B^c)$, exchanging the roles of A and B.

(iv) This is the definition of $\underset{P}{\perp\!\!\!\perp} A, B, C$.

(v) This proposition immediately follows from (iv) and (i).

(vi) We assume $P(B) > 0$. Then

$$\begin{aligned} A \underset{P}{\perp\!\!\!\perp} B &\Leftrightarrow P(A\cap B) = P(A)\cdot P(B) && [\text{(i)}] \\ &\Leftrightarrow \frac{P(A\cap B)}{P(B)} = P(A) \\ &\Leftrightarrow P(A\,|\,B) = P(A). && [\text{(4.2)}] \end{aligned}$$

(vii) This is the definition of $A \underset{P}{\perp\!\!\!\perp} C\,|\,B$.

(viii) We assume $P(B) > 0$. Therefore,

$$\begin{aligned} A \underset{P}{\perp\!\!\!\perp} C\,|\,B &\Leftrightarrow P(A\cap C\,|\,B) = P(A\,|\,B)\cdot P(C\,|\,B) && [\text{(vii)}] \\ &\Leftrightarrow P^B(A\cap C) = P^B(A)\cdot P^B(C) && [\text{(4.14)}] \\ &\Leftrightarrow A \underset{P^B}{\perp\!\!\!\perp} C && [\text{(i)}] \end{aligned}$$

(ix) We assume $P(B) > 0$. Therefore,

$$\begin{aligned} A \underset{P}{\perp\!\!\!\perp} C\,|\,B &\Leftrightarrow A \underset{P^B}{\perp\!\!\!\perp} C && [\text{(viii)}] \\ &\Leftrightarrow A \underset{P^B}{\perp\!\!\!\perp} C^c. && [\text{(ii)}] \end{aligned}$$

(x) We assume $P(B) > 0$. Then

$$\begin{aligned} P(A\cap C\,|\,B) &= \frac{P(A\cap B\cap C)}{P(B)} && [\text{(4.2)}] \\ &= \frac{P(A)\cdot P(B)\cdot P(C)}{P(B)} && [\underset{P}{\perp\!\!\!\perp} A, B, C, \ \text{(iv)}] \\ &= P(A\,|\,B)\cdot P(C\,|\,B). && [\text{(v), (vi)}] \end{aligned}$$

(xi) We assume $P(B), P(B^c) > 0$. Then

$$\begin{aligned} P(A\,|\,B) = P(A\,|\,B^c) &\Leftrightarrow \frac{P(A\cap B)}{P(B)} = \frac{P(A\cap B^c)}{1-P(B)} && [\text{(4.2), Box 4.1 (iv)}] \\ &\Leftrightarrow P(A\cap B)\cdot[1-P(B)] = P(A\cap B^c)\cdot P(B) \\ &\Leftrightarrow P(A\cap B) = [P(A\cap B) + P(A\cap B^c)]\cdot P(B) \\ &\Leftrightarrow P(A\cap B) = P(A)\cdot P(B) && [\text{(4.7)}] \\ &\Leftrightarrow A \underset{P}{\perp\!\!\!\perp} B. && [\text{(i)}] \end{aligned}$$

(xii) We assume $P(B\cap C) > 0$. This implies $P(B) > 0$ and

$$\begin{aligned} A \underset{P}{\perp\!\!\!\perp} C\,|\,B &\Leftrightarrow P(A\cap C\,|\,B) = P(A\,|\,B)\cdot P(C\,|\,B) && [\text{(vii)}] \\ &\Leftrightarrow \frac{P(A\cap B\cap C)}{P(B)} = \frac{P(A\cap B)}{P(B)}\cdot\frac{P(B\cap C)}{P(B)} && [\text{(4.2)}] \\ &\Leftrightarrow \frac{P(A\cap B\cap C)}{P(B\cap C)} = \frac{P(A\cap B)}{P(B)} \\ &\Leftrightarrow P(A\,|\,B\cap C) = P(A\,|\,B). && [\text{(4.2)}] \end{aligned}$$

(xiii) We assume $P(B \cap C^c) > 0$. This implies $P(B) > 0$ and

$$\begin{aligned} A \underset{P}{\perp\!\!\!\perp} C \,|\, B &\Leftrightarrow A \underset{P^B}{\perp\!\!\!\perp} C && [\text{(viii)}] \\ &\Leftrightarrow A \underset{P^B}{\perp\!\!\!\perp} C^c && [\text{(ii)}] \\ &\Leftrightarrow P^B(A \,|\, C^c) = P^B(A) && [\text{(vi)}] \\ &\Leftrightarrow P(A \,|\, B \cap C^c) = P(A \,|\, B) && [(4.15), (4.14)] \end{aligned}$$

(xiv) We assume $P(B \cap C), P(B \cap C^c) > 0$.

$$\begin{aligned} P(A \,|\, B \cap C) = P(A \,|\, B \cap C^c) &\Leftrightarrow P^B(A \,|\, C) = P^B(A \,|\, C^c) && [(4.15)] \\ &\Leftrightarrow A \underset{P^B}{\perp\!\!\!\perp} C && [\text{(xi)}] \\ &\Leftrightarrow A \underset{P}{\perp\!\!\!\perp} C \,|\, B. && [\text{(viii)}] \end{aligned}$$

(xv) We assume $P(B \cap C), P(B \cap C^c) > 0$.

$$\begin{aligned} B \underset{P}{\perp\!\!\!\perp} C &\Rightarrow P(C \,|\, B) = P(C), \;\; P(C^c \,|\, B) = P(C^c) && [\text{(vi), (ii)}] \\ &\Rightarrow P(A \,|\, B) = P(A \,|\, B \cap C) \cdot P(C) + P(A \,|\, B \cap C^c) \cdot P(C^c). && [(4.17)] \end{aligned}$$

(xvi) We assume $P(B \cap C), P(B \cap C^c) > 0$ and $A \underset{P}{\perp\!\!\!\perp} C \,|\, B$.

$$\begin{aligned} P(A \,|\, B) &= P(A \,|\, B \cap C) && [A \underset{P}{\perp\!\!\!\perp} C \,|\, B, \text{(xii)}] \\ &= P(A \,|\, B \cap C) \cdot [P(C) + P(C^c)] && [\text{Box 4.1 (iv)}] \\ &= P(A \,|\, B \cap C) \cdot P(C) + P(A \,|\, B \cap C) \cdot P(C^c)] \\ &= P(A \,|\, B \cap C) \cdot P(C) + P(A \,|\, B \cap C^c) \cdot P(C^c)]. && [\text{(xiv)}] \end{aligned}$$

Chapter 5
Random Variable, Distribution, Density, and Distribution Function

In chapter 4 we translated the concepts *measure* and *measure space* to probability theory introducing the notions *probability measure* and *probability space*. In this chapter we define a *random variable* as a measurable mapping and its *distribution* as the image measure of a measurable mapping with respect to a probability measure (see ch. 2). The distribution of a random variable contains the comprehensive information about its properties. It informs us about the probability of each event that can be represented by this random variable. Expectation, variance and other moments of a random variable are determined by its distribution (see ch. 6). For a multivariate random variable, the (joint) distribution also contains the information about the dependencies between its components. It also determines the conditional expectations (see ch. 10). In this chapter, we apply the concept of independence of families of events in order to introduce *independence of random variables* and *families of random variables*. Finally, the last sections of this chapter are devoted to the concept of a *probability function*, and, for a real-valued random variable, the notions of a *distribution function* and a *probability density*, which are very useful for describing a distribution, for calculations (see, e. g., ch. 6), and for providing instructive illustrations of the underlying distributions (see ch. 8).

5.1 Random Variable and its Distribution

In section 2.6 we introduced the notation

$$f\colon (\Omega, \mathscr{A}, \mu) \to (\Omega', \mathscr{A}'),$$

which expresses that $f\colon \Omega \to \Omega'$ is an $(\mathscr{A}, \mathscr{A}')$-measurable mapping and that μ is a measure on the measurable space $(\Omega, \mathscr{A})$. If μ is a probability measure, then a measurable mapping is also called a *random variable* and its image measure μ_f is also called its *distribution*. This change of terms goes along with a change in notation. Instead of f, g, and h, we preferably use letters such as X, Y, and Z.

Definition 5.1 (Random Variable)
If $(\Omega, \mathscr{A}, P)$ is a probability space and $X\colon (\Omega, \mathscr{A}) \to (\Omega'_X, \mathscr{A}'_X)$ a measurable mapping, i. e., if $X\colon \Omega \to \Omega'_X$ satisfies

$$X^{-1}(A') \in \mathscr{A}, \quad \forall\, A' \in \mathscr{A}'_X, \tag{5.1}$$

then X is called a $random\ variable\ on$ $(\Omega, \mathscr{A}, P)$ *with values in* $(\Omega'_X, \mathscr{A}'_X)$. *If* $(\Omega'_X, \mathscr{A}'_X) = (\mathbb{R}, \mathscr{B})$, *then X is called* $real\text{-}valued$, *and if* $(\Omega'_X, \mathscr{A}'_X) = (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, *then X is called* $numerical$.

Remark 5.2 (Measurability of Inverse Images) Equation (5.1) implies that all inverse images

$$X^{-1}(A') := \{\omega \in \Omega\colon X(\omega) \in A'\}, \quad A' \in \mathscr{A}'_X,$$

are elements of the σ-algebra $\mathscr{A}$ on Ω. Because the measure $P : \mathscr{A} \to [0, 1]$ assigns a probability to *all* elements of $\mathscr{A}$, the probabilities $P[X^{-1}(A')]$ of these inverse images are determined by P (see Exercises 5-1 and 5-2). ◁

Definition 5.3 (Distribution of a Random Variable)
Suppose that $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ *is a random variable. Then the function* $P_X\colon \mathscr{A}'_X \to [0,1]$ *defined by*

$$P_X(A') \;=\; P[X^{-1}(A')], \quad \forall\, A' \in \mathscr{A}'_X, \tag{5.2}$$

is called the $distribution$ *of X (with respect to P).*

Remark 5.4 (Notation $P(X \in A')$ and $P(X=x)$) If $A' \in \mathscr{A}'_X$, we use the notation

$$P(X \in A') \;:=\; P[X^{-1}(A')] \tag{5.3}$$

for the probability of the event $\{X \in A'\} = X^{-1}(A')$ and, if $\{x\} \in \mathscr{A}'_X$,

$$P(X=x) \;:=\; P[X^{-1}(\{x\})] \tag{5.4}$$

for the probability of the event $\{X=x\} = X^{-1}(\{x\}) = \{\omega \in \Omega\colon X(\omega) = x\}$. If we write $P(X=x)$, then we always assume $\{x\} \in \mathscr{A}'_X$, even if not mentioned explicitly. ◁

Remark 5.5 (A New Probability Space) Definition 5.1 implies that *every* random variable X on a probability space $(\Omega, \mathscr{A}, P)$ has a distribution P_X. Furthermore, $P_X\colon \mathscr{A}'_X \to [0,1]$ is also a measure, the *image measure of P under X* (see Th. 2.78 and Def. 2.79). Because $P_X(\Omega'_X) = P(\Omega) = 1$, we can conclude that P_X is a probability measure, and $(\Omega'_X, \mathscr{A}'_X, P_X)$ is also a probability space. Therefore, we use the notation

$$X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X, P_X)$$

expressing that

(a) $X\colon \Omega \to \Omega'_X$ is a random variable on the probability space $(\Omega, \mathscr{A}, P)$,
(b) $\mathscr{A}'_X$ is a σ-algebra on Ω'_X, and
(c) P_X is the distribution of X.

◁

Definition 5.6 (Identically Distributed Random Variables)
Let $X\colon (\Omega^{(1)}, \mathscr{A}^{(1)}, P^{(1)}) \to (\Omega', \mathscr{A}')$ and $Y\colon (\Omega^{(2)}, \mathscr{A}^{(2)}, P^{(2)}) \to (\Omega', \mathscr{A}')$ be random variables. If $P_X = P_Y$, then we say that X and Y are identically distributed.

Note that, oftentimes, $(\Omega^{(1)}, \mathscr{A}^{(1)}, P^{(1)}) = (\Omega^{(2)}, \mathscr{A}^{(2)}, P^{(2)})$. Now we consider the distribution of a composition $g(X)$ of a random variable $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ and a measurable function $g\colon (\Omega'_X, \mathscr{A}'_X) \to (\Omega', \mathscr{A}')$. According to Remark 5.5, the mapping g is a random variable on the probability space $(\Omega'_X, \mathscr{A}'_X, P_X)$. Furthermore, according to the following lemma, $g(X)$ is a random variable on $(\Omega, \mathscr{A}, P)$ and the distribution of $g(X)$ is the image measure of P_X under g. The notation of this image measure is $(P_X)_g$.

Lemma 5.7 (Distribution of a Composition)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable and $g\colon (\Omega'_X, \mathscr{A}'_X) \to (\Omega', \mathscr{A}')$ a measurable function. Then the composition $g(X)\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ is a random variable and

$$(P_X)_g = P_{g(X)}\,. \tag{5.5}$$

(Proof p. 185)

Example 5.8 (Indicator (Variable) of an Event) If $(\Omega, \mathscr{A}, P)$ is a probability space and $A \in \mathscr{A}$, then the mapping $1_A\colon (\Omega, \mathscr{A}, P) \to \big(\{0,1\}, \mathscr{P}(\{0,1\})\big)$ is a random variable. It is called the *indicator (variable) of A*. The distribution of 1_A is

$$P_{1_A}(\{0\}) = P(A^c), \quad P_{1_A}(\{1\}) = P(A), \quad P_{1_A}(\{0,1\}) = P(\Omega) = 1, \quad P_{1_A}(\emptyset) = P(\emptyset) = 0.$$

If we consider the same event A and the measurable space $(\mathbb{R}, \mathscr{B})$, then we can also write $1_A\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ in order to express that 1_A is also $(\mathscr{A}, \mathscr{B})$-measurable. Note, however, that now the distribution of 1_A is a probability measure on $(\mathbb{R}, \mathscr{B})$, and for all $B \in \mathscr{B}$,

$$\begin{aligned} P_{1_A}(B) &= P[1_A^{-1}(B)] = P[\{\omega \in \Omega\colon 1_A(\omega) \in B\}] \qquad [(5.2), (2.2)] \\ &= \begin{cases} P(\emptyset) = 0, & \text{if } 0 \notin B, 1 \notin B, \\ P(A), & \text{if } 0 \notin B, 1 \in B, \\ P(A^c), & \text{if } 0 \in B, 1 \notin B, \\ P(\Omega) = 1, & \text{if } \{0,1\} \subset B. \end{cases} \end{aligned}$$

◁

Example 5.9 (Indicator of an Inverse Image) If $(\Omega, \mathscr{A}, P)$ is a probability space, $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ a random variable, and $A' \in \mathscr{A}'_X$, then $1_{X^{-1}(A')}\colon (\Omega, \mathscr{A}, P) \to \big(\{0,1\}, \mathscr{P}(\{0,1\})\big)$ is a random variable on $(\Omega, \mathscr{A}, P)$ and

$$1_{X \in A'} := 1_{X^{-1}(A')} = 1_{A'}(X) = 1_{A'} \circ X \tag{5.6}$$

(see Exercise 5-3). The distribution of $1_{X \in A'}$ is

$$P_{1_{X\in A'}}(\{0\}) = P(X \notin A'), \quad P_{1_{X\in A'}}(\{1\}) = P(X \in A'),$$
$$P_{1_{X\in A'}}(\{0,1\}) = P(\Omega) = 1, \quad P_{1_{X\in A'}}(\emptyset) = P(\emptyset) = 0.$$

◁

Example 5.10 (Dichotomous Random Variable With Values 0 and 1) Let X be a real-valued random variable on $(\Omega, \mathscr{A}, P)$. Then it is called *dichotomous with values* 0 *and* 1 if $X \underset{P}{=} 1_{X=1}$ and $0 < P(X=1) < 1$. ◁

Example 5.11 (Flipping two Coins – continued) In Example 2.2.2, we considered flipping two coins and defined $X\colon (\Omega, \mathscr{A}, P) \to \big(\Omega'_X, \mathscr{P}(\Omega'_X)\big)$, a random variable representing with its values the *number of flipping heads*. Its possible values are 0, 1, or 2. Hence, we can choose $\Omega'_X := \{0,1,2\}$ and

$$P(X=0) = P_X(\{0\}) = P[X^{-1}(\{0\})] = P[\{(t,t)\}] = \frac{1}{4},$$
$$P(X=1) = P_X(\{1\}) = P[X^{-1}(\{1\})] = P[\{(h,t),(t,h)\}] = \frac{1}{2},$$
$$P(X=2) = P_X(\{2\}) = P[X^{-1}(\{2\})] = P[\{(h,h)\}] = \frac{1}{4}$$

are the probabilities assigned to the singletons $\{0\}$, $\{1\}$, and $\{2\}$, whereas

$$P(X \in \{0,1\}) = P_X(\{0,1\}) = P[X^{-1}(\{0,1\})] = P[\{(t,t),(h,t),(t,h)\}] = \frac{3}{4},$$
$$P(X \in \{0,2\}) = P_X(\{0,2\}) = P[X^{-1}(\{0,2\})] = P[\{(t,t),(h,h)\}] = \frac{2}{4},$$
$$P(X \in \{1,2\}) = P_X(\{1,2\}) = P[X^{-1}(\{1,2\})] = P[\{(h,t),(t,h)(h,h)\}] = \frac{3}{4}$$

are the probabilities assigned to the sets $\{0,1\}$, $\{0,2\}$, and $\{1,2\}$, which consist of two elements of Ω'_X. Finally, $P_X(\Omega'_X) = P[X^{-1}(\Omega'_X)] = P(\Omega) = 1$ and $P_X(\emptyset) = P[X^{-1}(\emptyset)] = P(\emptyset) = 0$. Generally speaking, $P_X(A')$ denotes the probability that X takes on a value in the subset A' of Ω'_X. ◁

Example 5.12 (Tom, Jim, and Kate) Now we consider an example that is similar to the experiment with Joe and Ann. However, the set of persons is now $\Omega_U := \{Tom, Jim, Kate\}$, and we consider three treatments, the elements of the set $\Omega_X := \{Con, BTh, PTh\}$, where *Con* could be *no treatment*. The random experiment consists of: drawing a unit u from the set Ω_U, assigning it to one of the three

Table 5.1. Tom, Jim, and Kate

Elements of Ω		Random variables		
Unit Treatment Success	Probabilities of elementary events $P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y
(*Tom*, *Con*, −)	10/99	*Tom*	0	0
(*Tom*, *Con*, +)	10/99	*Tom*	0	1
(*Tom*, *BTh*, −)	2/99	*Tom*	1	0
(*Tom*, *BTh*, +)	6/99	*Tom*	1	1
(*Tom*, *PTh*, −)	1/99	*Tom*	2	0
(*Tom*, *PTh*, +)	4/99	*Tom*	2	1
(*Jim*, *Con*, −)	5/99	*Jim*	0	0
(*Jim*, *Con*, +)	15/99	*Jim*	0	1
(*Jim*, *BTh*, −)	3/99	*Jim*	1	0
(*Jim*, *BTh*, +)	5/99	*Jim*	1	1
(*Jim*, *PTh*, −)	2/99	*Jim*	2	0
(*Jim*, *PTh*, +)	3/99	*Jim*	2	1
(*Kate*, *Con*, −)	12/99	*Kate*	0	0
(*Kate*, *Con*, +)	8/99	*Kate*	0	1
(*Kate*, *BTh*, −)	5/99	*Kate*	1	0
(*Kate*, *BTh*, +)	3/99	*Kate*	1	1
(*Kate*, *PTh*, −)	4/99	*Kate*	2	0
(*Kate*, *PTh*, +)	1/99	*Kate*	2	1

treatment conditions *Con*, *BTh*, or *PTh* and observing whether (+) or not (−) a success criterion is reached. Hence, the set of possible outcomes of this random experiment is

$$\Omega := \Omega_U \times \Omega_X \times \Omega_Y = \big\{ (Tom, Con, -),\ (Tom, Con, +), \ldots,\ (Kate, PTh, +) \big\}.$$

It consists of the 3·3·2 = 18 triples (u, ω_X, ω_Y) listed in the first column of Table 5.1. As the set of possible events $\mathscr{A}$ we consider the power set $\mathscr{P}(\Omega)$. This set has $2^{18} = 262144$ elements, where 18 is the number of elements of Ω. The probabilities of the 18 elementary events $\{\omega\}$, $\omega \in \Omega$, are displayed in the second column of the table. With these specifications, the probabilities $P(A)$ of all 2^{18} elements $A \in \mathscr{A}$ are determined [see Rule (x) of Box 4.1]. Hence, the probability space $(\Omega, \mathscr{A}, P)$ is completely specified.

Table 5.1 also displays the values of the three random variables $U\colon (\Omega, \mathscr{A}, P) \to [\Omega_U, \mathscr{P}(\Omega_U)]$, $X\colon (\Omega, \mathscr{A}, P) \to [\Omega'_X, \mathscr{P}(\Omega'_X)]$, and $Y\colon (\Omega, \mathscr{A}, P) \to [\Omega'_Y, \mathscr{P}(\Omega'_Y)]$, where

$\Omega'_X := \{0,1,2\}$ and $\Omega'_Y := \{0,1\}$. For the singletons $\{x\}$, $x \in \Omega'_X$, the values $P_X(\{x\}) = P[X^{-1}(\{x\})]$ of the distribution of X are

$$P_X(\{0\}) = 60/99, \qquad P_X(\{1\}) = 24/99, \qquad P_X(\{2\}) = 15/99,$$

for the sets that consist of two elements of Ω'_X, they are

$$P_X(\{0,1\}) = 84/99, \qquad P_X(\{0,2\}) = 75/99, \qquad P_X(\{1,2\}) = 39/99,$$

and for Ω'_X and $\emptyset$, they are $P_X(\Omega'_X) = 1$ and $P_X(\emptyset) = 0$.

For the singletons $\{u\}$, $u \in \Omega_U$, the values $P_U(\{u\}) = P[U^{-1}(\{u\})]$ of the distribution of U are

$$P_U(\{Tom\}) = P_U(\{Jim\}) = P_U(\{Kate\}) = 1/3,$$

for the sets that consist of two elements of Ω_U they are

$$P_U(\{Tom, Jim\}) = P_U(\{Tom, Kate\}) = P_U(\{Jim, Kate\}) = 2/3,$$

and for Ω_U and $\emptyset$, they are $P_U(\Omega_U) = 1$ and $P_U(\emptyset) = 0$. ◁

5.2 Equivalence of Two Random Variables With Respect to a Probability Measure

5.2.1 Identical and P-Equivalent Random Variables

Let $X, Y : (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ be two random variables. Then X and Y are called *identical* if

$$\forall\, \omega \in \Omega: \quad X(\omega) = Y(\omega). \tag{5.7}$$

Remark 5.13 (P-Equivalent Random Variables) Let X, Y: $(\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ be two random variables. Then X and Y are *almost surely identical* with respect to P or *P-equivalent*, denoted $X \underset{P}{=} Y$, if

$$\exists\, A \in \mathscr{A}: \quad \big(\forall\, \omega \in \Omega \setminus A: \; X(\omega) = Y(\omega) \quad \text{and} \quad P(A) = 0\big) \tag{5.8}$$

(see Def. 2.68). Another notation for $X \underset{P}{=} Y$ is $X(\omega) \underset{P\text{-a.a.}}{=} Y(\omega)$, which is a shortcut for

$$X(\omega) = Y(\omega), \quad \text{for } P\text{-a.a. } \omega \in \Omega, \tag{5.9}$$

meaning that the values of X and Y are identical for *P-almost all* $\omega \in \Omega$ (see Rem. 2.70). ◁

Remark 5.14 (Singleton With a Positive Probability) If $X \underset{P}{=} Y$ or, equivalently, if $X(\omega) \underset{P\text{-a.a.}}{=} Y(\omega)$, and $\omega^* \in \Omega$, with $P(\{\omega^*\}) > 0$, then $X(\omega^*) = Y(\omega^*)$ [see Rem. 2.71]. ◁

Example 5.15 (Indicator of a Null Set) Let $(\Omega, \mathscr{A}, P)$ be a probability space and $A \in \mathscr{A}$. If $P(A) = 0$, then

$$1_A \underset{P}{=} 0 \quad \text{and} \quad 1_{A^c} = 1 - 1_A \underset{P}{=} 1 \tag{5.10}$$

(see Example 5.8). ◁

Remark 5.16 (Q-Equivalence) Note that the definition of equivalence of two random variables X and Y with respect to a probability measure only presumes that X and Y are measurable with respect to a σ-algebra on Ω and that the measure considered is a probability measure on this σ-algebra. Hence, we can consider the equivalence of X and Y with respect to different probability measures, say P and Q and study their relationship. ◁

In the following lemma we consider the relationship between P-equivalence and Q-equivalence, presuming $Q \underset{\mathscr{C}}{\ll} P$ (absolute continuity), i. e., presuming

$$\forall\, C \in \mathscr{C}\colon \;\; P(C) = 0 \;\Rightarrow\; Q(C) = 0$$

(see Def. 3.70).

Lemma 5.17 (P-Equivalence and Q-Equivalence)
Let $X, Y\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ be random variables, let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, and assume that $\sigma(X), \sigma(Y) \subset \mathscr{C}$. If Q is a probability measure on $(\Omega, \mathscr{A})$ such that $Q \underset{\mathscr{C}}{\ll} P$, then $X \underset{P}{=} Y$ implies $X \underset{Q}{=} Y$.

(Proof p. 185)

According to Lemma 4.27 (i), $P^B \underset{\mathscr{A}}{\ll} P$, provided that $B \in \mathscr{A}$ is an event for which $P(B) > 0$. Hence, Lemma 5.17 immediately implies the following corollary:

Corollary 5.18 (P-Equivalence Implies P^B-Equivalence)
Let $X, Y\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ be two random variables and let $B \in \mathscr{A}$ with $P(B) > 0$. Then $X \underset{P}{=} Y$ implies $X \underset{P^B}{=} Y$.

Example 5.19 (P^B-Equivalence Does not Imply P-Equivalence) Consider the set $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ with the σ-algebra $\mathscr{A} = \mathscr{P}(\Omega)$, and the set $\Omega' = \{a, b, c, d\}$, with the σ-algebra $\mathscr{A}' = \mathscr{P}(\Omega')$. Furthermore, let $P\colon \mathscr{A} \to [0,1]$ satisfy $P(\{\omega_1\}) = .25$, $P(\{\omega_2\}) = .25$, $P(\{\omega_3\}) = 0$, and $P(\{\omega_4\}) = .50$. Finally, define $X, Y\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ by

$$X(\omega) = \begin{cases} a, & \text{if } \omega = \omega_1 \\ b, & \text{if } \omega = \omega_2 \\ c, & \text{if } \omega = \omega_3 \\ d, & \text{if } \omega = \omega_4 \end{cases} \qquad Y(\omega) = \begin{cases} a, & \text{if } \omega = \omega_1 \\ b, & \text{if } \omega = \omega_2 \\ d, & \text{if } \omega = \omega_3 \\ c, & \text{if } \omega = \omega_4 \end{cases}$$

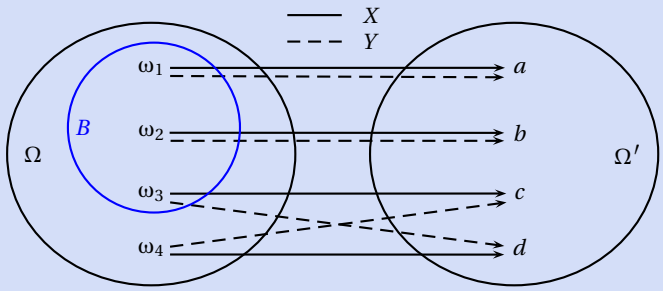


Figure 5.1. Two random variables that are P^B-equivalent if $P(\{\omega_3\}) = 0$

(see Fig. 5.1). If $B = \{\omega_1, \omega_2, \omega_3\}$, then $X \underset{P^B}{=} Y$, but neither $X = Y$ nor $X \underset{P}{=} Y$ (see Exercise 5-4). Therefore, equivalence with respect to P^B does not imply equivalence with respect to P. ◁

Theorem 2.84 on the equivalence of image measures immediately implies the following corollary on the equivalence of the distributions of two P-equivalent random variables:

Corollary 5.20 (P-Equivalence Implies Equal Distributions)
Let $X, Y : (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ be random variables with distributions P_X and P_Y, respectively. If $X \underset{P}{=} Y$, then $P_X = P_Y$.

In other words, if X and Y are P-equivalent, then they are identically distributed. Note, however, that identical distributions of X and Y do not imply that X and Y are P-equivalent.

In chapter 6 we shall see that Corollary 5.20 implies that also the expectations, variances, and other moments of X and Y are identical if $X, Y : (\Omega, \mathcal{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ are P-equivalent numerical random variables, provided that the expectations, variances, and other moments of X and Y exist.

The following corollary is an immediate implication of Theorem 2.85.

Corollary 5.21 (P-Equivalence of Compositions)
Let $X : (\Omega, \mathcal{A}, P) \to (\Omega'_X, \mathcal{A}'_X)$ be a random variable with distribution P_X and let $g, g^* : (\Omega'_X, \mathcal{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ be measurable functions. Then:

(i) $g(X) \underset{P}{=} g^*(X) \quad \Leftrightarrow \quad g \underset{P_X}{=} g^*$.

(ii) $g(X) \underset{P}{\leq} g^*(X) \quad \Leftrightarrow \quad g \underset{P_X}{\leq} g^*$.

(iii) $g(X) \underset{P}{\leqq} g^*(X) \quad \Leftrightarrow \quad g \underset{P_X}{\leqq} g^*$.

(Proof p. 185)

Remark 5.22 (Alternative Notations) Note that

$$g \underset{P_X}{=} g^* \quad \Leftrightarrow \quad g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X, \tag{5.11}$$

$$g \underset{P_X}{<} g^* \quad \Leftrightarrow \quad g(x) < g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X, \tag{5.12}$$

$$g \underset{P_X}{\leq} g^* \quad \Leftrightarrow \quad g(x) \leq g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \tag{5.13}$$

◁

5.2.2 *P*-Equivalence, P^B-Equivalence, and Absolute Continuity

Now we consider the relationship between equivalence of two random variables $X, Y\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ with respect to P and P^B, and absolute continuity of P_X with respect to P_X^B, the distribution of X with respect to the conditional-probability measure P^B. Remember, for $B \in \mathscr{A}$ and $P(B) > 0$, we defined the B-conditional probability measure P^B (see Def. 4.24). Referring to such a measure, $X \underset{P^B}{=} Y$ means

$$\exists\, A \in \mathscr{A}\colon \ \Big(\forall\, \omega \in \Omega \setminus A\colon \ X(\omega) = Y(\omega) \quad \text{and} \quad P^B(A) = 0\Big), \tag{5.14}$$

[see (5.8)]. If B denotes the event $\{X{=}x\} := \{\omega \in \Omega\colon X(\omega){=}x\}$, then we define $P^{X=x} := P^B$ and call it the $(X{=}x)$-conditional probability measure on $(\Omega, \mathscr{A})$.

Lemma 5.23 (An Implication of Absolute Continuity)
Let $X, Y\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ be random variables that are measurable with respect to the σ-algebra $\mathscr{C} \subset \mathscr{A}$, let $B \in \mathscr{A}$, and $P(B) > 0$. If $X \underset{P^B}{=} Y$ and $P \underset{\mathscr{C}}{\ll} P^B$, then $X \underset{P}{=} Y$.

(Proof p. 185)

Example 5.24 (No Treatment For Joe) Consider Table 5.2. In this example we define the set $\Omega_U = \{Joe, Jim, Ann\}$ and

$$\mathscr{A}_U = \mathscr{P}(\Omega_U) = \big\{\{Joe\}, \{Jim\}, \{Ann\}, \{Joe, Jim\}, \{Joe, Ann\}, \{Jim, Ann\}, \Omega_U, \emptyset\big\}.$$

Using these sets, not only $U\colon (\Omega, \mathscr{A}, P) \to (\Omega_U, \mathscr{A}_U)$ is a random variable, but also $U^*\colon (\Omega, \mathscr{A}, P) \to (\Omega_U, \mathscr{A}_U)$ defined in Table 5.2. Now the distribution of U is specified by $P_U(\{Joe\}) = .5$, $P_U(\{Jim\}) = 0$, and $P_U(\{Ann\}) = .5$. The probabilities of the other five elements of $\mathscr{A}_U$ are obtained using Rule (x) of Box 4.1. Furthermore, $P_{U^*} = P_U$.

Considering the measure $P^{X=x}$ we find $U \underset{P^{X=0}}{=} U^*$, because

$$P^{X=0}(\{U \neq U^*\}) = P^{X=0}\big(\{(Joe, yes, +), (Joe, yes, -)\}\big) = 0.$$

Furthermore, there are only two sets $A \in \sigma(U)$ with $P^{X=0}(A) = 0$. These are the sets $U^{-1}(\{Jim\}) = \{(Joe, yes, +), (Joe, yes, -)\}$ and $\emptyset$, and for these sets we find

$P\big(U^{-1}(\{Jim\})\big) = P(\emptyset) = 0$. Hence, $P \underset{\sigma(U)}{\ll} P^{X=0}$, and according to Lemma 5.23 this implies $U \underset{P}{=} U^*$. In fact, we find

$$P(\{U \neq U^*\}) \;=\; P\big(\{(Joe, yes, +), (Joe, yes, -)\}\big) \;=\; 0.$$

◁

Lemma 5.25 (Absolute Continuity)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable and $B \in \mathscr{A}$ with $P(B) > 0$. Then

$$P \underset{\sigma(X)}{\ll} P^B \;\Leftrightarrow\; P_X \underset{\mathscr{A}'_X}{\ll} P_X^B. \tag{5.15}$$

(Proof p. 186)

Example 5.26 (No Treatment For Joe – continued) In Example 5.24, we already found

$$P \underset{\sigma(U)}{\ll} P^{X=0}.$$

There are only two sets $A' \in \mathscr{A}_U$ with $P_U^{X=0}(A') = 0$, the sets $\{Jim\}$ and $\emptyset$, and for these sets we find $P_U(\{Jim\}) = P_U(\emptyset) = 0$. Hence, in this example,

$$P_U \underset{\mathscr{A}_U}{\ll} P_U^{X=0}$$

holds as well. ◁

Lemmas 5.25 and 4.27 immediately imply the following corollary.

Corollary 5.27 (Null-Set Equivalence)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable and $B \in \mathscr{A}$ with $P(B) > 0$. Then P and P^B are null-set equivalent on $\big(\Omega, \sigma(X)\big)$ if and only if P_X an P_X^B are null-set equivalent on $(\Omega'_X, \mathscr{A}'_X)$.

According to Lemma 5.25 (i), absolute continuity of P_X^B with respect to P_X always holds. In other words, $P_X^B \underset{\mathscr{A}'_X}{\ll} P_X$, which is equivalent to

$$\forall\, A' \in \mathscr{A}'_X\colon\; P_X(A') = 0 \;\Rightarrow\; P_X^B(A') = 0, \tag{5.16}$$

always holds. In contrast, $P_X \underset{\mathscr{A}'_X}{\ll} P_X^B$ is *not necessarily true*.

Example 5.28 (No Treatment For Joe – continued) Table 5.2 displays an example illustrating absolute continuity of P_U with respect to P_U^B for a discrete random variable. Consider the event $B = \{X=1\} = \{\omega \in \Omega\colon X(\omega) = 1\}$. Using this notation, P_U is not absolutely continuous with respect to $P_U^{X=1}$, i. e., $P_U \underset{\mathscr{A}_U}{\ll} P_U^{X=1}$ does *not hold*. In contrast, $P_U \underset{\mathscr{A}_U}{\ll} P_U^{X=0}$ *does hold*.

Table 5.2. No Treatment for Joe

Outcomes ω				Observables					
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Person variable U^*	Treatment variable X	Outcome variable Y	$P^{X=0}(\{\omega\})$ (rounded)	$P^{X=1}(\{\omega\})$
(*Joe*,	*no*,	−)	.152	*Joe*	*Joe*	0	0	.245	0
(*Joe*,	*no*,	+)	.348	*Joe*	*Joe*	0	1	.561	0
(*Joe*,	*yes*,	−)	0	*Joe*	*Jim*	1	0	0	0
(*Joe*,	*yes*,	+)	0	*Joe*	*Jim*	1	1	0	0
(*Ann*,	*no*,	−)	.096	*Ann*	*Ann*	0	0	.155	0
(*Ann*,	*no*,	+)	.024	*Ann*	*Ann*	0	1	.039	0
(*Ann*,	*yes*,	−)	.228	*Ann*	*Ann*	1	0	0	.60
(*Ann*,	*yes*,	+)	.152	*Ann*	*Ann*	1	1	0	.40

In this example, the eight elements of Ω are listed in the first column of the table. Furthermore, we choose $\mathscr{A} = \mathscr{P}(\Omega)$ and the probability measure on $(\Omega, \mathscr{A})$ is specified by the probabilities of the singletons $\{\omega\}$ specified in the second column of the table [see Box 4.1 (x)]. The random variables $U\colon (\Omega, \mathscr{A}, P) \to (\Omega_U, \mathscr{A}_U)$, with $\Omega_U = \{Joe, Ann\}$, $\mathscr{A}_U = \mathscr{P}(\Omega_U)$, and $X, Y\colon (\Omega, \mathscr{A}, P) \to \big(\Omega', \mathscr{P}(\Omega')\big)$ with $\Omega' = \{0,1\}$, are specified in Table 5.2. (The random variable U^* has been used in Example 5.24.) Note that the distribution of U is:

$$P_U(\{Joe\}) = P\big(\{(Joe, no, -), (Joe, no, +), (Joe, yes, +), (Joe, yes, -)\}\big) = .5,$$
$$P_U(\{Ann\}) = P\big(\{(Ann, no, -), (Ann, no, +), (Ann, yes, +), (Ann, yes, -)\}\big) = .5,$$

$P_U(\Omega_U) = 1$, and $P_U(\emptyset) = 0$.

Now, we compute the $(X=1)$-conditional probabilities of the elementary events:

$$P^{X=1}(\{\omega_1\}) = P^{X=1}\big(\{(Joe, no, -)\}\big) = \frac{P\big(\{(Joe, no, -)\} \cap \{X=1\}\big)}{P(X=1)}$$
$$= 0/(.228 + .152) = 0,$$

and the same result is obtained for ω_2 to ω_6. In contrast,

$$P^{X=1}(\{\omega_7\}) = P^{X=1}\big(\{(Ann, yes, -)\}\big) = \frac{P\big(\{(Ann, yes, -)\} \cap \{X=1\}\big)}{P(X=1)}$$
$$= .228/(.228 + .152) = .60,$$

and

$$P^{X=1}(\{\omega_8\}) = P^{X=1}\big(\{(Ann, yes, +)\}\big) = \frac{P\big(\{(Ann, yes, +)\} \cap \{X=1\}\big)}{P(X=1)}$$
$$= .152/(.228+.152) = .40.$$

These results are displayed in the last column of Table 5.2 and the last but one column shows the probabilities $P^{X=0}(\{\omega\})$ of the singletons with respect to $P^{X=0}$.

Now consider the element $\{Joe\} \in \mathscr{A}_U$. Inspecting the last and the second columns of Table 5.2 shows that

$$P_U^{X=1}(\{Joe\}) = 0 \quad \text{and} \quad P_U(\{Joe\}) = .5.$$

According to Definition 3.70 (i), this implies that $P_U \underset{\mathscr{A}_U}{\ll} P_U^{X=1}$ does *not hold*. In contrast, none of the four elements $A' \in \mathscr{A}_U$ satisfies

$$P_U^{X=0}(A') = 0 \quad \text{and} \quad P_U(A') \neq 0.$$

Therefore, in this example, $P_U \underset{\mathscr{A}_U}{\ll} P_U^{X=0}$ *does hold*. ◁

5.3 Multivariate Random Variable

Univariate random variables take on their values in sets such as $\Omega' \subset \overline{\mathbb{R}}$, $\Omega' = \{male, female\}$, or $\Omega' = \{low, medium, high\}$, whereas bivariate random variables take on their values in sets such as $\Omega' \subset \overline{\mathbb{R}}^2$ or

$$\Omega' = \{male, female\} \times \{low, medium, high\}.$$

The values of bivariate random variables are pairs such as $(5,8)$ or $(male, low)$. The values of n-variate random variables are n-tuples. If X takes on values such as *male* or $(male, low)$, then we call X *qualitative*. If X takes on values in a subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, we call it *n-variate real-valued*. If X takes on values in a subset of $\overline{\mathbb{R}}^n$, $n \in \mathbb{N}$, we call it *n-variate numerical*.

Remark 5.29 (Joint and Marginal Distributions) Definition 5.1 also applies to an n-variate random variable X, i. e., to a random variable

$$X = (X_1, \ldots, X_n) \colon (\Omega, \mathscr{A}, P) \to \Big(\bigtimes_{i=1}^n \Omega_i', \bigotimes_{i=1}^n \mathscr{A}_i'\Big) \tag{5.17}$$

that consists of n univariate random variables $X_i \colon (\Omega, \mathscr{A}, P) \to (\Omega_i', \mathscr{A}_i')$. Hence,

$$X(\omega) = [X_1(\omega), \ldots, X_n(\omega)], \quad \forall\, \omega \in \Omega. \tag{5.18}$$

The distribution $P_X = P_{X_1,\ldots,X_n}$ of X is also called the *joint distribution* of the random variables X_i, $i = 1, \ldots, n$.

Because $\pi_i(X_1, \ldots, X_n) = X_i$,

$$P_{X_i} = P_{\pi_i(X_1,\ldots,X_n)}, \qquad i = 1,\ldots,n, \tag{5.19}$$

[see Eq. (5.5) and Eq. (2.20) defining the projection π_i]. In this context, P_{X_i} is called the (one-dimensional) *marginal distribution of* X_i. Equation (5.19) shows that the joint distribution uniquely determines all marginal distributions, but not vice versa! More specifically, for $i = 1,\ldots,n$,

$$P_{X_i}(A_i') = P_{X_1,\ldots,X_n}(\Omega_1' \times \ldots \times \Omega_{i-1}' \times A_i' \times \Omega_{i+1}' \times \ldots \times \Omega_n'), \quad \forall\, A_i' \in \mathscr{A}_i'. \tag{5.20}$$

Analogously, we may also describe the marginal distribution of $(X_{i_1},\ldots,X_{i_m})$, where $\{i_1,\ldots,i_m\} \subset \{1,\ldots,n\}$. ◁

Remark 5.30 (Joint Distribution vs. Other Quantities) The joint distribution of a multivariate random variable contains the essential information about the random variables $X_1,\ldots,X_n$. All other quantities such as expectations $E(X_i)$, variances $Var(X_i)$, covariances $Cov(X_i,X_j)$, or regressions such as $E(X_1 \mid X_2,\ldots,X_n)$, which is introduced in succeeding chapters, are determined by the joint distribution, and usually they contain less information. Nevertheless, these other quantities often reveal certain properties of a multivariate random variable more clearly than the joint distribution. ◁

Example 5.31 (Flipping two Coins – continued) In Example 2.2.2, we considered the random experiment of flipping two coins and defined a random variable X representing with its values the number of flipping heads. Additional to X we may also define the random variables $X_1, X_2 \colon (\Omega,\mathscr{A},P) \to (\{0,1\},\mathscr{P}(\{0,1\})$ by

$$X_1(\omega) = \begin{cases} 1, \text{ if } \omega \in \{(h,t),(h,h)\} \\ 0, \text{ if } \omega \in \{(t,h),(t,t)\} \end{cases} \tag{5.21}$$

and

$$X_2(\omega) = \begin{cases} 1, \text{ if } \omega \in \{(t,h),(h,h)\} \\ 0, \text{ if } \omega \in \{(h,t),(t,t)\}\,. \end{cases} \tag{5.22}$$

They indicate with their value 1 if *heads* are flipped at the first and second flip, respectively. Obviously, $X = X_1 + X_2$. Furthermore,

$$(X_1,X_2) \colon (\Omega,\mathscr{A},P) \to \big(\{0,1\}\times\{0,1\}, \mathscr{P}(\{0,1\}) \otimes \mathscr{P}(\{0,1\})\big)$$

is a two-dimensional random variable with values $(0,0),(0,1),(1,0)$, and $(1,1)$. The joint distribution P_{X_1,X_2} is uniquely defined by

$$P_{X_1,X_2}\big(\{(i,j)\}\big) = \frac{1}{4}, \quad \forall\, i,j = 0,1.$$

The marginal distribution of X_1 is

$$P_{X_1}(\{i\}) = P_{X_1,X_2}\big(\{(i,0)\}\big) + P_{X_1,X_2}\big(\{(i,1)\}\big) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \qquad i = 0,1,$$

$P_{X_1}(\{0,1\}) = 1$, and $P_{X_1}(\emptyset) = 0$. Obviously, P_{X_1} is completely determined by the joint distribution P_{X_1,X_2}, and the same applies to the marginal distribution P_{X_2}. ◁

Example 5.32 (Tom, Jim, and Kate – continued) The second column of Table 5.1 also displays the probabilities $P_{U,X,Y}\big(\{(u,x,y)\}\big) = P(\{\omega\})$ of the three-dimensional random variable (U,X,Y) that maps the elements $\omega \in \Omega$ onto the set

$$\Omega' := \{\textit{Tom, Jim, Kate}\} \times \{0,1,2\} \times \{0,1\}$$

on which we consider the σ-algebra

$$\mathscr{A}' := \mathscr{P}(\{\textit{Tom, Jim, Kate}\}) \otimes \mathscr{P}(\{0,1,2\}) \otimes \mathscr{P}(\{0,1\}).$$

The probabilities $P_{U,X,Y}\big(\{(u,x,y)\}\big)$, $(u,x,y) \in \Omega'$, uniquely determine the joint distribution $P_{U,X,Y}$ as well as the one-dimensional marginal distributions P_U, P_X, P_Y, and the two-dimensional marginal distributions $P_{U,X}$, $P_{U,Y}$, and $P_{X,Y}$. ◁

5.4 Independence of Random Variables

The concepts of independence of events and of set systems, i. e., of sets of events, which have been introduced in Definition 4.35, can be used to define *stochastic independence of random variables*. Remember that

$$\sigma(X) := X^{-1}(\mathscr{A}') := \{X^{-1}(A') \colon A' \in \mathscr{A}'\}$$

is a σ-algebra on Ω, called the σ-*algebra generated by* X (see Def. 2.26). Hence, we can define the random variables $X_1 \colon (\Omega, \mathscr{A}, P) \to (\Omega_1', \mathscr{A}_1')$ and $X_2 \colon (\Omega, \mathscr{A}, P) \to (\Omega_2', \mathscr{A}_2')$ to be *P-independent* if $X_1^{-1}(\mathscr{A}_1')$ and $X_2^{-1}(\mathscr{A}_2')$ are P-independent. In other words, X_1 and X_2 are defined to be P-independent, if

$$P(A \cap B) = P(A) \cdot P(B), \quad \forall\, (A,B) \in X_1^{-1}(\mathscr{A}_1') \times X_2^{-1}(\mathscr{A}_2'). \tag{5.23}$$

Using the notation introduced in Remark 5.4 and

$$P(X_1 \in A',\, X_2 \in B') := P(\{X_1 \in A'\} \cap \{X_2 \in B'\}), \tag{5.24}$$

this equation is equivalent to

$$P(X_1 \in A',\, X_2 \in B') = P(X_1 \in A') \cdot P(X_2 \in B'), \quad \forall (A', B') \in \mathscr{A}_1' \times \mathscr{A}_2'. \tag{5.25}$$

Independence of the random variables X_1 and X_2 with respect to P is denoted by $X_1 \perp\!\!\!\perp_P X_2$.

Example 5.33 (Joe and Ann With Random Assignment – continued) In Table 2.2 we presented the random experiment of drawing a person from a set of persons, $\Omega_U = \{\textit{Joe}, \textit{Ann}\}$, randomly assigning the drawn person to one of two treatment conditions represented by the elements of the set $\Omega_X = \{\textit{yes}, \textit{no}\}$, and observing success or failure, represented by the elements of the set $\Omega_Y = \{-,+\}$. Hence, the set of possible outcomes of this random experiment is

$$\Omega = \Omega_U \times \Omega_X \times \Omega_Y,$$

which consists of the eight triples listed in the first column of Table 2.2. In that table we defined the three random variables

$$U\colon \big(\Omega,\mathscr{P}(\Omega)\big)\to\big(\Omega_U,\mathscr{P}(\Omega_U)\big) \quad \text{and} \quad X,Y\colon \big(\Omega,\mathscr{P}(\Omega)\big)\to(\Omega',\mathscr{A}'),$$

where $\Omega'=\{0,1\}$ and $\mathscr{A}'=\big\{\{0\},\{1\},\Omega',\emptyset\big\}$. In order to check if Equation (5.25) actually holds, we choose the two sets $\{Joe\}\in\mathscr{P}(\Omega_U)$ and $\{0\}\in\mathscr{A}'$ and compare the probability

$$P(X{=}0,U{=}Joe) \;=\; P(\{(Joe,no,-),(Joe,no,+)\}) \;=\; .3$$

(see the first two rows of Table 2.2) to the product of the two probabilities

$$P(X{=}0) \;=\; P(\{(Joe,no,-),(Joe,no,+),(Ann,no,-),(Ann,no,+)\} \;=\; .6$$

and

$$P(U{=}Joe) \;=\; P(\{(Joe,no,-),(Joe,no,+),(Joe,no,-),(Joe,no,+)\} \;=\; .5.$$

Obviously, Equation (5.25) holds for the pair $(\{0\},\{Joe\})\in\mathscr{A}'\times\mathscr{P}(\Omega_U)$. Repeating the corresponding comparisons for all pairs of $\mathscr{A}'\times\mathscr{P}(\Omega_U)$ shows that Equation (5.25) actually holds in this example (see also Exercise 5-5). ◁

Remark 5.34 (A Methodological Note on Random Assignment) In random experiments such as the one presented in Example 5.33, in which the drawn person is randomly assigned to one of several treatment conditions, we create independence of X and the person variable U. This implies that we create independence of X and all U-measurable mappings $f(U)$, because $\sigma[f(U)]\subset\sigma(U)$. More generally, random assignment of an observational unit (such as a person) creates independence of X and all pretreatment variables. ◁

Using definition 4.35, the following definition extends the concept of independence of two random variables to a family of random variables. This includes a finite sequence of random variables X_i, $i\in I:=\{1,\dots,n\}$, an infinite sequence of random variables X_i, $i\in I:=\{1,2,\dots\}$, and a family $(X_i,i\in I)$ of random variables in which the index set I may be *any* set, including, e. g., $I\subset\mathbb{R}$.

Definition 5.35 (Family of Independent Random Variables)
A family $(X_i,i\in I)$ of random variables $X_i\colon(\Omega,\mathscr{A},P)\to(\Omega_i',\mathscr{A}_i')$ is called a family of P-independent random variables, denoted $\perp\!\!\!\perp_P(X_i,i\in I)$, if $\big(X_i^{-1}(\mathscr{A}_i'),i\in I\big)$ is a family of P-independent σ-algebras.

Remark 5.36 (Independence of Three Random Variables) Hence, three random variables X_1,X_2,X_3 are *independent*, denoted $\perp\!\!\!\perp_P X_1,X_2,X_3$, if and only if

$$\begin{gathered}P(X_1\in A',X_2\in B',X_3\in C') \;=\; P(X_1\in A')\cdot P(X_2\in B')\cdot P(X_3\in C'),\\ \forall(A',B',C')\in\mathscr{A}_1'\times\mathscr{A}_2'\times\mathscr{A}_3'\end{gathered} \tag{5.26}$$

(see Rem. 4.33). Note that pairwise independence of X_1, X_2, X_3 follows from choosing $A' = \Omega_1'$, $B' = \Omega_2'$, or $C' = \Omega_3'$, respectively. ◁

Remark 5.37 (Independence of n Random Variables) Correspondingly, the random variables $X_1, \ldots, X_n$ are *independent*, denoted $\perp\!\!\!\perp_P X_1, \ldots, X_n$, if and only if

$$\begin{aligned} P(X_1 \in A_1', \ldots, X_n \in A_n') \;&=\; P(X_1 \in A_1') \cdot \ldots \cdot P(X_n \in A_n'), \\ &\forall (A_1', \ldots, A_n') \in \mathscr{A}_1' \times \ldots \times \mathscr{A}_n'. \end{aligned} \tag{5.27}$$

◁

Remark 5.38 (Sample) Oftentimes, we assume that $X_1, \ldots, X_n$ is a sequence of independent and identically distributed (abbreviated i. i. d.) random variables (see, e. g., chs. 6 and 8). In statistics, a sequence $X_1, \ldots, X_n$ of i. i. d. random variables is called a *random sample*.

An important example of i. i. d. random variables is treated in the section on Bernoulli trials (see section 8.1.2). ◁

Remark 5.39 (Independence With Respect to a Probability Measure) If there is no ambiguity we also use the term *independence* of events, sets of events, random variables, and sets of random variables. Note that, if Q is another probability measure on $(\Omega, \mathscr{A})$, then events, sets of events, and random variables can be P-independent although they are not Q-independent. ◁

Remark 5.40 (A Random Variable and a Set System) Independence of a set system and a random variable is defined in the same way. A set system $\mathscr{E} \subset \mathscr{A}$ and a random variable $X\colon (\Omega, \mathscr{A}, P) \to (\Omega_X', \mathscr{A}_X')$ are called *independent*, denoted $\mathscr{E} \perp\!\!\!\perp_P X$, if $\mathscr{E}$ and $\sigma(X)$ are independent. Of course, $\mathscr{E}$ can also be a sub-σ-algebra of $\mathscr{A}$. ◁

Remark 5.41 (A Random Variable and a Family of Random Variables) Independence of a random variable X and a family $(Y_i, i \in I)$ of random variables, denoted $X \perp\!\!\!\perp_P (Y_i, i \in I)$, is defined by $X \perp\!\!\!\perp_P \sigma(Y_i, i \in I)$. Note that $X \perp\!\!\!\perp_P (Y_i, i \in I)$ implies $X \perp\!\!\!\perp_P \sigma(Y_i)$, for all $i \in I$. ◁

Remark 5.42 (Equivalent Propositions) The following propositions are equivalent to each other: $1_B \perp\!\!\!\perp_P \mathscr{C}$, $\sigma(\{B\}) \perp\!\!\!\perp_P \mathscr{C}$, $\{B\} \perp\!\!\!\perp_P \mathscr{C}$, $B \perp\!\!\!\perp_P \mathscr{C}$ (see Rem. 4.36 and Exercise 5-6). ◁

In Corollary 5.20 we noted that P-equivalent random variables have identical distributions. According to the following lemma this also has implications for independence of random variables.

Lemma 5.43 (P-Equivalence and Independence)
Let $X_i\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$, $i = 1, 2$, and $Y\colon (\Omega, \mathscr{A}, P) \to (\Omega_Y', \mathscr{A}_Y')$ be random variables. Then

$$\left(X_1 \underset{P}{=} X_2 \;\wedge\; X_1 \perp\!\!\!\perp_P Y\right) \quad \Rightarrow \quad X_2 \perp\!\!\!\perp_P Y. \tag{5.28}$$

(Proof p. 186)

Now we consider the probability measure P and the B-conditional-probability measure P^B on $(\Omega, \mathscr{A})$ (see Def. 4.24). In Lemma 4.27 we have shown that P^B is absolutely continuous (see Def. 3.70) with respect to P on $(\Omega, \mathscr{A})$. In the following lemma we show that P is absolutely continuous with respect to P^B on $(\Omega, \mathscr{C})$, $\mathscr{C} \subset \mathscr{A}$, provided that B and $\mathscr{C}$ are independent.

Lemma 5.44 (Independence and Absolute Continuity)
Let $(\Omega, \mathscr{A}, P)$ be a probability space, let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, and $B \in \mathscr{A}$ with $P(B) > 0$. Then $1_B \underset{P}{\perp\!\!\!\perp} \mathscr{C}$ implies $P \underset{\mathscr{C}}{\ll} P^B$.

(Proof p. 186)

In the following lemma, $P_{X_1} \otimes \ldots \otimes P_{X_n}$ denotes the product measure of the marginal distributions (see Def. 1.66 and Rem. 5.29).

Lemma 5.45 (Independence and Product Measure)
Let $X = (X_1, \ldots, X_n)$ be an n-variate random variable as specified in (5.17). *Then*

$$\underset{P}{\perp\!\!\!\perp} X_1, \ldots, X_n \quad \Leftrightarrow \quad P_{X_1,\ldots,X_n} = P_{X_1} \otimes \ldots \otimes P_{X_n}. \tag{5.29}$$

(Proof p. 187)

Example 5.46 (Tom, Jim, and Kate – continued) In example 5.12 we considered the random variables X and U, which have been constructed such that they are independent. All $8 \cdot 8 = 64$ pairs (A, B) of elements $A \in X^{-1}(\mathscr{A}'_X)$ and $B \in U^{-1}(\mathscr{A}_U)$ satisfy $P(A \cap B) = P(A) \cdot P(B)$. Let us consider, e. g., $A_1 := X^{-1}(\{0\})$ and $B_1 := U^{-1}(\{Tom\})$, $B_2 := U^{-1}(\{Jim\})$, and $B_3 := U^{-1}(\{Kate\})$. Then

$$P(A_1 \cap B_j) = \frac{20}{99}, \quad j = 1,2,3$$

and

$$P(A_1) \cdot P(B_j) = \frac{60}{99} \cdot \frac{1}{3} = \frac{20}{99}, \quad j = 1,2,3.$$

Similarly, considering the events $A_2 := X^{-1}(\{1\})$ and B_j,

$$P(A_2 \cap B_j) = \frac{8}{99}, \quad j = 1,2,3$$

and

$$P(A_2) \cdot P(B_j) = \frac{24}{99} \cdot \frac{1}{3} = \frac{8}{99}, \quad j = 1,2,3.$$

Finally, considering the events $A_3 := X^{-1}(\{2\})$ and B_j, yields

$$P(A_3 \cap B_j) = \frac{5}{99}, \quad j = 1,2,3$$

and

$$P(A_3) \cdot P(B_j) = \frac{15}{99} \cdot \frac{1}{3} = \frac{5}{99}, \quad j = 1,2,3.$$

Because $\emptyset$ and all sets $A \in \mathscr{A}$ are independent, this implies that independence holds for all pairs $(A,B) \in \{A_1, A_2, A_3, \emptyset\} \times \{B_1, B_2, B_3, \emptyset\}$. Furthermore, because

(a) $\mathscr{E}_1 := \{A_1, A_2, A_3, \emptyset\}$ and $\mathscr{E}_2 := \{B_1, B_2, B_3, \emptyset\}$ are $\cap$-stable set systems on $\mathscr{A}$,
(b) $\sigma(\mathscr{E}_1) = X^{-1}(\mathscr{A}'_X)$ and $\sigma(\mathscr{E}_2) = U^{-1}(\mathscr{A}_U)$,

we can conclude that $P(A \cap B) = P(A) \cdot P(B)$ holds for *all* elements $A \in X^{-1}(\mathscr{A}'_X)$ and $B \in U^{-1}(\mathscr{A}_U)$ (see Th. 4.39). Therefore, according to Equation (5.23), X and U are independent. ◁

Lemma 5.47 (Independence of a Constant and a Set of Events)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable and $\mathscr{C} \subset \mathscr{A}$. If $X \underset{P}{=} \alpha$, $\alpha \in \Omega'_X$, then X and $\mathscr{C}$ are independent.

(Proof p. 187)

Now we consider mappings of independent random variables. If two random variables X_1 and X_2 are independent and $f_i\colon (\Omega'_i, \mathscr{A}'_i) \to (\Omega''_i, \mathscr{A}''_i)$, $i = 1,2$, are measurable mappings, then the two random variables $f_1(X_1)$ and $f_2(X_2)$ are independent as well. More generally, if $f_i\colon (\Omega'_i, \mathscr{A}'_i) \to (\Omega''_i, \mathscr{A}''_i)$, $i = 1,\ldots,n$, is a sequence of measurable mappings, then

$$\perp\!\!\!\perp_P f_1(X_1), \ldots, f_n(X_n),$$

i. e., then $f_1(X_1), \ldots, f_n(X_n)$ is a sequence of independent random variables on $(\Omega, \mathscr{A}, P)$, provided that $X_1, \ldots, X_n$ are independent. In the following theorem we generalize this proposition.

Theorem 5.48 (Mappings of Families of Independent Random Variables)
Let $X_i\colon (\Omega, \mathscr{A}, P) \to (\Omega'_i, \mathscr{A}'_i)$, $i = 1,\ldots,n$, be random variables and, for $m \in \mathbb{N}$, let $I_1 = \{1,\ldots,i_1\}$, $I_2 = \{i_1+1,\ldots,i_2\}, \ldots, I_m = \{i_{m-1}+1,\ldots,n\}$. Furthermore, let

$$f_j\colon \Big(\bigtimes_{i \in I_j} \Omega'_i, \bigotimes_{i \in I_j} \mathscr{A}'_i \Big) \to (\Omega''_j, \mathscr{A}''_j), \quad j = 1,\ldots,m,$$

be measurable mappings. If $X_1, \ldots, X_n$ are independent, then

$$f_1(X_1,\ldots,X_{i_1}),\ f_2(X_{i_1+1},\ldots,X_{i_2}),\ldots,\ f_m(X_{i_{m-1}+1},\ldots,X_n)$$

are independent.

For a generalization and a proof see Bauer (1996, Theorem 9.6, p. 63).

Example 5.49 (Sums of Independent Random Variables) Let $X_1, \ldots, X_{2n}$, $n \in \mathbb{N}$, be independent real-valued random variables, then the n random variables

$$X_1 + X_2,\ X_3 + X_4,\ \ldots,\ X_{2n-1} + X_{2n}$$

are independent as well. ◁

Example 5.50 (Tom, Jim, and Kate – continued) In Example 5.46 we showed that the random variables X and U are independent. Now we consider the mappings $f: \{0,1,2\} \to \{0,1\}$ and $g: \Omega_U \to \{male, female\}$ defined by

$$f(x) = \begin{cases} 0, & \text{if } x = 0 \\ 1, & \text{if } x = 1 \text{ or } x = 2 \end{cases}$$

and

$$g(u) = \begin{cases} male, & \text{if } u = Tom \text{ or } u = Jim \\ female, & \text{if } u = Kate, \end{cases}$$

respectively. According to Theorem 5.48 the mappings $f(X)$ (control vs. any of the two treatments) and $g(U)$ (*sex*) are independent as well (see Exercise 5-7). ◁

Remark 5.51 (Conditional Independence of Random Variables Given an Event) In chapter 4 we also considered *conditional* independence of events and families of events *given an event* B. If, for random variables $X_1, \ldots, X_n$ (or, more generally, families of random variables) we consider the set systems $\sigma(X_1), \ldots, \sigma(X_n)$, then we can use Definition 4.44 in order to define conditional independence of $X_1, \ldots, X_n$ given an event B, presuming $P(B) > 0$. According to Remark 4.42, conditional independence given B is equivalent to independence (see Def. 5.35) with respect to the probability measure P^B. In chapter 16 we generalize this concept and study it in more detail. ◁

5.5 Probability Function of a Discrete Random Variable

The distribution of a discrete random variable can be described by its *probability function* that is now introduced. Remember, if $X: (\Omega, \mathcal{A}, P) \to (\Omega'_X, \mathcal{A}'_X)$ is a random variable, then the distribution P_X of X is a probability measure on $(\Omega'_X, \mathcal{A}'_X)$. Furthermore, if there is a finite or countable set $\Omega'_0 \subset \Omega'_X$ with $P_X(\Omega'_0) = 1$ and $\{x\} \in \mathcal{A}'_X$ for all $x \in \Omega'_X$, then σ-additivity of P_X implies that P_X is already defined by the values $P_X(\{x\})$, $x \in \Omega'_0$ [see Rule (x) in Box 4.1]. This justifies the following definition:

Definition 5.52 (Discrete Random Variable and its Probability Function)
Let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable and assume that $\Omega'_0\subset\Omega'_X$ is finite or countable with $P_X(\Omega'_0)=1$ and $\{x\}\in\mathscr{A}'_X$ for all $x\in\Omega'_0$. Then X and its distribution P_X are called $discrete$, and the function $p_X:\Omega'_X\to[0,1]$ defined by

$$p_X(x) = \begin{cases} P_X(\{x\}), & \text{if } x\in\Omega'_0,\\ 0, & \text{if } x\in\Omega'_X\setminus\Omega'_0, \end{cases} \tag{5.30}$$

is called the $probability\ function$ of X.

Remark 5.53 (Notation) Note that $P(X{=}x) = p_X(x)$, using the notation introduced in Remark 5.4. ◁

Remark 5.54 (Probability Function vs. Distribution) The *distribution* P_X is defined for *every* random variable, whereas the *probability function* p_X only applies to *discrete* random variables. While P_X assigns probabilities to *subsets* of the codomain Ω'_X of X, the probability function p_X assigns a probability to each *element* x in Ω'_X. Note that p_X is a real-valued random variable on the probability space $(\Omega'_X,\mathscr{A}'_X,P_X)$. ◁

Remark 5.55 (The Probability Function Uniquely Determines the Distribution) Note that σ-additivity of the probability measure P_X implies that P_X is uniquely determined by the probability function p_X. Vice versa, according to Definition 5.52, P_X defines p_X. Hence, if $X,Y\colon(\Omega,\mathscr{A},P)\to(\Omega',\mathscr{A}')$ are discrete random variables, then

$$p_X = p_Y \quad\Leftrightarrow\quad P_X = P_Y. \tag{5.31}$$

◁

Remark 5.56 (Probability Function of a Discrete Distribution) Note that (5.31) allows us to use the term *probability function of a discrete distribution* instead of *probability function of a discrete random variable.* ◁

Lemma 5.57 (Characterizations of a Discrete Random Variable)
Let $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable.

(i) Then X is discrete if and only if there is a finite or countable $\Omega'_0\subset\Omega'_X$ such that $\{x\}\in\mathscr{A}'_X$ for all $x\in\Omega'_0$, $P_X(\Omega'_0)=1$, and

$$1_{X=x_0} \underset{P}{=} 1 - \sum_{x\in\Omega'_0\setminus\{x_0\}} 1_{X=x}, \quad \forall x_0\in\Omega'_0. \tag{5.32}$$

(ii) Now assume that $X\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ is real-valued. Then X is discrete if and only if there is a finite or countable $\Omega'\subset\mathbb{R}$ such that

$$X \underset{P}{=} \sum_{x\in\Omega'} x\cdot 1_{X=x}. \tag{5.33}$$

(Proof p. 188)

Remark 5.58 (A Caveat) Note that Equation (5.32) is equivalent to $P(\Omega_0') = 1$. In proposition (ii), we can choose Ω' such that $0 \notin \Omega'$ even if $P(X=0) > 0$. In this case, the set Ω_0' referred to in (i) can be chosen such that $\Omega_0' := \Omega' \cup \{0\}$. ◁

Corollary 5.59 (Discrete Real-Valued Random Variable)
Assume that $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ is a real-valued random variable. Then X is discrete if and only if the following two conditions hold:

(a) $\Omega_>' := \{x \in \Omega_X' \colon P(X=x) > 0\}$ *is finite or countable.*

(b) $X \underset{P}{=} \sum_{x \in \Omega_>'} x \cdot 1_{X=x}$.

(Proof p. 188)

Condition (b) may equivalently be written

$$X \underset{P}{=} \sum_{\substack{x \in \Omega_X' \\ P(X=x)>0}} x \cdot 1_{X=x}. \tag{5.34}$$

Example 5.60 (Flipping two Coins – continued) Consider again Example 5.31 and let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega_X', \mathscr{A}_X')$ denote the *number of flipping heads*. If we assume that $P(\{\omega\}) = \frac{1}{4}$ for all $\omega \in \Omega$, then

$$p_X(0) = P_X(\{0\}) = \frac{1}{4} \qquad p_X(1) = P_X(\{1\}) = \frac{1}{2} \qquad p_X(2) = P_X(\{2\}) = \frac{1}{4}$$

are the values of the probability function p_X of X. They are the probabilities of the events that X takes on the value 0, 1, and 2, respectively. For simplicity, we also denote these probabilities by $P(X=0)$, $P(X=1)$, and $P(X=2)$. In this example, we may choose different measurable spaces $(\Omega_X', \mathscr{A}_X')$. If we choose $(\Omega_X', \mathscr{A}_X') = \big(\{0,1,2\}, \mathscr{P}(\{0,1,2\})\big)$, then $\Omega_0' = \Omega_X'$ (see Def. 5.52). If we choose $(\Omega_X', \mathscr{A}_X') = (\mathbb{R}, \mathscr{B})$, then $\Omega_0' = \{0,1,2\}$ and $\mathscr{A}_X'|_{\Omega_0'} = \mathscr{B}|_{\Omega_0'} = \mathscr{P}(\Omega_0')$ (see Remark 1.29). According to Equation (5.33),

$$X = 0 \cdot 1_{X=0} + 1 \cdot 1_{X=1} + 2 \cdot 1_{X=2}.$$

This example is a special case of a random variable with a binomial distribution. The general case is treated in Definition 8.7. Other examples of a discrete random variable and their probability function are random variables that have a *Poisson distribution* or a *geometric distribution*. In both cases, the random variable considered takes on an infinite and countable number of values, each of which has a probability greater than 0. These examples are treated in chapter 8 (see Defs. 8.14 and 8.20). ◁

Example 5.61 (Flipping two Coins – continued) In Example 5.31 we introduced the random variables X_1 and X_2, which indicate if we flip *heads* in the first and second trial, respectively. The probability function of the bivariate random variable $X = (X_1, X_2)$ is

$$p_{X_1,X_2}(x_1,x_2) = \frac{1}{4}, \quad \forall\,(x_1,x_2) \in \{0,1\}^2.$$

◁

Lemma 5.62 (Probability Function of a Marginal Distribution)
Consider a multivariate random variable $X = (X_1,\dots,X_n)$ as specified in (5.17) *and assume that there is a finite or countable set $\Omega_0' \subset \bigtimes_{i=1}^n \Omega_i'$ with $\{x\} \in \bigotimes_{i=1}^n \mathscr{A}_i'$ for all $x \in \Omega_0'$. Furthermore, for all $x_i \in \Omega_i'$, define*

$$\Omega_{0,x_i}' := \big\{(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n) : (x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n) \in \Omega_0'\big\}.$$

Then, for all $x_i \in \Omega_i'$,

$$p_{X_i}(x_i) = \sum_{(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)\in\Omega_{0,x_i}'} p_X(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n), \tag{5.35}$$

where p_{X_i} denotes the probability function of X_i, $i = 1,\dots,n$, which is also called the marginal probability function of X_i.

(Proof p. 189)

Now we turn to a condition that is equivalent to independence of discrete random variables.

Remark 5.63 (Support Sets of Discrete Random Variables) Under the assumptions of Lemma 5.62 we define the 'support sets'

$$\Omega_{0,i}' := \{x_i \in \Omega_i' : p_{X_i}(x_i) > 0\}, \quad i = 1,\dots,n. \tag{5.36}$$

Obviously, $\Omega_{0,i}'$ is finite or countable for all $i = 1,\dots,n$. Hence, $\Omega_{sn}' := \bigtimes_{i=1}^n \Omega_{0,i}'$ is finite or countable as well. Furthermore, $P(X \in \Omega_{sn}') = 1$, because, for $(x_1,\dots,x_n) \in \Omega_0' \setminus \Omega_{sn}'$, there is at least one i such that $X_i \notin \Omega_{0,i}'$ and therefore $p_{X_i} = P(X_i = x_i) = 0$, which implies $P(X_1 = x_1,\dots,X_i = x_i,\dots,X_n = x_n) = 0$. ◁

Lemma 5.64 (A Condition Equivalent to Independence)
Let X be a multivariate random variable as specified in (5.17) *and assume that there is a finite or countable set $\Omega_0' \subset \bigtimes_{i=1}^n \Omega_i'$ with $P(\Omega_0') = 1$ and $\{x\} \in \mathscr{A}_X'$, for all $x \in \Omega_0'$. Furthermore, let $p_X, p_{X_1},\dots,p_{X_n}$ denote the probability functions of $X, X_1,\dots,X_n$, respectively, and let $\Omega_{0,1}',\dots,\Omega_{0,n}'$ be the sets defined in* (5.36). *Then $X_1,\dots,X_n$ are independent if and only if*

$$p_X(x_1,\ldots,x_n) \;=\; p_{X_1}(x_1)\cdot\ldots\cdot p_{X_n}(x_n), \quad \forall (x_1,\ldots,x_n)\in \bigtimes_{i=1}^{n}\Omega'_{0,i}\,. \tag{5.37}$$

(Proof p. 189)

Note that, in Lemmas 5.62 and 5.64, the set Ω'_0 is not necessarily a Cartesian product. We only require that it is *a subset* of a Cartesian product (see Exercise 5-9).

In the following section we shall see that a probability function is a special probability density (see Th. 5.73).

5.6 Probability Density With Respect to a Measure

Some probability measures can also be described by a density with respect to the Lebesgue measure on $(\mathbb{R},\mathscr{B})$ or the counting measure on $\big(\Omega,\mathscr{P}(\Omega)\big)$, where Ω is a finite or countable set (see Th. 3.65). Such a density is useful for explicit numerical calculations and comparing distributions to each other. Furthermore, the conditional-probability measure P^B can be interpreted as a measure with density with respect to P. We start by translating some concepts and results of chapter 3 to probability measures.

5.6.1 General Concepts and Properties

According to Theorem 3.65 and Definition 3.66, a nonnegative measurable function $f\colon(\Omega,\mathscr{A},\mu)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ is called a *density of ν with respect to μ*, if

$$\nu(A) \;=\; \int_A f\;d\mu, \quad \forall\, A\in\mathscr{A}. \tag{5.38}$$

The function $\nu\colon\mathscr{A}\to\overline{\mathbb{R}}$ defined by (5.38) is a measure, which is also denoted by $f\odot\mu$. Hence, $f\odot\mu(A) \;=\int_A f\;d\mu$, $\forall\, A\in\mathscr{A}$. Theorems 3.65, 3.72, and Remark 3.73 imply the following corollary.

Corollary 5.65 (Probability Measure With Density)
Let $(\Omega,\mathscr{A},\mu)$ be a measure space. If $f\colon(\Omega,\mathscr{A},\mu)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ is a nonnegative μ-integrable function with $\int f\,d\mu=1$, then $P:=f\odot\mu$ is a probability measure on $(\Omega,\mathscr{A})$.

Corollary 5.65 justifies the following definition.

Definition 5.66 (Probability Density)
Let $(\Omega,\mathscr{A},P)$ be a probability space and μ a measure on $(\Omega,\mathscr{A})$. If $P=f\odot\mu$, then f is called a probability density of P with respect to μ.

Remark 5.67 (Probability Density of a Random Variable) Consider the random variable $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ and let μ be a measure on $(\Omega'_X,\mathscr{A}'_X)$. If $P_X = f_X\odot\mu$, then f_X is also referred to as a *probability density of X with respect to μ*. ◁

Applying Equation (5.38) yields the following corollary.

Corollary 5.68 (Characterizing the Probability Measure by a Density) *Let $f\colon(\Omega,\mathscr{A},\mu)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be μ-integrable and nonnegative and P a probability measure on $(\Omega,\mathscr{A})$. Then f is a (probability) density of P with respect to μ if and only if it satisfies*

$$P(A) = \int_A f\,d\mu, \quad \forall\, A\in\mathscr{A}. \tag{5.39}$$

Theorem 3.68 (a) and (c) imply the following corollary.

Corollary 5.69 (Probability Densities are μ-Equivalent) *Let μ and P be measures on the measurable space $(\Omega,\mathscr{A})$, where P is a probability measure. If $f, f^*\colon(\Omega,\mathscr{A},\mu)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ are probability densities of P with respect to μ, then $f \underset{\mu}{=} f^*$.*

Now we translate the Radon-Nikodym Theorem (see Th. 3.72), which yields the following corollary.

Corollary 5.70 (An Implication of the Radon-Nikodym Theorem) *Let $(\Omega,\mathscr{A},\mu)$ be a measure space. If μ is σ-finite, P is a probability measure on $(\Omega,\mathscr{A})$, and $P\underset{\mathscr{A}}{\ll}\mu$, then there is a probability density f of P with respect to μ (also called a Radon-Nikodym derivative), i. e.,*

$$f = \frac{dP}{d\mu}. \tag{5.40}$$

Example 5.71 (Conditional-Probability Measure) In Lemma 4.27 (ii) we showed that $1_B/P(B)$ is a density of P^B with respect to P. ◁

5.6.2 Density of a Discrete Random Variable

As a special case, we consider a discrete random variable (see section 5.5).

Remark 5.72 (A Sum of Dirac Measures) Let the assumptions of Definition 5.52 hold, i. e., let $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable and suppose $\Omega'_0\subset\Omega'_X$

is finite or countable with $P_X(\Omega'_0) = 1$ and $\{x\} \in \mathscr{A}'_X$ for all $x \in \Omega'_0$. Furthermore, define

$$\mu := \sum_{x \in \Omega'_0} \delta_x, \tag{5.41}$$

the sum of Dirac measures at x, $x \in \Omega'_0$. According to Example 1.57, μ is a measure on $(\Omega'_X, \mathscr{A}'_X)$, which is σ-finite. ◁

The following theorem asserts that the probability function p_X is a density of P_X with respect to μ.

Theorem 5.73 (The Probability Function is a Density)
Let the assumptions of Definition 5.52 be satisfied and let μ be defined by Equation (5.41)*. Then:*

(i) $P_X \underset{\mathscr{A}'_X}{\ll} \mu$.

(ii) The probability function p_X is a density of P_X with respect to μ, i. e.,

$$p_X = \frac{dP_X}{d\mu} \tag{5.42}$$

and

$$\begin{aligned} P_X(A') &= \int_{A'} p_X \, d\mu && (5.43)\\ &= \sum_{x \in \Omega'_0} 1_{A'}(x) \cdot p_X(x) && (5.44)\\ &= \sum_{x \in A'} p_X(x), \quad \forall\, A' \in \mathscr{A}'_X. && (5.45) \end{aligned}$$

(Proof p. 190)

Hence, each probability $P_X(A')$, $A' \in \mathscr{A}'_X$, can be computed from the probability function p_X.

5.6.3 Density of a Bivariate Random Variable

Now we consider bivariate random variables. However, extending the following notation and propositions to general multivariate random variables is straightforward.

Lemma 5.74 (Absolute Continuity of Marginal Distributions)
Let $(X,Y)\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X \times \Omega'_Y, \mathscr{A}'_X \otimes \mathscr{A}'_Y)$ be a random variable and suppose that $(\Omega'_X, \mathscr{A}'_X, \mu)$ and $(\Omega'_Y, \mathscr{A}'_Y, \nu)$ are σ-finite measure spaces. Then

$$P_{X,Y} \underset{\mathscr{A}'_X \otimes \mathscr{A}'_Y}{\ll} \mu \otimes \nu \;\Rightarrow\; P_X \underset{\mathscr{A}'_X}{\ll} \mu \text{ and } P_Y \underset{\mathscr{A}'_Y}{\ll} \nu.$$

(Proof p. 190)

Let

$$f_{X,Y} := \frac{dP_{X,Y}}{d\mu \otimes \nu}, \qquad f_X := \frac{dP_X}{d\mu}, \qquad f_Y := \frac{dP_Y}{d\nu}$$

denote Radon-Nikodym derivatives (see Th. 3.72 and Remark 3.74). In the following lemma, we use the notation '$\underset{\mu\text{-a.a.}}{=}$' introduced in Remark 2.70.

Lemma 5.75 (Marginal Densities)
Let $(X,Y)\colon (\Omega,\mathscr{A},P) \to (\Omega'_X \times \Omega'_Y, \mathscr{A}'_X \otimes \mathscr{A}'_Y)$ *be a random variable,* $(\Omega'_X,\mathscr{A}'_X,\mu)$, $(\Omega'_Y,\mathscr{A}'_Y,\nu)$ *be measure spaces,* μ,ν *be* σ*-finite, and assume* $P_{X,Y} \underset{\mathscr{A}'_X\otimes\mathscr{A}'_Y}{\ll} \mu\otimes\nu$. *Then*

$$f_X(x) \underset{\mu\text{-a.a.}}{=} \int f_{X,Y}(x,y)\ \nu(dy), \tag{5.46}$$

and

$$f_Y(y) \underset{\nu\text{-a.a.}}{=} \int f_{X,Y}(x,y)\ \mu(dx). \tag{5.47}$$

The functions f_X *and* f_Y *are also called* $marginal\ densities$ *of* X *and* Y, *respectively.*

(Proof p. 191)

Remark 5.76 (Marginal and Joint Density) Suppose that X and Y are real-valued. Then for the Lebesgue measure $\mu = \nu = \lambda$ and a Riemann integrable density $f_{X,Y}$, Equations (5.46) and (5.47) yield

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy \tag{5.48}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx, \tag{5.49}$$

respectively (see Th. 3.62). ◁

Note that, for a discrete random variable $X = (X_1,\ldots,X_n)$, the probability function p_X is a density of P_X with respect to the measure μ specified in Equation (5.41), and the marginal probability functions p_{X_i}, $i = 1,\ldots,n$, are marginal densities. In section 5.7.4 we consider multivariate densities with respect to the Lebesgue measure $\mu = \nu = \lambda$.

5.7 Uni- or Multivariate Real-Valued Random Variable

The remaining sections of this chapter show how to describe distributions of *real-valued* uni- and multivariate random variables.

5.7.1 Distribution Function of a Univariate Real-Valued Random Variable

If we consider a univariate real-valued random variable X, then the *distribution function* F_X assigns to each $x \in \mathbb{R}$ the probability $P(X \le x)$ of the event $\{X \le x\} = \{\omega \in \Omega\colon X(\omega) \le x\}$ that X takes on a value *smaller or equal* than x. As we shall see, the distribution function uniquely determines the distribution P_X.

Definition 5.77 (Distribution Function)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B}, P_X)$ denote a real-valued random variable. Then the (cumulative) distribution function $F_X\colon \mathbb{R} \to [0,1]$ of X is defined by:

$$F_X(x) \;:=\; P_X(]-\infty, x]) \;=\; P(X \le x), \quad \forall\, x \in \mathbb{R}. \tag{5.50}$$

Remark 5.78 (Probabilities of Intervals) This definition implies that we can compute the probability $P(a < X \le b)$ of X taking a value in the interval $]a, b]$ by

$$P(a < X \le b) \;=\; F_X(b) - F_X(a), \quad \text{if } a < b, \tag{5.51}$$

because

$$P(a < X \le b) \;=\; P_X\big(]-\infty, b] \setminus\,]-\infty, a]\big) \;=\; P_X(]-\infty, b]) \;-\; P_X(]-\infty, a])$$

[see Rule (vi), Box 4.1]. ◁

Remark 5.79 (The Distribution Function Determines the Distribution) Every random variable X has a distribution P_X. Therefore, the distribution function F_X exists for all real-valued random variables. The distribution function uniquely determines the distribution P_X of a real-valued random variable, because a finite measure on $(\Omega, \mathscr{A})$ is already uniquely specified by its values on a $\cap$-stable generating system of $\mathscr{A}$ (see Th. 1.71) and the set system $\{]-\infty, x]\colon x \in \mathbb{R}\}$ is a $\cap$-stable generating system of $\mathscr{B}$, the Borel σ-algebra on $\mathbb{R}$ [see Eq. (1.19)]. Hence, P_X uniquely determines F_X, which implies the following theorem. ◁

Theorem 5.80 (Uniqueness)
Let P_X, P_Y denote the distributions and F_X, F_Y the distribution functions of two real-valued random variables $X, Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$. Then

$$F_X = F_Y \quad \Leftrightarrow \quad P_X = P_Y. \tag{5.52}$$

This theorem facilitates describing distributions and calculations involving distributions considerably, because the distribution function is defined on the set $\mathbb{R}$ of real numbers, whereas P_X is defined on a much more complex domain, the Borel σ-algebra $\mathscr{B}$.

Example 5.81 (Flipping two Coins – continued) In Example 5.11, we considered flipping two coins and specified the distribution P_X of $X\colon (\Omega, \mathscr{A}, P) \to \left(\Omega'_X, \mathscr{P}(\Omega'_X)\right)$, representing with its values x the number of flipping heads. The distribution P_X assigns a probability to all $2^3 = 8$ subsets of $\Omega'_X := \{0,1,2\}$. Because $\{0,1,2\} \subset \mathbb{R}$, the random variable X is also a random variable $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ with values in $\mathbb{R}$. In this case, F_X is a step function and we only have to specify

$$F_X(x) \;=\; P_X\big(]-\infty, x]\big) \;=\; P(X \le x) \;=\; \begin{cases} 0, & \text{if } x < 0 \\ 1/4, & \text{if } 0 \le x < 1 \\ 3/4, & \text{if } 1 \le x < 2 \\ 1, & \text{if } x \ge 2\,. \end{cases}$$

According to Theorem 5.80 the distribution P_X is uniquely defined by these four values. In other words, if we know these four values, then we know the probabilities $P_X(A')$ for all elements A' of the Borel σ-algebra $\mathscr{B}$ (see Exercise 5-10). ◁

Now we turn to the *quantile function*, which, in some sense, is the inverse of the distribution function. Sometimes this function is also called the *pseudo-inverse* of F_X. It assigns to each $p \in [0,1]$ the smallest real number x for which $P(X \le x) = F_X(x) \ge p$.

Definition 5.82 (Quantile Function)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a real-valued random variable with distribution function F_X. Then the quantile function $Q_X\colon [0,1] \to \overline{\mathbb{R}}$ of X is defined by:

$$\begin{aligned} Q_X(p) \;&=\; \inf\{x \in \mathbb{R}\colon F_X(x) \ge p\}, \quad \forall\, p \in]0,1[\,, \\ Q_X(0) \;&=\; \begin{cases} \inf\{x \in \mathbb{R}\colon F_X(x) > p\}, & \text{if } \exists\, x \in \mathbb{R} \text{ with } F_X(x) = 0 \\ -\infty, & \text{if } F_X(x) > 0,\ \forall\, x \in \mathbb{R}, \end{cases} \\ Q_X(1) \;&=\; \begin{cases} \inf\{x \in \mathbb{R}\colon F_X(x) = 1\}, & \text{if } \exists\, x \in \mathbb{R} \text{ with } F_X(x) = 1 \\ \infty, & \text{if } F_X(x) < 1,\ \forall\, x \in \mathbb{R}. \end{cases} \end{aligned} \tag{5.53}$$

A value $Q_X(p)$, $p \in [0,1]$, is called the quantile of p with respect to F_X.

Remark 5.83 (Inverse Function of F_X) If F_X is continuous and strictly monotone, i. e., if $x_1 < x_2$ implies $F_X(x_1) < F_X(x_2)$, then

$$Q_X(p) \;=\; F_X^{-1}(p), \quad \forall\, p \in]0,1[\,. \tag{5.54}$$

where F_X^{-1} denotes the inverse function of F_X. ◁

Example 5.84 (Flipping two Coins – continued) In Example 5.81, we specified the distribution function of $X :=$ *number of flipping heads* for the random experiment of flipping two coins. The corresponding quantile function takes on the three values

$$Q_X(p) = \begin{cases} 0, & \text{if } 0 < p \le 1/4 \\ 1, & \text{if } 1/4 < p \le 3/4 \\ 2, & \text{if } 3/4 < p \le 1. \end{cases}$$

◁

5.7.2 Distribution Function of a Multivariate Real-Valued Random Variable

Now we extend the concept of a distribution function to the multivariate case. In the following definition we use the notation introduced in Equation (5.24).

Definition 5.85 (Joint Distribution Function)
Let $(X_1,\dots,X_n)\colon (\Omega,\mathscr{A},P) \to (\mathbb{R}^n,\mathscr{B}_n)$ be a random variable. Its distribution function $F_{X_1,\dots,X_n}\colon \mathbb{R}^n \to [0,1]$ is defined by

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) := P(X_1 \le x_1,\dots,X_n \le x_n), \quad \forall (x_1,\dots,x_n) \in \mathbb{R}^n. \tag{5.55}$$

It is also called the joint distribution function of $X_1,\dots,X_n$.

Example 5.86 (Flipping two Coins – continued) In example 5.31, we considered flipping two coins and defined the random variables X_1 and X_2 indicating whether or not we flip *heads* at first and second flip, respectively. In this example, the bivariate distribution function F_{X_1,X_2} takes on the values

$$F_{X_1,X_2}(x_1,x_2) = \begin{cases} 0 \text{ if } x_1 < 0 \text{ or } x_2 < 0, \\ 1/4 \text{ if } 0 \le x_1 < 1,\ 0 \le x_2 < 1, \\ 2/4 \text{ if } x_1 \ge 1,\ 0 \le x_2 < 1, \\ 2/4 \text{ if } 0 \le x_1 < 1,\ x_2 \ge 1, \\ 1 \text{ if } x_1 \ge 1,\ x_2 \ge 1. \end{cases}$$

◁

Just like in Theorem 5.80 we can prove uniqueness, using a $\cap$-stable generating system for $\mathscr{B}_n$, now referring to Equation (1.21).

Theorem 5.87 (Uniqueness)
Let P_X, P_Y denote the distributions and F_X, F_Y the distribution functions of two n-variate real-valued random variables $X,Y\colon (\Omega,\mathscr{A},P) \to (\mathbb{R}^n,\mathscr{B}_n)$. Then

$$F_X = F_Y \quad \Leftrightarrow \quad P_X = P_Y. \tag{5.56}$$

As a special case of Equation (5.20) with $A_i' =]-\infty, X_i]$ and $\Omega_j' = \mathbb{R}$, $i \neq j$, we obtain the next corollary. In the special case of a bivariate real-valued random variable (X_1, X_2), this corollary asserts that the value of the *marginal distribution function of* X_1 for the argument x_1, i. e., $\lim_{x_2 \to \infty} F_{X_1,X_2}(x_1, x_2)$, is identical to the value $F_{X_1}(x_1)$ of the distribution function of X_1 for the argument x_1. In this corollary,

$$\lim_{\substack{x_j \to \infty \\ j \neq i}} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \tag{5.57}$$

denotes the limit of the distribution function of $(X_1, \ldots, X_n)$ for $x_j \to \infty$, for all $j = 1, \ldots, i-1, i+1, \ldots, n$. This limit is the value of the *marginal distribution function of* X_i for the argument x_i, and the corollary asserts that this limit is identical to the value $F_{X_i}(x_i)$ of the distribution function of X_i for the argument X_i.

Corollary 5.88 (Joint and Marginal Distribution Function)
Let $(X_1, \ldots, X_n) \colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}^n, \mathscr{B}_n)$ *be a random variable. Then*

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to \infty \\ j \neq i}} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \quad \forall x_i \in \mathbb{R}. \tag{5.58}$$

(Proof p. 191)

The next corollary shows how independence of the random variables $X_1, \ldots, X_n$ can be formulated in terms of their distribution functions.

Corollary 5.89 (Independence and Joint Distribution Function)
Let $(X_1, \ldots, X_n) \colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}^n, \mathscr{B}_n)$ *be a random variable. Then the following two propositions are equivalent to each other:*

(a) $X_1, \ldots, X_n$ *are independent*
(b) $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n.$

(Proof p. 191)

5.7.3 Density of a Continuous Univariate Real-Valued Random Variable

As a special case, we consider a random variable $X \colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ for which there is a nonnegative measurable function $f_X : (\mathbb{R}, \mathscr{B}, P_X) \to (\mathbb{R}, \mathscr{B})$ such that

$$P_X(B) = \int_B f_X \, d\lambda, \quad \forall B \in \mathscr{B}, \tag{5.59}$$

where λ denotes the Lebesgue measure on $(\mathbb{R}, \mathscr{B})$ (see Def. 5.66). According to Theorem 3.68 (ii), this equation is equivalent to $P_X = f_X \odot \lambda$.

The following definition is a special case of Definition 5.66.

Definition 5.90 (Continuous Random Variable and its Density)
Let $X\colon (\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be a real-valued random variable with distribution P_X. We call X $continuous$, if there is a nonnegative function $f_X:\mathbb{R}\to\mathbb{R}$ that is integrable with respect to the Lebesgue measure λ and

$$P_X = f_X\odot\lambda\,. \tag{5.60}$$

A function f_X satisfying (5.60) *is called a $(probability)$ $density$ of X.*

Note that Equation (5.60) is equivalent to

$$F_X(x) = \int_{]-\infty,x]} f_X\,d\lambda,\quad \forall\, x\in\mathbb{R}\,, \tag{5.61}$$

because

$$F_X(x) = P_X\big(]-\infty,x]\big) = \int 1_{]-\infty,x]}\,dP_X = \int_{]-\infty,x]} f_X\,d\lambda,\quad \forall\, x\in\mathbb{R}\,, \tag{5.62}$$

[see Ths. 3.67 and 3.68 (iv)]. Furthermore, Theorem 3.62 immediately implies the following corollary:

Corollary 5.91 (Riemann Integral of the Density)
If f_X is a density of the random variable $X\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ and f_X is Riemann integrable, then

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt,\quad \forall\, x\in\mathbb{R}\,. \tag{5.63}$$

Remark 5.92 (Interpretation of Densities) Note that the term $f_X(t)$ in Equation (5.63) is not a probability, instead it is a value of the density for $t\in\mathbb{R}$. However, the probability $P(a<X\le b)$ that X takes on a value in the interval $]a,b]$ can be computed using Equation (5.51) and the density f_X, provided that it exists and that it is Riemann integrable:

$$P(a<X\le b) = F_X(b)-F_X(a) = \int_a^b f_X(x)\,dx,\quad \text{if } a<b. \tag{5.64}$$

This probability can be represented as the area between the density and the x-axis above the interval $[a,b]$ (see Fig. 3.6). ◁

Remark 5.93 (Continuity of X Implies $P(X=x)=0$) Consider a continuous random variable $X\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$. Definition 5.90 and Remark 3.71 imply that $P_X\underset{\mathscr{B}}{\ll}\lambda$. Because $\lambda(\{x\})=0$ [see Eq. (1.52)] we can conclude $P(X{=}x)=0$, for all $x\in\mathbb{R}$. Hence, additivity of P yields, for all $a,b\in\mathbb{R},\ a<b$,

$$P(a < X \leq b) \;=\; P(a \leq X \leq b) \;=\; P(a \leq X < b) \;=\; P(a < X < b), \tag{5.65}$$

provided that X is continuous. ◁

Example 5.94 (Continuous Random Variables and Their Densities) In section 8.2 we present some examples of continuous random variables and their densities, such as the densities of normal distributions, central χ^2-distributions, central t-distributions, and central F-distributions. ◁

5.7.4 Density of a Continuous Multivariate Real-Valued Random Variable

Remark 5.95 (Multivariate Case) Let $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}^n, \mathscr{B}_n)$ be a multivariate random variable with distribution P_X. If $f_X\colon \mathbb{R}^n \to \mathbb{R}$ is nonnegative and integrable with respect to the Lebesgue measure λ_n on $(\mathbb{R}^n, \mathscr{B}_n)$, and $P_X = f_X \odot \lambda_n$, then X is continuous with probability density f_X, and

$$F_X(x_1, \ldots, x_n) \;=\; \int_B f_X \, d\lambda_n, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{5.66}$$

where $B := \bigtimes_{i=1}^n]-\infty, X_i]$. If f_X is Riemann integrable, then

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f_X(t_1, \ldots, t_n) \, dt_1 \ldots dt_n, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{5.67}$$

More generally, for any $B \in \mathscr{B}_n$,

$$P_X(B) \;=\; P(X \in B) \;=\; \int 1_B \cdot f_X \, d\lambda_n. \tag{5.68}$$

The probability $P_X(B)$ can be represented as the $(n+1)$-dimensional *volume* between the joint density and the $(x_1, \ldots, x_n)$-hyperplane above B. ◁

In analogy to Corollary 5.89, independence of continuous real-valued random variables can also be formulated in terms of probability densities, using the marginal densities $f_{X_1}, \ldots, f_{X_n}$ (see Lemma 5.75).

Corollary 5.96 (Independence and Probability Densities)
Let $(X_1, \ldots, X_n)\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}^n, \mathscr{B}_n)$, $n > 1$, be a random variable and suppose that all random variables X_i, $i = 1, \ldots, n$, have a density f_{X_i} with respect to the Lebesgue measure λ on $(\mathbb{R}, \mathscr{B})$. Then $X_1, \ldots, X_n$ are independent if and only if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \;:=\; f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{5.69}$$

is a joint density of $(X_1, \ldots, X_n)$ with respect to λ_n.

(Proof p. 192)

Remark 5.97 (Independence, Densities, and Riemann Integrals) If all densities f_{X_i}, $i = 1, \ldots, n$, are Riemann integrable, then $f_{X_1,\ldots,X_n}$ in (5.69) is Riemann integrable as well (see, e. g., Ellis & Gulick, 2006). ◁

Remark 5.98 (Other Random Variables) Some special distribution functions are treated in more detail in chapter 8. There we shall see that the distribution functions of discrete random variables are step functions (see Fig. 8.1) and discrete random variables do not have probability densities with respect to the Lebesgue measure. The distribution functions of continuous random variables do not have discontinuities or jumps (see Fig. 8.7). Note, however, that there are random variables that are neither discrete nor continuous. Their distribution functions have jumps but are not step functions. ◁

5.8 Proofs

Proof of Lemma 5.7

For all $A \in \mathscr{A}$,

$$\begin{aligned}
P_{g(X)}(A) &= P\big([g(X)]^{-1}(A)\big) && \text{[(5.2)]}\\
&= P\big(\{\omega \in \Omega\colon g[X(\omega)] \in A\}\big) && \text{[Def. 2.1]}\\
&= P\big(\{\omega \in \Omega\colon X(\omega) \in g^{-1}(A)\}\big) && \text{[Def. 2.1]}\\
&= P\big(X^{-1}[g^{-1}(A)]\big) && \text{[Def. 2.1]}\\
&= P_X[g^{-1}(A)] && \text{[(5.2)]}\\
&= (P_X)_g(A). && \text{[(5.2)]}
\end{aligned}$$

Proof of Lemma 5.17

If $Q \underset{\mathscr{C}}{\ll} P$, then $P(A) = 0$ implies $Q(A) = 0$ for all $A \in \mathscr{C}$. This implication also holds for the event $A = \{X \neq Y\} := \{\omega \in \Omega\colon X(\omega) \neq Y(\omega)\}$. Therefore, $P(\{X \neq Y\}) = 0$ implies that $Q(\{X \neq Y\}) = 0$.

Proof of Corollary 5.21

The definition of the distribution P_X of X as an image measure (see Def. 5.3) yields

$$P_X\big(\{x \in \Omega'_X\colon g(x) \neq g^*(x)\}\big) = P\big(\{\omega \in \Omega\colon g[X(\omega)] \neq g^*[X(\omega)]\}\big)$$

Hence, $P_X\big(\{x \in \Omega'_X\colon g(x) \neq g^*(x)\}\big) = 0$ if and only if $P\big(\{\omega \in \Omega\colon g[X(\omega)] \neq g^*[X(\omega)]\}\big) = 0$. The same arguments hold if '$\underset{P}{=}$' is replaced by '$\underset{P}{<}$' or '$\underset{P}{\leqq}$' and '$\underset{P_X}{=}$' by '$\underset{P_X}{<}$' or '$\underset{P_X}{\leqq}$', respectively.

Proof of Lemma 5.23

Let $A := \big\{\omega \in \Omega\colon X(\omega) \neq Y(\omega)\big\}$. Note that A is $\mathscr{C}$-measurable. Hence, if $X \underset{P^B}{=} Y$, then the

conjunction of $P^B(A) = 0$ and $P \underset{\mathscr{C}}{\ll} P^B$ implies $P(A) = 0$.

Proof of Lemma 5.25

$P \underset{\sigma(X)}{\ll} P^B \Rightarrow P_X \underset{\mathscr{A}'_X}{\ll} P^B_X$: If $P \underset{\sigma(X)}{\ll} P^B$, then, for all $C' \in \mathscr{A}'_X$:

$$\begin{aligned} P^B_X(C') = 0 &\Rightarrow P^B[X^{-1}(C')] = 0 && [(5.2)] \\ &\Rightarrow P[X^{-1}(C')] = 0 && [P \underset{\sigma(X)}{\ll} P^B] \\ &\Rightarrow P_X(C') = 0. && [(5.2)] \end{aligned}$$

$P_X \underset{\mathscr{A}'_X}{\ll} P^B_X \Rightarrow P \underset{\sigma(X)}{\ll} P^B$: If $P_X \underset{\mathscr{A}'_X}{\ll} P^B_X$ and $C \in \sigma(X)$, then there is a $C' \in \mathscr{A}'_X$ such that

$$C = X^{-1}(C').$$

Hence,

$$\begin{aligned} P^B(C) = 0 &\Rightarrow P^B[X^{-1}(C')] = 0 \\ &\Rightarrow P^B_X(C') = 0 && [(5.2)] \\ &\Rightarrow P_X(C') = 0 && [P_X \underset{\mathscr{A}'_X}{\ll} P^B_X] \\ &\Rightarrow P[X^{-1}(C')] = 0 && [(5.2)] \\ &\Rightarrow P(C) = 0. \end{aligned}$$

Proof of Lemma 5.43

If $X_1 \underset{P}{=} X_2$, then

$$\begin{aligned} & X_1 \underset{P}{\perp\!\!\!\perp} Y \\ \Rightarrow\ & \sigma(X_1) \underset{P}{\perp\!\!\!\perp} \sigma(Y) && [\text{Def. 5.35}] \\ \Rightarrow\ & \forall (A', B') \in \mathscr{A}' \times \mathscr{A}'_Y: \\ & P[X_1^{-1}(A') \cap Y^{-1}(B')] = P[X_1^{-1}(A')] \cdot P[Y^{-1}(B')] && [\text{Defs. 4.44, 4.35, 4.32 (i)}] \\ \Rightarrow\ & \forall (A', B') \in \mathscr{A}' \times \mathscr{A}'_Y: \\ & P[X_2^{-1}(A') \cap Y^{-1}(B')] = P[X_2^{-1}(A')] \cdot P[Y^{-1}(B')] && [X_1 \underset{P}{=} X_2, \text{Def. 5.3, Cor. 5.20}] \\ \Rightarrow\ & \sigma(X_2) \underset{P}{\perp\!\!\!\perp} \sigma(Y) && [\text{Defs. 4.44, 4.35, 4.32 (i)}] \\ \Rightarrow\ & X_2 \underset{P}{\perp\!\!\!\perp} Y. && [\text{Def. 5.35}] \end{aligned}$$

Proof of Lemma 5.44

According to Remark 5.42, $1_B \underset{P}{\perp\!\!\!\perp} \mathscr{C} \Leftrightarrow B \underset{P}{\perp\!\!\!\perp} \mathscr{C}$. Hence, for all $C \in \mathscr{C}$,

$$\begin{aligned} P^B(C) = 0 \;&\Rightarrow\; P(B \cap C) = 0 \\ &\Rightarrow\; P(B) \cdot P(C) = 0 \qquad [B \underset{P}{\perp\!\!\!\perp} C] \\ &\Rightarrow\; P(C) = 0. \qquad [P(B) > 0] \end{aligned}$$

Proof of Lemma 5.45

$\Rightarrow$ If $X_1, \ldots, X_n$ are independent, then, for all $A_i' \in \mathscr{A}_i'$, $i = 1, \ldots, n$,

$$\begin{aligned} P_{X_1,\ldots,X_n}(A_1' \times \ldots \times A_n') &= P[X_1^{-1}(A_1') \cap \ldots \cap X_n^{-1}(A_n')] \\ &= P[X_1^{-1}(A_1')] \cdot \ldots \cdot P[X_n^{-1}(A_n')] \\ &= P_{X_1}(A_1') \cdot \ldots \cdot P_{X_n}(A_n') \\ &= (P_{X_1} \otimes \ldots \otimes P_{X_n})(A_1' \times \ldots \times A_n'). \end{aligned}$$

Hence, according to the definition of the product measure (see Def. 1.66),

$$P_{X_1,\ldots,X_n} = P_{X_1} \otimes \ldots \otimes P_{X_n}.$$

$\Leftarrow$ If the right-hand side of (5.29) holds, then, for all $A_i' \in \mathscr{A}_i'$, $i = 1, \ldots, n$,

$$\begin{aligned} P[X_1^{-1}(A_1') \cap \ldots \cap X_n^{-1}(A_n')] &= P_{X_1,\ldots,X_n}(A_1' \times \ldots \times A_n') \\ &= P_{X_1}(A_1') \cdot \ldots \cdot P_{X_n}(A_n') \\ &= P[X_1^{-1}(A_1')] \cdot \ldots \cdot P[X_n^{-1}(A_n')]. \end{aligned}$$

According to Definitions 5.35 and 4.35, this implies independence of $X_1, \ldots, X_n$.

Proof of Lemma 5.47

Assume that $X\colon (\Omega, \mathscr{A}, P) \to (\Omega_X', \mathscr{A}_X')$ is a random variable and that there is an $\alpha \in \Omega_X'$ such that $X \underset{P}{=} \alpha$. If $A := \{\omega \in \Omega\colon X(\omega) \neq \alpha\}$, then $P(A) = 0$ and $P(A^c) = 1$. This implies, for all $A' \in \mathscr{A}_X'$,

$$\begin{aligned} P[X^{-1}(A')] &= P[X^{-1}(A') \cap A^c] && \text{[Box 4.1 (viii)]} \\ &= P[\{\omega \in \Omega\colon X(\omega) \in A',\, X(\omega) = \alpha\}] \\ &= \begin{cases} P(\emptyset), & \text{if } \alpha \notin A' \\ P(A^c), & \text{if } \alpha \in A' \end{cases} \\ &= \begin{cases} 0, & \text{if } \alpha \notin A' \\ 1, & \text{if } \alpha \in A'. \end{cases} && (5.70) \end{aligned}$$

This implies, for all $A' \in \mathscr{A}_X'$ and all $C \in \mathscr{C}$,

$$\begin{aligned} P[X^{-1}(A') \cap C] &= \begin{cases} 0, & \text{if } \alpha \notin A', \\ P(C), & \text{if } \alpha \in A', \end{cases} && \text{[Box 4.1 (v), (viii)]} \\ &= P[X^{-1}(A')] \cdot P(C). && \text{[(5.70)]} \end{aligned}$$

Proof of Lemma 5.57

(i) ($\Rightarrow$) If X is discrete, then there is a finite or countable $\Omega_0' \subset \Omega_X'$ with $\{x\} \in \mathscr{A}_X'$ for all $x \in \Omega_0'$ and $P_X(\Omega_0') = 1$. This implies

$$P\left(1 \neq \sum_{x \in \Omega_0'} 1_{X=x}\right) = P(X \notin \Omega_0') = 0,$$

and therefore

$$1 \underset{P}{=} \sum_{x \in \Omega_0'} 1_{X=x}.$$

($\Leftarrow$) If Equation (5.32) holds and $\Omega_0' \subset \Omega_X'$ is finite or countable with $\{x\} \in \mathscr{A}_X'$ for all $x \in \Omega_0'$, then $1 \neq \sum_{x \in \Omega_0'} 1_{X=x}$ and therefore

$$1 = P\left(1 = \sum_{x \in \Omega_0'} 1_{X=x}\right) = P(X \in \Omega_0') = P_X(\Omega_0').$$

(ii) If X is real-valued, then X is discrete if and only if there is a finite or countable $\Omega' \subset \mathbb{R}$ such that

$$\begin{aligned}
X &= X \cdot 1 \\
&\underset{P}{=} X \cdot \sum_{x \in \Omega'} 1_{X=x} && \text{[(i)]} \\
&\underset{P}{=} \sum_{x \in \Omega'} X \cdot 1_{X=x} \\
&\underset{P}{=} \sum_{x \in \Omega'} x \cdot 1_{X=x}. && \text{[}1_{X=x}(\omega) = 0 \text{ if } X(\omega) \neq x, \text{ (5.6)]}
\end{aligned}$$

Proof of Corollary 5.59

$\Rightarrow$. If X is discrete and $\Omega_0' \subset \Omega_X'$ (see Def. 5.52), i. e., if Ω_0' is finite or countable and $P_X(\Omega_0') = 1$, then $\Omega_>' \subset \Omega_0'$, because $P_X(\Omega_X' \setminus \Omega_0') = 0$. Hence, $\Omega_>'$ is finite or countable as well, which proves (a). Furthermore, for finite or countable sets $\Omega_0', \Omega_>'$:

$$P_X(\Omega_0' \setminus \Omega_>') = \sum_{x \in \Omega_0' \setminus \Omega_>'} P(X{=}x) = 0. \qquad \text{[Def. of } \Omega_>'\text{]}$$

Hence,

$$P_X(\Omega_X' \setminus \Omega_>') = P_X(\Omega_X' \setminus \Omega_0') + P_X(\Omega_0' \setminus \Omega_>') = 0$$

and, according to (5.10),

$$1_{X \in (\Omega_X' \setminus \Omega_>')} \underset{P}{=} 0. \tag{5.71}$$

Therefore,

$$\begin{aligned} X &= 1_{X\in\Omega'_>}\cdot X + 1_{X\in(\Omega'_X\setminus\Omega'_>)}\cdot X && [1 = 1_{X\in\Omega'_>} + 1_{X\in(\Omega'_X\setminus\Omega'_>)}] \\ &= X\cdot \sum_{x\in\Omega'_>} 1_{X=x} && [(5.71),\ \Omega'_> \text{ is finite or countable}] \\ &= \sum_{x\in\Omega'_>} X\cdot 1_{X=x} \\ &= \sum_{x\in\Omega'_>} x\cdot 1_{X=x}. && [X\cdot 1_{X=x} = x\cdot 1_{X=x}] \end{aligned}$$

$\Leftarrow$. This is an immediate implication of Lemma 5.57 (ii).

Proof of Lemma 5.62

If $X_i \in \Omega'_i$ and $\Omega'_{0,x_i} = \emptyset$, then $p_{X_i}(X_i) = P(X_i{=}x_i) = 0$. If $\Omega'_{0,x_i} \neq \emptyset$, then it is finite or countable, because Ω'_0 is finite or countable. Then, using $P_X(\Omega'_0) = 1$,

$$\begin{aligned} & p_{X_i}(X_i) \\ &= P(X_i{=}x_i) \\ &= P(X_1\in\Omega'_1,\dots,X_{i-1}\in\Omega'_{i-1}, X_i{=}x_i, X_{i+1}\in\Omega'_{i+1},\dots,X_n\in\Omega'_n) && \text{[Box 4.1 (viii)]} \\ &= P\big((X_1,\dots,X_n)\in(\Omega'_1\times\dots\times\Omega'_{i-1}\times\{X_i\}\times\Omega'_{i+1}\times\dots\times\Omega'_n)\big) \\ &= P\Big((X_1,\dots,X_n)\in\big((\Omega'_1\times\dots\times\Omega'_{i-1}\times\{X_i\}\times\Omega'_{i+1}\times\dots\times\Omega'_n)\cap\Omega'_0\big)\Big) && \text{[Box 4.1 (viii)]} \\ &= \sum_{(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)\in\Omega'_{0,x_i}} P(X_1{=}x_1,\dots,X_{i-1}{=}x_{i-1}, X_i{=}x_i, X_{i+1}{=}x_{i+1},\dots,X_n{=}x_n) && \text{[Box 4.1 (x)]} \\ &= \sum_{(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)\in\Omega'_{0,x_i}} p_X(x_1,\dots,x_{i-1},X_i,x_{i+1},\dots,x_n). && [(5.35)] \end{aligned}$$

Proof of Lemma 5.64

If $X = (X_1,\dots,X_n)$ and $X_1,\dots,X_n$ are independent, then, for all $(x_1,\dots,x_n)\in\Omega'_0$,

$$\begin{aligned} p_X(x_1,\dots,x_n) &= P\big((X_1,\dots,X_n) = (x_1,\dots,x_n)\big) && [(5.30)] \\ &= P(X_1{=}x_1,\dots,X_n{=}x_n) \\ &= P(X_1{=}x_1)\cdot\ldots\cdot P(X_n{=}x_n) && [(5.27)] \\ &= p_{X_1}(x_1)\cdot\ldots\cdot p_{X_n}(x_n). && [(5.30)] \end{aligned}$$

For all $A'_1\in\mathscr{A}'_1,\dots,A'_n\in\mathscr{A}'_n$,

$$(A'_1\times\dots\times A'_n)\ \cap\ (\Omega'_{0,1}\times\dots\times\Omega'_{0,n})\ =\ (A'_1\cap\Omega'_{0,1})\times\dots\times(A'_n\cap\Omega'_{0,n}). \tag{5.72}$$

Now assume that Equation (5.37) holds and define $\Omega'_{sn} := \bigtimes_{i=1}^{n}\Omega'_{0,i}$. Then,

$$\begin{aligned} & P(X_1\in A'_1,\dots,X_n\in A'_n) \\ &= P[(X_1,\dots,X_n)\in(A'_1\times\dots\times A'_n)] \\ &= P[(X_1,\dots,X_n)\in(A'_1\times\dots\times A'_n)\cap\Omega'_{sn}] && [P_X(\Omega'_{sn}) = 1, \text{Box 4.1, (viii)}] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{(x_1,\dots,x_n)\in(A'_1\times\dots\times A'_n)\cap\Omega'_{sn}} P[(X_1,\dots,X_n)=(x_1,\dots,x_n)] && \text{[Box 4.1 (x)]}\\
&= \sum_{(x_1,\dots,x_n)\in(A'_1\times\dots\times A'_n)\cap\Omega'_{sn}} p_X(x_1,\dots,x_n) && \text{[(5.35)]}\\
&= \sum_{(x_1,\dots,x_n)\in(A'_1\times\dots\times A'_n)\cap\Omega'_{sn}} p_{X_1}(x_1)\cdot\ldots\cdot p_{X_n}(x_n) && \text{[(5.37)]}\\
&= \left(\sum_{x_1\in A'_1\cap\Omega'_{0,1}} p_{X_1}(x_1)\right)\cdot\ldots\cdot\left(\sum_{x_n\in A'_n\cap\Omega'_{0,n}} p_{X_n}(x_n)\right) && [P_{X_i}(\Omega'_{0,i})=1,\ (5.72)]\\
&= P(X_1\in A'_1)\cdot\ldots\cdot P(X_n\in A'_n), && \text{[Box 4.1 (x)]}
\end{aligned}
$$

which implies independence of $X_1,\dots,X_n$ [see (5.27)].

Proof of Theorem 5.73

For $A'\in\mathscr{A}'_X$,

$$
\begin{aligned}
P_X(A') &= \sum_{x\in A'\cap\Omega'_0} P_X(\{x\}) && \text{[Box 4.1, (x)]}\\
&= \sum_{x\in A'\cap\Omega'_0} p_X(x) && \text{[(5.30)]}\\
&= \sum_{x\in\Omega'_0} p_X(x)\cdot 1_{A'}(x)\\
&= \sum_{x\in\Omega'_0}\int p_X\cdot 1_{A'}\,d\delta_x && \text{[(3.23)]}\\
&= \int 1_{A'}\cdot p_X\,d\left(\sum_{x\in\Omega'_0}\delta_x\right) && \text{[(3.53)]}\\
&= \int_{A'} p_X\,d\left(\sum_{x\in\Omega'_0}\delta_x\right) && \text{[(3.30)]}\\
&= \int_{A'} p_X\,d\mu. && \text{[(5.41)]}
\end{aligned}
$$

Hence, according to Definition 5.66, the probability function p_X is the density of P_X with respect to the measure μ on the measurable space $(\Omega'_X,\mathscr{A}'_X)$ and $p_X=\frac{dP}{d\mu}$ (see Remark 3.73).

Proof of Lemma 5.74

Let $A'\in\mathscr{A}'_X$ with $\mu(A')=0$. Then

$$\mu\otimes\nu(A'\times\Omega'_Y) = \mu(A')\cdot\nu(\Omega'_Y) = 0\cdot\nu(\Omega'_Y) = 0,$$

and this holds even if $\nu(\Omega'_Y)=\infty$. This implies

$$P_X(A') = P(X\in A', Y\in\Omega'_Y) = P_{X,Y}(A'\times\Omega'_Y) = 0.$$

Therefore, $P_X\ll\mu$. The proof for P_Y is analogous.

Proof of Lemma 5.75

For all $A' \in \mathscr{A}'_X$,

$$\begin{aligned}
&\int 1_{A'}(x) \left(\int f_{X,Y}(x,y)\ \nu(dy) \right) \mu(dx) \\
&= \int 1_{A'}(x)\, f_{X,Y}(x,y)\ \mu \otimes \nu\,[d(x,y)] && \text{[Th. 3.76]} \\
&= \int 1_{A'}(x)\ P_{X,Y}[d(x,y)] && \text{[Th. 3.67]} \\
&= P_{X,Y}(A' \times \Omega'_Y) && \text{[(3.30), (3.8)]} \\
&= P_X(A'). && \text{[(5.20)]}
\end{aligned}$$

Theorem 3.65 then implies the lemma. The proof for f_Y is analogous.

Proof of Corollary 5.88

Note that

$$\mathbb{R}^{n-1} = \bigcup_{m=1}^{\infty}]-\infty, m]^{n-1}. \tag{5.73}$$

For all $X_i \in \mathbb{R}$,

$$\begin{aligned}
F_{X_i}(X_i) &= P(X_i \le X_i) \\
&= P_{X_1,\ldots,X_n}(\mathbb{R} \times \ldots \times \mathbb{R} \times]-\infty, X_i] \times \mathbb{R} \times \ldots \times \mathbb{R}) && \text{[(5.20)]} \\
&= P\big(X_1 \in \mathbb{R}, \ldots, X_{i-1} \in \mathbb{R}, X_i \in]-\infty, X_i], X_{i+1} \in \mathbb{R}, \ldots, X_n \in \mathbb{R}\big) && \text{[(5.2)]} \\
&= P\Big((X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \in \mathbb{R}^{n-1}, X_i \in]-\infty, X_i]\Big) \\
&= P\Big(X_j \in \bigcup_{m_j=1}^{\infty}]-\infty, m_j], j \in \{1, \ldots, i-1, i+1, \ldots, n\}, X_i \in]-\infty, X_i]\Big) && \text{[(5.73)]} \\
&= \lim_{m_j \to \infty,\, j \ne i} P\Big(X_j \in]-\infty, m_j], j \in \{1, \ldots, i-1, i+1, \ldots, n\}, X_i \in]-\infty, X_i]\Big) \\
& && \text{[Th. 1.67 (i)]} \\
&= \lim_{m_j \to \infty,\, j \ne i} F_{X_1,\ldots,X_n}(m_1, \ldots, m_{i-1}, X_i, m_{i+1}, \ldots, m_n) \\
&= \lim_{x_j \to \infty,\, j \ne i} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n).
\end{aligned}$$

The limits exist, because $F_{X_1,\ldots,X_n}$ is monotone in all coordinates.

Proof of Corollary 5.89

(a) $\Rightarrow$ (b) For all $(x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$\begin{aligned}
F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= P_{X_1,\ldots,X_n}\big((-\infty, x_1] \times \ldots \times (-\infty, x_n]\big) && \text{[(5.55)]} \\
&= P_{X_1} \otimes \ldots \otimes P_{X_n}\big((-\infty, x_1] \times \ldots \times (-\infty, x_n]\big) && \text{[Lemma 5.45 and (a)]} \\
&= P_{X_1}\big((-\infty, x_1]\big) \cdot \ldots \cdot P_{X_n}\big((-\infty, x_n]\big) && \text{[(1.50)]} \\
&= F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n). && \text{[Def. 5.77]}
\end{aligned}$$

(b) $\Rightarrow$ (a) For all $(x_1,\dots,x_n)\in\mathbb{R}^n$,

$$\begin{aligned}
&P_{X_1,\dots,X_n}\big((-\infty,x_1]\times\ldots\times(-\infty,x_n]\big)\\
&= F_{X_1,\dots,X_n}(x_1,\dots,x_n) && [(5.55)]\\
&= F_{X_1}(x_1)\cdot\ldots\cdot F_{X_n}(x_n) && [(b)]\\
&= P_{X_1}\big((-\infty,x_1]\big)\cdot\ldots\cdot P_{X_n}\big((-\infty,x_n]\big) && [\text{Def. 5.77}]\\
&= P_{X_1}\otimes\ldots\otimes P_{X_n}\big((-\infty,x_1]\times\ldots\times(-\infty,x_n]\big). && [(1.50)]
\end{aligned}$$

Because $\{(-\infty,x_1]\times\ldots\times(-\infty,x_n]\colon(x_1,\dots,x_n)\in\mathbb{R}^n\}$ is a $\cap$-stable generating system of $\mathscr{B}_n$ (see Example 1.73 and Def. 1.31), Theorem 1.71 yields $P_{X_1,\dots,X_n}=P_{X_1}\otimes\ldots\otimes P_{X_n}$. Applying Lemma 5.45 then completes the proof.

Proof of Corollary 5.96

(i) $(X_1,\dots,X_n)$ are independent $\Rightarrow$ $f_{X_1,\dots,X_n}$ in (5.69) is a joint density.
Because $f_{X_1,\dots,X_n}$ defined in (5.69) is nonnegative and integrable with respect to λ_n, we can conclude that $f_{X_1,\dots,X_n}\odot\lambda_n$ defines a finite measure on $(\mathbb{R}^n,\mathscr{B}_n)$ (see Th. 3.65). Furthermore, if $X_1,\dots,X_n$ are independent and f_{X_i} is a density of X_i for all $i=1,\dots,n$, then, for all $(x_1,\dots,x_n)\in\mathbb{R}^n$ and $B=\bigtimes_{i=1}^n\,]-\infty,X_i]$,

$$\begin{aligned}
&\int_B f_{X_1,\dots,X_n}(t_1,\dots,t_n)\;\lambda_n[d(t_1,\dots,t_n)]\\
&= \int_B f_{X_1}(t_1)\cdot\ldots\cdot f_{X_n}(t_n)\;\lambda_n[d(t_1,\dots,t_n)] && [(5.69)]\\
&= \prod_{i=1}^n\int_{(-\infty,X_i]} f_{X_i}(t_i)\;\lambda(dt_i) && [\text{Th. 3.76}]\\
&= \prod_{i=1}^n F_{X_i}(X_i) && [(5.61)]\\
&= F_{X_1,\dots,X_n}(x_1,\dots,x_n) && [\text{Cor. 5.89}]\\
&= P_{X_1,\dots,X_n}(B).
\end{aligned}$$

This shows that $f_{X_1,\dots,X_n}\odot\lambda_n(B)=P_{X_1,\dots,X_n}(B)$, and this implies that $f_{X_1,\dots,X_n}$ defined in (5.69) is a density of $(X_1,\dots,X_n)$ with respect to λ_n (see Def. 5.66).

(ii) $f_{X_1,\dots,X_n}$ in (5.69) is a density of $(X_1,\dots,X_n)$ $\Rightarrow$ $X_1,\dots,X_n$ are independent.
If Equation (5.69) holds and $B=\bigtimes_{i=1}^n\,]-\infty,X_i]$, then for all $(x_1,\dots,x_n)\in\mathbb{R}^n$

$$\begin{aligned}
F_{X_1,\dots,X_n}(x_1,\dots,x_n) &= \int_B f_{X_1,\dots,X_n}(t_1,\dots,t_n)\;\lambda_n[d(t_1,\dots,t_n)]\\
&= \int_{]-\infty,x_n]}\dots\int_{]-\infty,x_1]} f_{X_1}(t_1)\cdot\ldots\cdot f_{X_n}(t_n)\;\lambda(dt_1)\dots\lambda(dt_n)\\
&= \prod_{i=1}^n\int_{]-\infty,X_i]} f_{X_i}(t_i)\,\lambda(dt_i)\\
&= \prod_{i=1}^n F_{X_i}(X_i).
\end{aligned}$$

Now, Corollary 5.89 implies that $X_1,\dots,X_n$ are independent.

5.9 Exercises

▷ **Exercise 5-1** Consider the random variable X defined in Example 2.34. Which are the elements ω in the inverse image $X^{-1}(\{1\})$ and which are the probabilities of the events $\{\omega\}$?

▷ **Exercise 5-2** Consider again the random variable X defined in Example 2.34. What are the values of the distribution of X and the distribution of U?

▷ **Exercise 5-3** Consider Example 5.9. Show that $\mathbb{1}_{X \in A'}$ is a random variable on $(\Omega, \mathscr{A}, P)$ and that $\mathbb{1}_{X \in A'} = \mathbb{1}_{A'}(X) = \mathbb{1}_{A'} \circ X$.

▷ **Exercise 5-4** Show that $X \underset{P^B}{=} Y$ for $B = \{\omega_1, \omega_2, \omega_3\}$ and that $X \underset{P}{=} Y$ does *not* hold in Example 5.19.

▷ **Exercise 5-5** Show that the random variables U and X presented in Table 2.2 (p. 54) are independent.

▷ **Exercise 5-6** Let $(\Omega, \mathscr{A}, P)$ be a probability space, $B \in \mathscr{B}$, and $\mathscr{C} \subset \mathscr{A}$ a σ-algebra. Prove that the following propositions are equivalent to each other: $\mathbb{1}_B \underset{P}{\perp\!\!\!\perp} \mathscr{C}$, $\sigma(\{B\}) \underset{P}{\perp\!\!\!\perp} \mathscr{C}$, $\{B\} \underset{P}{\perp\!\!\!\perp} \mathscr{C}$, $B \underset{P}{\perp\!\!\!\perp} \mathscr{C}$.

▷ **Exercise 5-7** In Example 5.46 we showed that X and U are independent and in Example 5.50 we defined the mappings $f\colon \{0,1,2\} \to \{0,1\}$ and $g\colon \Omega_U \to \{male, female\}$. Use Equation (5.23) to show that the mappings $f(X)$ (control vs. any of the two treatments) and $g(U)$ (*sex*) are independent as well.

▷ **Exercise 5-8** Consider the random variable $(U, X)\colon \Omega \to \Omega_U \times \{0,1\}$ defined in Example 2.34 (p. 53). Which are the elements ω in the inverse image $(U,X)^{-1}(\{(Joe, 1)\})$ and which are the probabilities to the events $\{\omega\}$?

▷ **Exercise 5-9** Assume that $\Omega_0' \subset \bigtimes_{i=1}^n \Omega_i'$ is finite or countable. Construct a finite or countable set $\bigtimes_{i=1}^n \Omega_{0,i}'$ with

$$\Omega_0' \subset \bigtimes_{i=1}^n \Omega_{0,i}' \subset \bigtimes_{i=1}^n \Omega_i'.$$

▷ **Exercise 5-10** In Example 5.81 we specified the values of the distribution function F_X for the random variable *number of flipping heads*. Use these values to compute all eight values of the distribution P_X.

Solutions

▷ **Solution 5-1** The inverse image of the set $\{1\}$ under X is

$$X^{-1}(\{1\}) = \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -)\, (Ann, yes, +)\}.$$

The events $\{\omega\}$, $\omega \in X^{-1}(\{1\})$ have the probabilities $P[\{(Joe, yes, -)\}] = .04$, $P[\{(Joe, yes, +)\}] = .16$, $P[\{(Ann, yes, -)\}] = .16$, and $P[\{(Ann, yes, +)\}] = .04$.

▷ **Solution 5-2** The random variable X (the treatment variable) has the following distribution:

$$P_X(\{1\}) = P[X^{-1}(\{1\})] = .40, \qquad P_X(\{0\}) = P[X^{-1}(\{0\})] = .60,$$
$$P_X(\Omega') = P[X^{-1}(\Omega')] = 1, \qquad P_X(\emptyset) = P[X^{-1}(\emptyset)] = 0,$$

whereas the distribution of the random variable U (the observational-unit variable), is:

$$P_U(\{Joe\}) = P[U^{-1}(\{Joe\})] = .50 \qquad P_U(\{Ann\}) = P[U^{-1}(\{Ann\})] = .50$$
$$P_U(\Omega_U) = P[U^{-1}(\Omega_U)] = 1 \qquad P_U(\emptyset) = P[U^{-1}(\emptyset)] = 0.$$

▷ **Solution 5-3** Let $A' \in \mathscr{A}'_X$ and consider the indicator function $1_{X\in A'}\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$. Measurability: For all $B \in \mathscr{B}$,

$$\begin{aligned}
\left(1_{X\in A'}\right)^{-1}(B) &= \left(1_{X^{-1}(A')}\right)^{-1}(B) && \text{[(5.6)]}\\
&= \left\{\omega \in \Omega\colon 1_{X^{-1}(A')}(\omega) \in B\right\} && \text{[(2.2)]}\\
&= \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B,\\ X^{-1}(A'), & \text{if } 0 \notin B, 1 \in B\\ \Omega \setminus X^{-1}(A'), & \text{if } 0 \in B, 1 \notin B\\ \Omega, & \text{if } \{0,1\} \subset B, \end{cases}
\end{aligned}$$

and all these sets are elements of $\mathscr{A}$, because X is assumed to be a random variable.

Furthermore, for all $\omega \in \Omega$,

$$\begin{aligned}
1_{X\in A'}(\omega) &= 1_{X^{-1}(A')}(\omega) && \text{[(5.6)]}\\
&= \begin{cases} 1, & \text{if } \omega \in X^{-1}(A')\\ 0, & \text{if } \omega \notin X^{-1}(A') \end{cases}\\
&= \begin{cases} 1, & \text{if } X(\omega) \in A'\\ 0, & \text{if } X(\omega) \notin A' \end{cases} && \text{[(2.2)]}\\
&= 1_{A'}[X(\omega)]\\
&= 1_{A'} \circ X(\omega). && \text{[(2.25)]}
\end{aligned}$$

▷ **Solution 5-4** In order to prove $X \underset{P^B}{=} Y$ for $B = \{\omega_1, \omega_2, \omega_3\}$, we have to show $P^B(A) = 0$, where $A := \{\omega \in \Omega\colon X(\omega) \neq Y(\omega)\}$. In this example, $A = \{\omega_3, \omega_4\}$. Now $P(\{\omega_3\}) = 0$, which implies

$$P^B(\{\omega_3\}) = \frac{P(B \cap \{\omega_3\})}{P(B)} = \frac{P(\{\omega_3\})}{P(B)} = \frac{0}{P(B)} = 0.$$

Furthermore,

$$P^B(\{\omega_4\}) = \frac{P(B \cap \{\omega_4\})}{P(B)} = \frac{P(\emptyset)}{P(B)} = \frac{0}{P(B)} = 0.$$

Additivity of the measure P^B then implies $P^B(A) = 0$.

In order to prove that $X \underset{P}{=} Y$ *does not hold*, we have to show $P(A) \neq 0$. Now $P(\{\omega_3\}) = 0$ and $P(\{\omega_4\}) = .50$. Therefore, $P(A) \geq P(\{\omega_4\}) = .50 > 0$.

▷ **Solution 5-5** In Example 4.38, we already showed that the events $A := \{Joe\} \times \Omega_X \times \Omega_Y$ (that Joe is sampled) and $B := \Omega_U \times \{yes\} \times \Omega_Y$ (that the person sampled is treated) are independent. According to Box 4.2 (iii), this implies that the σ-algebras $\{A, A^c, \Omega, \emptyset\}$ and $\{B, B^c, \Omega, \emptyset\}$ are independent as well. If $X\colon (\Omega, \mathscr{A}, P) \to (\Omega', \mathscr{A}')$ and $U\colon (\Omega, \mathscr{A}, P) \to$

$(\Omega_U, \mathscr{A}_U)$ are the random variables defined Table 2.2 (p. 54), then $X^{-1}(\mathscr{A}') = \{A, A^c, \Omega, \emptyset\}$ and $U^{-1}(\mathscr{A}_U) = \{B, B^c, \Omega, \emptyset\}$ are the σ-algebras generated by U and X, respectively. Hence, in this example, X and U are independent.

▷ **Solution 5-6** First of all,

$$\{B\} \underset{P}{\perp\!\!\!\perp} \mathscr{C} \;\Leftrightarrow\; B \underset{P}{\perp\!\!\!\perp} \mathscr{C} \quad \text{[Rem. 4.36]}.$$

Furthermore, $\sigma(1_B) = \{\Omega, \emptyset, B, B^c\}$ (see Example 2.31) and $\sigma(\{B\}) = \{\Omega, \emptyset, B, B^c\}$. Hence, $\sigma(1_B) = \sigma(1_B)$, and this implies

$$1_B \underset{P}{\perp\!\!\!\perp} \mathscr{C} \;\Leftrightarrow\; \sigma(\{B\}) \underset{P}{\perp\!\!\!\perp} \mathscr{C} \quad \text{[Rem. 5.40]}.$$

Finally,

$$\sigma(\{B\}) \underset{P}{\perp\!\!\!\perp} \mathscr{C} \;\Leftrightarrow\; \{B\} \underset{P}{\perp\!\!\!\perp} \mathscr{C} \quad \text{[Box 4.2 (ii)]}.$$

▷ **Solution 5-7** Consider the two events $A := f(X)^{-1}(\{0\})$ and $B := g(U)^{-1}(\{male\})$.

$$P(A \cap B) \;=\; \frac{40}{99}$$

and

$$P(A) \cdot P(B) \;=\; \frac{60}{99} \cdot \frac{2}{3} \;=\; \frac{40}{99}.$$

Because $f(X)^{-1}\big(\mathscr{P}(\{0,1\})\big) = \{A, A^c, \Omega, \emptyset\}$ and $g(U)^{-1}\big(\mathscr{P}(\{male, female\})\big) = \{B, B^c, \Omega, \emptyset\}$, this proves that $f(X)$ and $g(U)$ are independent (see Exercise 5-6).

▷ **Solution 5-8** The inverse image of the set $\{1\}$ under X is

$$X^{-1}(\{1\}) \;=\; \{(Joe, yes, -), (Joe, yes, +), (Ann, yes, -)\,(Ann, yes, +)\}.$$

The events $\{\omega\}$, $\omega \in X^{-1}(\{1\})$ have the probabilities $P[\{(Joe, yes, -)\}] = .04$, $P[\{(Joe, yes, +)\}] = .16$, $P[\{(Ann, yes, -)\}] = .16$, and $P[\{(Ann, yes, +)\}] = .04$.

▷ **Solution 5-9** Define

$$\pi_i(\Omega_0') := \big\{X_i \in \Omega_i' \colon \exists (x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_n) \in \Omega_0'\big\}, \quad i = 1, \ldots, n,$$

the projection of Ω_0' onto Ω_i'. If Ω_0' is finite or countable, the all $\pi_i(\Omega_0')$, $i = 1, \ldots, n$, are finite or countable. Hence, $\bigtimes_{i=1}^n \pi_i(\Omega_0')$ is finite or countable and $\bigtimes_{i=1}^n \pi_i(\Omega_0') \subset \bigtimes_{i=1}^n \Omega_i'$. If $(x_1, \ldots, x_n) \in \Omega_0'$, then, by definition, $X_i \in \pi_i(\Omega_0')$, for all $i = 1, \ldots, n$, and therefore $(x_1, \ldots, x_n) \in \bigtimes_{i=1}^n \pi_i(\Omega_0')$. This implies $\Omega_0' \subset \bigtimes_{i=1}^n \pi_i(\Omega_0')$.

▷ **Solution 5-10** We consider all elements $B \in \mathscr{B}$ of the Borel σ-algebra on $\mathbb{R}$ and assign the following values:

$$P_X(B) \;=\; \begin{cases} F_X(\alpha) = 0,\ \alpha < 0, & \text{if } 0,1,2 \notin B \\ F_X(0) = \frac{1}{4}, & \text{if } 0 \in B,\ 1,2 \notin B \\ F_X(1) = \frac{3}{4}, & \text{if } 0,1 \in B,\ 2 \notin B \\ F_X(2) = 1, & \text{if } 0,1,2 \in B \\ F_X(1) - F_X(0) = \frac{2}{4}, & \text{if } 1 \in B,\ 0,2 \notin B \\ F_X(2) - F_X(0) = \frac{3}{4}, & \text{if } 1,2 \in B,\ 0 \notin B \\ F_X(2) - F_X(1) = \frac{1}{4}, & \text{if } 2 \in B,\ 0,1 \notin B \\ F_X(2) - F_X(1) + F_X(0) = \frac{1}{2}, & \text{if } 0,2 \in B,\ 1 \notin B. \end{cases}$$

Chapter 6
Expectation, Variance, and Other Moments

In chapter 4 we introduced a probability measure as a special finite measure and in chapter 5 we defined a random variable as a measurable mapping on a probability space. In this chapter we will translate integration theory (see ch. 3) to probability theory introducing *expectations* of numerical random variables and other important concepts that are special expectations: *central* and *noncentral moments*, and *variances*. Even *covariances* and *correlations* are special expectations (see ch. 7). All these concepts are important properties of random variables, although, in general, they do not determine the complete distribution.

6.1 Expectation

6.1.1 Definition

Reading the following definition, remember that a random variable is called *quasi-integrable* with respect to P if $\int Y^+ dP$ or $\int Y^- dP$ are finite, where Y^+ and Y^- denote the positive and negative parts of Y, respectively (see Rem. 2.62 and Def. 3.28).

Definition 6.1 (Expectation)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable that is quasi-integrable with respect to P. Then we define

$$E(Y) := \int Y \, dP, \tag{6.1}$$

call it the expectation of Y (w.r.t. P), and say that it exists. Instead of expectation w.r.t. P we also use the term P-expectation.

Remark 6.2 (Existence of the Expectation) Note that $E(Y)$ can be infinite. Furthermore, if $E(Y)$ *exists* we also say that Y is a random variable *with expectation* $E(Y)$. If Y is not quasi-integrable with respect to P and therefore also not P-integrable, then we say that the expectation of Y with respect to P does *not exist*. ◁

Remark 6.3 (Notation and Synonymous Terms) A synonym for expectation is *expectation value*. The reference to the measure P is usually omitted if the context is unambiguous. If we consider the expectation with respect to another probability measure on $(\Omega,\mathscr{A})$, e. g., the conditional-probability measure P^B (see Def. 4.24), then we adapt the notation correspondingly:

$$E^B(Y) := \int Y \, dP^B. \tag{6.2}$$

Expectation with respect to P^B is used synonymously to *P^B-expectation.* ◁

Remark 6.4 (Random Variables With Finite Expectations) Without substantial loss of generality, we will often assume that a random variable Y is *real-valued* if its expectation is finite. According to Remark 3.42, if the random variable $Y\colon (\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ has a finite expectation, then there is random variable $Y^*\colon (\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ such that $Y \underset{P}{=} Y^*$. ◁

6.1.2 Expectation of Discrete Random Variables

Remark 6.5 (Random Variable With a Finite Number of Real Values) Suppose $y_1,\dots,y_n\in\mathbb{R}$, $n\in\mathbb{N}$, denote all (negative, 0, or positive) values of a real-valued random variable $Y\colon (\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$, let P_Y denote the distribution, and p_Y the probability function of Y. Then the expectation $E(Y)$ exists, and

$$E(Y) = \sum_{i=1}^{n} y_i\cdot P(Y{=}y_i) = \sum_{i=1}^{n} y_i\cdot P_Y(\{y_i\}) = \sum_{i=1}^{n} y_i\cdot p_Y(y_i) \tag{6.3}$$

(see Cor. 3.59 and Def. 5.52). Hence, if Y has only a finite number n of values, then its expectation is simply the sum of its values, each one weighted by its probability $P(Y{=}y_i)=P_Y(\{y_i\})=p_Y(y_i)$. ◁

Example 6.6 (Expectation of an Indicator) If $(\Omega,\mathscr{A},P)$ is a probability space and 1_A is the indicator of $A\in\mathscr{A}$, then

$$E(1_A) = 0\cdot P(1_A{=}0) + 1\cdot P(1_A{=}1) = P(1_A{=}1) = P(A) \tag{6.4}$$

is the expectation of 1_A (see also Example 3.14). Considering the event $\{Y{=}y\}$ and using the notation $1_{Y=y} := 1_{\{Y=y\}}$, this yields

$$E(1_{Y=y}) = P(Y{=}y). \tag{6.5}$$

◁

Example 6.7 (Joe and Ann With Random Assignment – continued) In Example 1.9 we defined the set

$$B = \big\{(\mathit{Joe},\mathit{yes},-),(\mathit{Joe},\mathit{yes},+),(\mathit{Ann},\mathit{yes},-),(\mathit{Ann},\mathit{yes},+)\big\} = \Omega_U\times\{\mathit{yes}\}\times\Omega_Y,$$

the event that the drawn person is treated, and the set

$$C = \{(Joe, no, +), (Joe, yes, +), (Ann, no, +), (Ann, yes, +)\} = \Omega_U \times \Omega_X \times \{+\},$$

the event that $\{+\}$ (success) occurs, irrespective of which person is drawn and whether or not the person is treated. In Table 2.2 (p. 54) we assigned probabilities to each elementary event $\{\omega\}$, $\omega \in \Omega$ and defined $X := 1_B$, the treatment variable, as well as $Y := 1_C$ the outcome variable. Applying Equation (6.4) to the indicator 1_B yields:

$$\begin{aligned} E(X) &= E(1_B) = P(B) \\ &= P[\{(Joe, yes, -)\}] + P[\{(Joe, yes, +)\}] + P[\{(Ann, yes, -)\}] + P[\{(Ann, yes, +)\}] \\ &= .04 + .16 + .12 + .08 = .40. \end{aligned}$$

Similarly, for the indicator 1_C, we obtain

$$\begin{aligned} E(Y) &= E(1_C) = P(C) \\ &= P[\{(Joe, no, +)\}] + P[\{(Joe, yes, +)\}] + P[\{(Ann, no, +)\}] + P[\{(Ann, yes, +)\}] \\ &= .21 + .16 + .06 + .08 = .51. \end{aligned}$$

◁

Example 6.8 (Tossing a Dice – continued) In Example 3.6 we considered the random variable $X =$ *number of dots*. In this example we specified the probability space $(\Omega, \mathscr{A}, P)$ by $\Omega := \{\omega_1, \ldots, \omega_6\}$, $\mathscr{A} := \mathscr{P}(\Omega)$, and $P(\{\omega_1\}) = \ldots = P(\{\omega_6\}) = 1/6$. Because $P(X{=}x_i) = P_X(\{i\}) = P(\{\omega_i\})$, $i = 1, \ldots, 6$, Equation (6.3) yields

$$E(X) = \sum_{i=1}^{6} i \cdot P(X{=}i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = 3.5.$$

◁

Remark 6.9 (Random Variable With a Countable Number of Real Values) Let $y_1, y_2, \ldots \in \mathbb{R}$ denote the values of a real-valued random variable $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ and suppose that the expectation of Y exists. Then

$$E(Y) = \sum_{i=1}^{\infty} y_i \cdot P(Y{=}y_i) = \lim_{n\to\infty} \sum_{i=1}^{n} y_i \cdot P(Y{=}y_i), \tag{6.6}$$

using the notation introduced in Remark 5.4. Examples in which the expectation of a random variable is the 'infinite sum' of its values weighted by their probabilities are random variables with a Poisson distribution and with a geometric distribution (see Theorems 8.16 and 8.22). ◁

Example 6.10 (A Discrete Random Variable Without Expectation) Suppose that $\Omega = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ and consider the random variable $Y\colon \big(\Omega, \mathscr{P}(\Omega), P\big) \to (\mathbb{R}, \mathscr{B})$ defined by

$$Y(i) = y_i = (-1)^i i!\, e, \quad \forall\, i \in \mathbb{N}_0, \tag{6.7}$$

with

$$P(Y{=}y_i) \;=\; \frac{1}{e}\cdot\frac{1}{i!}, \quad \forall\, i \in \mathbb{N}_0\,. \tag{6.8}$$

Note that $e = \sum_{i=0}^{\infty} \frac{1}{i!}$. Hence, dividing both sides by e yields $\sum_{i=0}^{\infty} \frac{1}{i!\,e} = 1$. Therefore, Equation (6.8) specifies a probability distribution. Now consider

$$\sum_{i=0}^{n} y_i \cdot P(Y{=}y_i) \;=\; \sum_{i=0}^{n} (-1)^i \;=\; \begin{cases} 1, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd.} \end{cases} \tag{6.9}$$

Obviously, in this example, the limit

$$\lim_{n\to\infty} \sum_{i=0}^{n} y_i \cdot P(Y{=}y_i)$$

occurring in Equation (6.6) does not exist. Therefore, according to Definitions 6.1 and 3.28 $E(Y)$ is not defined. ◁

6.1.3 Computing Expectations Using Densities

According to the following theorem, the expectation of a continuous real-valued random variable can also be computed using the *Riemann integral*.

Theorem 6.11 (Expectation of a Continuous Random Variable)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a continuous random variable with expectation $E(Y)$ and a density f_Y that is Riemann integrable. Then

$$E(Y) \;=\; \int_{-\infty}^{\infty} y\; f_Y(y)\, dy. \tag{6.10}$$

(Proof p. 209)

Examples of continuous random variables and their expectations are treated in chapter 8, section 8.2.

Example 6.12 (A Continuous Random Variable Without Expectation) Consider the continuous random variable $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ with density

$$f_Y(y) = \frac{1}{\pi}\cdot\frac{1}{1+y^2}, \quad \forall\, y \in \mathbb{R}, \tag{6.11}$$

and distribution function

$$F_Y(y) \;=\; \frac{1}{2} + \frac{1}{\pi}\cdot\arctan y, \quad \forall\, y \in \mathbb{R}. \tag{6.12}$$

A distribution P_Y with density (6.11) is called a *standard Cauchy distribution* or *central t-distribution* with 1 degree of freedom.

The integral of the positive part of $y \cdot f_Y(y)$ is

$$\int_{-\infty}^{\infty} \big[y \cdot f_Y(y)\big]^+ dy = \int_0^{\infty} y \cdot \frac{1}{\pi} \cdot \frac{1}{1+y^2}\, dy = \frac{1}{2\pi} \ln(1+y^2)\Big|_0^{-\infty} = \infty,$$

and the integral of the negative part is

$$\int_{-\infty}^{\infty} \big[y \cdot f_Y(y)\big]^- dy = \int_{-\infty}^{0} -y \cdot \frac{1}{\pi} \cdot \frac{1}{1+y^2}\, dy = -\frac{1}{2\pi} \ln(1+y^2)\Big|_{-\infty}^{0} = \infty$$

[for the notation cf. Eq. (3.69)]. Hence, $y \cdot f_Y(y)$ is not quasi-P-integrable on $\mathbb{R}$ (see Def. 3.28) and $E(Y)$ does not exist. ◁

6.1.4 Transformation Theorem

The following corollary immediately follows from Theorem 3.57, translating the measure theory terms to probability theory. This corollary is relevant whenever we consider the expectation of a composition $g \circ Y = g(Y)$ of a numerical function g [see Eq. (2.25)] and a mapping Y or the expectation $E_Y(g)$ of g with respect to the distribution P_Y [see Eq. (5.2)].

Theorem 6.13 (Transformation Theorem)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Y, \mathscr{A}'_Y)$ be a random variable and $g\colon (\Omega'_Y, \mathscr{A}'_Y) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable.

(i) If g is nonnegative or has a finite expectation $E_Y(g)$, then

$$E_Y(g) = \int g\, dP_Y = \int g(y)\, P_Y(dy) = \int g(Y)\, dP = E[g(Y)]. \quad (6.13)$$

(ii) $E_Y(g)$ is finite if and only if $E[g(Y)]$ is finite.

Remark 6.14 (A Special Case) If we consider the special case in which g is the identity function $id\colon \overline{\mathbb{R}} \to \overline{\mathbb{R}}$, defined by $id(y) = y$, for all $y \in \overline{\mathbb{R}}$, then $id(Y) = Y$ and Equations (6.13) yield

$$E(Y) = \int Y\, dP = \int y\, P_Y(dy). \quad (6.14)$$

◁

Remark 6.15 (Finite Number of Values) If Y takes on only a finite number of different values $y_1, \ldots, y_n \in \mathbb{R}$, then Equation (6.13) simplifies to

$$E[g(Y)] = E_Y(g) = \int g\, dP_Y = \sum_{i=1}^{n} g(y_i) \cdot P(Y{=}y_i), \quad (6.15)$$

where $P(Y{=}y_i) = P_Y(\{y_i\}) = p_Y(y_i)$, $i = 1, \ldots, n$, and p_Y denotes the probability function of Y. ◁

Remark 6.16 (Countable Number of Values) If Y takes on a countable number of different values $y_1, y_2, \ldots \in \mathbb{R}$ and $\sum_{i=1}^{\infty} g^+(y_i) < \infty$ or $\sum_{i=1}^{\infty} g^-(y_i) < \infty$, then

$$E[g(Y)] \;=\; E_Y(g) \;=\int g \, dP_Y \;=\sum_{i=1}^{\infty} g(y_i) \cdot P(Y{=}y_i). \tag{6.16}$$

Note that (6.16) also applies if g is nonnegative, because in this case $g^- = 0$ holds for the negative part of g, which implies $\sum_{i=1}^{\infty} g^-(y_i) = 0 < \infty$. ◁

Equation (6.13) immediately implies the following corollary according to which the expectations of two random variables X and Y are identical if they have identical distributions, provided that the expectations exist (see also Remark 6.27).

Corollary 6.17 (Identical Distributions Imply Identical Expectations) *Let $X, Y : (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ be random variables and $g : (\Omega', \mathcal{A}') \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ a measurable function that is nonnegative or with expectation $E_X(g)$. If $P_X = P_Y$, then $E[g(X)] = E[g(Y)]$.*

This property allows us to use the term *expectation of a distribution* instead of expectation of a random variable.

In the following lemma we consider a bivariate random variable (X, Y) and a numerical function $g(X)$, i. e., a function that only depends on X. According to this lemma, the expectation of g with respect to the joint distribution $P_{X,Y}$ is identical to the expectation of g with respect to the marginal distribution P_X.

Lemma 6.18 (Expectation w.r.t. Joint and Marginal Distributions) *Let $(X, Y) : (\Omega, \mathcal{A}, P) \to (\Omega'_X \times \Omega'_Y, \mathcal{A}'_X \otimes \mathcal{A}'_Y)$ be a bivariate random variable with joint distribution $P_{X,Y}$ and let $g : (\Omega'_X, \mathcal{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ be a measurable function that is nonnegative or with expectation $E_X(g)$. Then*

$$E_{X,Y}(g) \;=\; E_X(g), \tag{6.17}$$

which is equivalent to

$$\int g(x) \, P_{X,Y}\,[d(x,y)] \;=\int g(x) \, P_X\,(dx). \tag{6.18}$$

(Proof p. 209)

Example 6.19 (Flipping two Coins – continued) Consider again the random variable $X =$ *number of flipping heads* and the indicator $1_H : \Omega \to \mathbb{R}$ of the event that *at least one heads is tossed*. In Example 2.47 we showed that $1_H = g \circ X$, where $g : \mathbb{R} \to \mathbb{R}$ is defined by

$$g(x) \;=\; \begin{cases} 1 & \text{if } x \in \{1,2\}, \\ 0, & \text{otherwise} \end{cases} \qquad \text{for all } x \in \mathbb{R}.$$

According to (6.15),

$$E(1_H) = E[g(X)] = E_X(g) = 0 \cdot P(\{X=0\}) + 1 \cdot P(X \in \{1,2\}) = P(X \in \{1,2\}) = \frac{3}{4}.$$

◁

Example 6.20 (Expectation of Y^2) Let $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a real-valued random variable, let $g\colon (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$ be measurable and let $g(Y) := Y^2$. Then, according to Theorem 6.13 and Equation (6.13),

$$E(Y^2) = E[g(Y)] = E_Y(g) = \int g \, dP_Y = \int y^2 \, P_Y(dy).$$

[Note that $E(Y^2)$ can be infinite.] This equation shows that the expectation of Y^2 solely depends on the distribution P_Y of Y. Using the integral $\int g \; dP_Y$ is often the most convenient way of computing the expectation $E(Y^2)$. If Y takes on only a finite number of values $y_1, \ldots, y_n \in \mathbb{R}$, then this equation simplifies to

$$E(Y^2) = \sum_{i=1}^{n} y_i^2 \cdot P_Y(\{y_i\}) = \sum_{i=1}^{n} y_i^2 \cdot P(Y=y_i). \tag{6.19}$$

These equations only involve the probabilities $P(Y=y_i) = P_Y(\{y_i\})$, not the probabilities $P(Y^2=y^2)$. ◁

Example 6.21 (Multiplication With Indicators) Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable with expectation $E(Y)$. If $A \in \mathscr{A}$ and $P(A) = 0$, then $(1_A \, Y) \underset{P}{=} 0$ and Rule (i) of Box 6.1 implies

$$E(1_A \, Y) = 0. \tag{6.20}$$

If $C = A \cup B$, $A \cap B = \emptyset$, and $A, B \in \mathscr{A}$, then $1_C \, Y = 1_A \, Y + 1_B \, Y$ and Rule (vi) of Box 6.1 implies

$$E(1_C \, Y) = E(1_A \, Y) + E(1_B \, Y). \tag{6.21}$$

◁

The following corollary shows how to compute the expectation of the composition $g(Y)$ using the density of Y. The virtue of Equation (6.22) is that we do not have to know the density of $g(Y)$, the density of Y suffices. (For a special choice of g see Remark 6.26.)

Corollary 6.22 (Transformation Theorem, Continuous Random Variable) *Let $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a continuous random variable with a Riemann integrable density f_Y. If $g\colon (\mathbb{R}, \mathscr{B}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a measurable function that is nonnegative or numerical with finite expectation $E_Y(g) = \int g \, dP_Y$, then*

$$E[g(Y)] = E_Y(g) = \int_{-\infty}^{\infty} g(y) \; f_Y(y) \, dy. \tag{6.22}$$

(Proof p. 209)

Box 6.1 Rules of Computation for Expectations

Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a random variable with expectation $E(Y)$ and let $\alpha \in \mathbb{R}$. Then:

$$Y \underset{P}{=} \alpha \quad \Rightarrow \quad E(Y) = \alpha. \tag{i}$$

$$E(\alpha + Y) = \alpha + E(Y). \tag{ii}$$

$$E(\alpha \cdot Y) = \alpha \cdot E(Y). \tag{iii}$$

Let $A, B \in \mathscr{A}$. Then

$$E(1_A \cdot 1_B) = P(A \cap B). \tag{iv}$$

$$E(1_A \cdot Y) = 0, \quad \text{if } P(A) = 0. \tag{v}$$

If Y_1, Y_2 are nonnegative or real-valued with finite expectation, then

$$E(Y_1 + Y_2) = E(Y_1) + E(Y_2). \tag{vi}$$

For $i = 1, \ldots, n$, let $Y_i\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be random variables with finite expectations $E(Y_i)$ and $\alpha_i \in \mathbb{R}$. Then

$$E\Big(\sum_{i=1}^{n} \alpha_i \cdot Y_i\Big) = \sum_{i=1}^{n} \alpha_i \cdot E(Y_i). \tag{vii}$$

Let $X, Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be random variables that are nonnegative or with finite expectations. Then:

$$X \underset{P}{=} Y \quad \Rightarrow \quad E(X) = E(Y). \tag{viii}$$

$$X \underset{P}{=} Y \quad \Leftrightarrow \quad \forall A \in \mathscr{A}\colon\ E(1_A X) = E(1_A Y). \tag{ix}$$

$$X \underset{P}{\perp\!\!\!\perp} Y \quad \Rightarrow \quad E(X \cdot Y) = E(X) \cdot E(Y). \tag{x}$$

6.1.5 Rules of Computation

Some rules of computation for expectations are gathered in Box 6.1 (for proofs see Exercise 6-2).

Example 6.23 (Expectation of a Sample Mean) Let $Y_1, \ldots, Y_n$ be a random sample. This means that the random variables $Y_1, \ldots, Y_n$ are i. i. d. (see Rem. 5.38). Further, define

$$\overline{Y} = \frac{1}{n} \cdot \sum_{i=1}^{n} Y_i, \tag{6.23}$$

the *arithmetic mean*, which in statistics is also called the *sample mean*. If Y_1 is nonnegative or with finite expectation and

$$\mu_Y := E(Y_1) \tag{6.24}$$

denotes the identical expectations of the variables $Y_1, \ldots, Y_n$, then

$$E(\overline{Y}) = \mu_Y \tag{6.25}$$

(see Exercise 6-4). ◁

Now we turn to a generalization of Rule (x) of Box 6.1.

Theorem 6.24 (Expectation of the Product of Random Variables)
Let $Y_i : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, $i = 1, \ldots, n$, be real-valued random variables that are nonnegative or with finite expectations and assume that the $Y_1, \ldots, Y_n$ are independent. Then

$$E\Big(\prod_{i=1}^{n} Y_i\Big) = \prod_{i=1}^{n} E(Y_i). \tag{6.26}$$

If we additionally assume that the expectations $E(Y_i)$, $i = 1, \ldots, n$, are finite, then $E\left(\prod_{i=1}^{n} Y_i\right)$ is finite, too.

For a proof see, e. g., Bauer (1996, Theorem 8.1, p. 49). Later we will weaken the independence assumption [see Box 7.1 (i)]. However, if the variables Y_i are *not independent*, then Equation (6.26) does *not necessarily hold* (see Remark 7.9).

6.2 Moments, Variance, and Standard Deviation

The expectation $E(Y)$ of a numerical random variable Y is also called the *first moment of Y*, provided that this expectation exists, whereas the expectation $E(Y^2)$ is called the *second moment of Y* (see Example 6.20). For second (and higher) moments we distinguish between *moments* and *central moments*.

Definition 6.25 (Moments)
Let $Y : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable, let $n \in \mathbb{N}$, and let Y be nonnegative or such that $E(Y^n)$ is finite. Then $E(Y^n)$ is called the nth moment and $E(|Y|^n)$ the nth absolute moment. Furthermore, if $n \geq 2$, and the expectation $E(Y)$ is finite, then we call $E\big([Y - E(Y)]^n\big)$ the nth central moment of Y.

Remark 6.26 (Higher Moments of Y) Analogously to Example 6.20, we may consider $g(Y) = Y^n$ with $n \geq 1$. If n is even, then the random variable Y^n is nonnegative and $E(Y^n)$ exists. In contrast, if n is odd, then the expectation $E(Y^n)$ does not necessarily exist. Provided that it exists, the expectation $E(Y^n)$ is called the *nth moment of* Y, $n \in \mathbb{N}$. If, for $n \in \mathbb{N}$, the expectation $E(Y^n)$ exists and is finite, then $E(Y^m)$ exists and is finite as well for all m with $1 \leq m \leq n$ (see Exercise 6-3). $\triangleleft$

Remark 6.27 (Moments Under P-Equivalence) If the expectations of Y^n and of $[Y - E(Y)]^n$ are finite, then they can be represented as expectations of functions $g(Y)$ of Y, where $g\colon (\overline{\mathbb{R}}, \overline{\mathscr{B}}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a measurable function with finite expectation $E_Y(g)$ [see Proposition (ii) of Theorem 6.13]. Therefore, according to Corollary 6.17, all moments (central or noncentral) of a numerical random variable Y solely depend on its distribution P_Y. Hence, if two random variables Y_1 and Y_2 have the same distribution $P_{Y_1} = P_{Y_2}$, then they have the same moments. For instance, if $Y_1 \underset{P}{=} Y_2$ and the expectations $E({Y_1}^2)$ and $E({Y_2}^2)$ are finite, then $E(Y_1) = E(Y_2)$ and $E({Y_1}^2) = E({Y_2}^2)$. This allows us to use the terms (central) moments of a distribution instead of (central) moments of a random variable. $\triangleleft$

Variance and standard deviation are the most important parameters describing the *variability* of a random variable. They are defined as follows:

Definition 6.28 (Variance and Standard Deviation)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable and assume that $E(Y^2) < \infty$. Then the variance of Y is defined by

$$Var(Y) := E\big([Y - E(Y)]^2\big), \tag{6.27}$$

and the standard deviation of Y by the positive square root of the variance, i. e.,

$$SD(Y) := \sqrt{Var(Y)}. \tag{6.28}$$

According to this definition, $Var(Y)$ is the expectation of the squared *mean centered* random variable $Y - E(Y)$. Hence, the variance of Y is the second central moment of Y. Note that variances and standard deviations are nonnegative. The variance of Y is also denoted by σ_Y^2 and the standard deviation by σ_Y. Box 6.2 summarizes some important properties of variances (see Exercise 7-5).

Example 6.29 (Variance of an Indicator) Let $(\Omega, \mathscr{A}, P)$ be a probability space and let 1_A denote the indicator of $A \in \mathscr{A}$. Then

$$\begin{aligned} Var(1_A) &= E({1_A}^2) - [E(1_A)]^2 && \text{[Box 6.2, (i)]} \\ &= E(1_A) - [E(1_A)]^2 && [{1_A}^2 = 1_A] \\ &= E(1_A) \cdot [1 - E(1_A)] \\ &= P(A) \cdot [1 - P(A)]. && \text{[(6.4)]} \end{aligned} \tag{6.29}$$

Box 6.2 Rules of Computation for Variances

Let $X, Y : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be random variables with finite second moments and let $\alpha \in \mathbb{R}$. Then:

$$Var(Y) = E(Y^2) - E(Y)^2. \tag{i}$$

$$Var(\alpha + Y) = Var(Y). \tag{ii}$$

$$Var(\alpha \cdot Y) = \alpha^2 \cdot Var(Y). \tag{iii}$$

$$\exists\, \alpha \in \mathbb{R} : Y \underset{P}{=} \alpha \quad \Leftrightarrow \quad Var(Y) = 0. \tag{iv}$$

$$X \underset{P}{=} Y \quad \Rightarrow \quad Var(X) = Var(Y). \tag{v}$$

For $i = 1, \ldots, n$, let the random variables $Y_i : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be independent with finite second moments and $\alpha_i \in \mathbb{R}$. Then

$$Var\left(\sum_{i=1}^{n} \alpha_i \cdot Y_i\right) = \sum_{i=1}^{n} \alpha_i^2 \cdot Var(Y_i). \tag{vi}$$

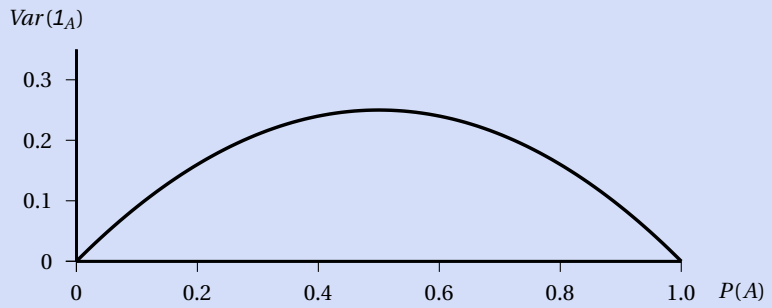


Figure 6.1. Variance of an indicator of an event as a function of its probability

According to Equation (6.4), the expectation of the indicator 1_A is $P(A)$, and Equation (6.29) shows that its variance is $P(A) \cdot [1 - P(A)]$. Obviously, the variance of an indicator variable does not contain any information additional to the expectation $E(1_A) = P(A)$. In fact, in this case, $E(1_A)$ contains the full information about the distribution of 1_A. This is not surprising because the distribution of 1_A is completely determined by the single parameter $P(A)$. Unlike the expectation, the variance of 1_A *does not* comprise the full information about the distribution of 1_A. For example, $Var(1_A) = .1 \cdot .9 = .09$ if $P(A) = .1$ or $P(A) = .9$. The variance of 1_A has its maximum for $P(A) = 1/2$ and goes to 0 if $P(A)$ approaches 0 or 1 (see Fig. 6.1). $\lhd$

Example 6.30 (Location vs. Variability) Consider two real-valued random variables $X, Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ with $P(X{=}-1) = P(X{=}1) = 0.5$ and $P(Y{=}-10) = P(Y{=}10) = 0.5$. Then $E(X) = E(Y) = 0$ but $Var(X) = 1 \neq Var(Y) = 100$. This illustrates that the expectation describes the 'location' of a random variable [see Rule (ii) of Box 6.1], while the variance is invariant with respect to translations [see Rule (ii) of Box 6.2]. In contrast, the variance describes the 'variability' of a random variable, whereas, in general, the expectation does not. ◁

Example 6.31 (Joe and Ann – continued) In the example presented in Table 2.2 (p. 54), X is an indicator variable. Hence, its variance is most easily computed as follows:

$$Var(X) \;=\; P(X{=}1) \cdot [1 - P(X{=}1)] \;=\; .40 \cdot .60 \;=\; .24.$$

Similarly, the variance of Y is obtained by

$$Var(Y) \;=\; P(Y{=}1) \cdot [1 - P(Y{=}1)] \;=\; .51 \cdot (1 - .51) \;=\; 0.2499.$$

◁

Example 6.32 (Variance and Standard Error of the Sample Mean) Let $Y_1, \ldots, Y_n$ be a sample (see Rem. 5.38), and $\overline{Y}$ the sample mean [see Eq. (6.23)]. If $E({Y_1}^2) < \infty$, and

$$\sigma_Y^2 \;:=\; Var(Y_1) \tag{6.30}$$

denotes the identical variances of the $Y_1, \ldots, Y_n$, then

$$Var(\overline{Y}) \;=\; \frac{\sigma_Y^2}{n} \tag{6.31}$$

(see Exercise 6-4). Hence,

$$SD(\overline{Y}) \;=\; \frac{\sigma_Y}{\sqrt{n}}. \tag{6.32}$$

In statistics $SD(\overline{Y})$ is also denoted $SE(\overline{Y})$ or $\sigma_{\overline{Y}} := \sqrt{\sigma_{\overline{Y}}^2}$ and called the *standard error of the sample mean.* ◁

Remark 6.33 (Z-Transformation) Note that every real-valued random variable $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ with finite and positive variance $Var(Y)$ can be transformed by

$$Z_Y \;:=\; \frac{Y - E(Y)}{SD(Y)}. \tag{6.33}$$

Note that $E(Z_Y) = 0$ and $Var(Z_Y) = 1$ (see Exercise 6-5). This transformation is called the *Z-transformation.*

As a special case consider the mean $\overline{Y}$ of a sample $Y_1, \ldots, Y_n$ with expectation $\mu_Y := E(Y_1)$ and standard deviation $\sigma_Y := SD(Y_1)$ (see Example 6.32). Then the Z-transformation of $\overline{Y}$ is

$$Z_{\overline{Y}} = \sqrt{n} \cdot \frac{\overline{Y} - \mu_Y}{\sigma_Y}. \tag{6.34}$$

The random variable $Z_{\overline{Y}}$ will be used in the Central Limit Theorem (see Th. 8.34). ◁

Definition 6.34 (Skewness)
Let $Y: (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable. If $E(Y^3)$ is finite, $Var(Y) > 0$, and $Z_Y := [Y - E(Y)]/SD(Y)$, then $E(Z_Y^3)$ is called the skewness of Y.

6.3 Proofs

Proof of Theorem 6.11

$$\begin{aligned}
E(Y) &= \int Y \, dP && [(6.1)] \\
&= \int y \, P_Y(dy) && [(3.59)] \\
&= \int y \, f_Y(y) \, \lambda(dy) && [(3.72), (5.60)] \\
&= \int_{-\infty}^{\infty} y \, f_Y(y) \, dy. && [\text{Riemann integrability of } f_Y, (3.68)]
\end{aligned}$$

Proof of Lemma 6.18

For all $A' \in \mathscr{A}'_X$,

$$\begin{aligned}
\int 1_{A'}(x) \, P_X(dx) &= P_X(A') && [(3.9)] \\
&= P_{X,Y}(A' \times \Omega'_Y) && [(5.20)] \\
&= \int 1_{A'}(x) \cdot 1_{\Omega'_Y}(y) \, P_{X,Y}[d(x,y)] && [(3.9), (1.37)] \\
&= \int 1_{A'}(x) \, P_{X,Y}[d(x,y)]. && [1_{\Omega'_Y}(y) = 1]
\end{aligned}$$

Now the proposition follows applying the standard methods of proofs described in Remark 3.30.

Proof of Corollary 6.22

$$\begin{aligned} E[g(Y)] &= E_Y(g) = \int g(y)\, P_Y(dy) && [(6.13)] \\ &= \int g(y)\, f_Y(y)\, \lambda(dy) && [(3.72), (5.60)] \\ &= \int_{-\infty}^{\infty} y\, f_Y(y)\, dy. && [\text{Riemann integrability of } f_Y, (3.68)] \end{aligned}$$

6.4 Exercises

▷ **Exercise 6-1** Assume $E(Y^2) < \infty$ and show that $\alpha \in \mathbb{R}$ minimizes $MSE(a) := E[(Y-a)^2]$ if and only if $\alpha = E(Y)$.

▷ **Exercise 6-2** Prove the rules of computation of Box 6.1.

▷ **Exercise 6-3** Show: If, for $n \in \mathbb{N}$, the expectation $E(Y^n)$ exists and is finite, then $E(Y^m)$ exists and is finite as well for all $1 \le m \le n$.

▷ **Exercise 6-4** Prove Equations (6.25) and (6.31).

▷ **Exercise 6-5** Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable with finite second moment $E(Y^2)$ and $Var(Y) > 0$. Show that the expectation of $Z := [Y - E(Y)]/SD(Y)$ is 0 and its variance is 1.

Alternativ Nagel

▷ **Exercise 6-6** Let $X_i\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, $i = 1, 2, \ldots, n$, be a finite sequence of independent identically distributed random variables with finite expectations $E(X_i) = \mu$ and finite and positive variances $Var(X_i) = \sigma^2$, $i = 1, \ldots, n$. Furthermore, let $\overline{X} := S/n$ be the arithmetic mean, where $S := \sum_{i=1}^{n} X_i$, and

$$\overline{Z} := \frac{(\overline{X} - \mu) \cdot \sqrt{n}}{\sigma}.$$

Show that $E(\overline{Z}) = 0$ and $Var(\overline{Z}) = 1$.

Solutions

▷ **Solution 6-1** For all $\alpha \in \mathbb{R}$,

$$\begin{aligned} E[(Y-\alpha)^2] &= E\big([Y - E(Y) + E(Y) - \alpha]^2\big) \\ &= E\big([Y - E(Y)]^2\big) + [E(Y) - \alpha]^2 + 2 \cdot E\big([Y - E(Y)] \cdot [E(Y) - \alpha]\big) \\ &= E\big([Y - E(Y)]^2\big) + [E(Y) - \alpha]^2 && [\text{Box 6.1 (iii), (ii)}] \\ &\ge E\big([Y - E(Y)]^2\big), \end{aligned}$$

and '=' holds if and only if $E(Y) = \alpha$.

▷ **Solution 6-2** Because the expectation of a numerical random variable is defined as an integral, we simply can refer to the corresponding propositions of chapter 3.

(i) This is Equation (3.8) with $A = \Omega$ and $\mu(A) = P(\Omega)$.

(ii) This is a special case of Equation (3.34)

(iii) This is a special case of Equation (3.32).

(iv) Note that $1_A \cdot 1_B = 1_{A\cap B}$. Hence, this rule follows from Equation (3.8), with $\alpha = 1$.

(vi) If Y_1, Y_2 are nonnegative, then this equation is a special case of Equation (3.65). If Y_1 or Y_2 has a finite expectation, then this equation is a special case of Equation (3.34).

(v) This is a special case of Lemma 3.45.

(vi) This rule follows from Equation (3.34) and complete induction.

(viii) This is Lemma 3.47.

(ix) This is Theorem 3.48.

(x) This is a special case of Theorem 6.24.

▷ **Solution 6-3** Assume that, for $n \in \mathbb{N}$, the expectation $E(Y^n)$ exists and is finite. Furthermore, let $A := \{\omega \in \Omega\colon |Y(\omega)| > 1\}$. Now, for all m with $1 \le m \le n$,

$$|Y(\omega)^m| \le |Y(\omega)^n|, \quad \forall\, \omega \in A \qquad \text{and} \qquad |Y(\omega)^m| \le 1, \quad \forall\, \omega \in A^c.$$

Therefore, applying Corollary 3.13, for $1 \le m \le n$:

$$\begin{aligned}
E(|Y^m|) &= \int |Y^m|\, dP \\
&= \int_A |Y^m|\, dP + \int_{A^c} |Y^m|\, dP && \text{[(3.36)]} \\
&\le \int_A |Y^n|\, dP + \int_{A^c} 1\, dP && \text{[Lemma 3.26]} \\
&\le \int_A |Y^n|\, dP + 1 && \text{[(3.30), } 1_{A^c} \le 1\text{, Lemma 3.26]} \\
&< \infty.
\end{aligned}$$

▷ **Solution 6-4**

Equation (6.25) can be derived as follows:

$$\begin{aligned}
E(\overline{Y}) &= E\left(\frac{1}{n} \cdot \sum_{i=1}^{n} Y_i\right) && \text{[(6.23)]} \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} E(Y_i) && \text{[Box 6.1 (vi)]} \\
&= \frac{1}{n} \cdot n \cdot \mu_Y = \mu_Y. && [Y_1, \ldots, Y_n \text{ are identically distributed, (6.24)}]
\end{aligned}$$

Equation (6.31) can be derived as follows:

$$\begin{aligned}
Var(\overline{Y}) &= Var\left(\frac{1}{n} \cdot \sum_{i=1}^{n} Y_i\right) && \text{[(6.23)]} \\
&= \frac{1}{n^2} \cdot \sum_{i=1}^{n} Var(Y_i) && [\perp\!\!\!\perp_P Y_1, \ldots, Y_n, \text{ Box 6.2 (vi)}] \\
&= \frac{1}{n^2} \cdot n \cdot \sigma_Y^2 = \frac{\sigma_Y^2}{n}. && [Y_1, \ldots, Y_n \text{ are identically distributed, (6.30)}]
\end{aligned}$$

▷ **Solution 6-5** Let $\mu := E(Y)$ and $\sigma := SD(Y)$. Then

$$\begin{aligned} E(Z) &= E\left(\frac{Y-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma}\cdot(Y-\mu)\right) = \frac{1}{\sigma}\cdot E(Y-\mu) && \text{[Box 6.1 (iii)]} \\ &= \frac{1}{\sigma}\cdot[E(Y)-E(\mu)] = \frac{1}{\sigma}\cdot(\mu-\mu) = 0 && \text{[Box 6.1 (vi), (i)].} \end{aligned}$$

$$\begin{aligned} Var(Z) &= Var\left(\frac{Y-\mu}{\sigma}\right) = Var\left(\frac{1}{\sigma}\cdot(Y-\mu)\right) \\ &= \frac{1}{\sigma^2}\cdot Var(Y) = \frac{1}{\sigma^2}\cdot\sigma^2 = 1 && \text{[Box 6.2 (iii), (ii)].} \end{aligned}$$

▷ **Solution 6-6**

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{S}{n}\right) = \frac{1}{n}\cdot E(S) && \text{[def. of } \overline{X}\text{, Box 6.1 (iii)]} \\ &= \frac{1}{n}\cdot E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\cdot\sum_{i=1}^{n} E(X_i) && \text{[def. of } S\text{, Box 6.1 (vi)]} \\ &= \frac{1}{n}\cdot\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\cdot\mu = \mu. \end{aligned}$$

$$\begin{aligned} Var(\overline{X}) &= Var\left(\frac{S}{n}\right) = \frac{1}{n^2}\cdot Var(S) = \frac{1}{n^2}\cdot Var\left(\sum_{i=1}^{n} X_i\right) && \text{[Box 6.2 (iii), def. of } S\text{]} \\ &= \frac{1}{n^2}\cdot\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\cdot\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\cdot\sigma^2 && \text{[Box 7.1 (viii), def. of } S\text{]} \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

In the second line we also used the assumption that the X_i are independent, which implies $Cov(X_i, X_j) = 0$, if $i \neq j$ (see Remark 7.9 and Th. 6.24). Furthermore, we used the assumption that the X_i are identically distributed, which implies $Var(X_i) = \sigma^2,\ i = 1,\ldots,n$.

Using these results and Exercise 6-5 yields

$$E(\overline{Z}) = E\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) = 0.$$

and

$$Var(\overline{Z}) = Var\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) = 1.$$

Chapter 7
Linear Quasi-Regression, Covariance, and Correlation

Expectation and variance are parameters that describe important properties of a univariate numerical random variable and its distribution. Now we consider *two* numerical random variables, say X and Y, and their joint distribution. In other words, we consider the distribution of the bivariate real-valued random variable $(X,Y)\colon(\Omega,\mathscr{A},P)\to(\mathbb{R}^2,\mathscr{B}_2)$. We also introduce a new random variable that can be used to describe a specific kind of dependence of Y on X. It is the kind of dependence of Y on X that is represented by the best fitting linear function $\alpha_0+\alpha_1 X$, 'best fitting' in terms of the minimal mean squared error. This function is the composition of X and the *linear quasi-regression* or the *linear least-squares regression. Covariance* and *correlation* are important parameters quantifying the strength of the kind of dependence that can be described by a linear quasi-regression.

7.1 Linear Quasi-Regression

Remark 7.1 (Implications of Finite Second Moments) Reading the following definition note that $E(X^2),E(Y^2)<\infty$ implies that $E(X)$, $E(Y)$, and $E(X\cdot Y)$ are finite (see Klenke, 2008, Remark 5.2, p. 102). Hence, according to Remark 3.42, there is no substantial loss of generality if we additionally assume that X and Y are real-valued. ◁

Definition 7.2 (Linear Quasi-Regression)
Let $X,Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be two real-valued random variables, assume $E(X^2),E(Y^2)<\infty$, and $\mathrm{Var}(X)>0$. Then the function $f\colon\mathbb{R}\to\mathbb{R}$ defined by

$$f(x) \;=\; \alpha_0+\alpha_1 x,\quad \forall\, x\in\mathbb{R}, \tag{7.1}$$

where the pair (α_0,α_1) minimizes the function $MSE\colon\mathbb{R}^2\to\mathbb{R}$ with

$$MSE(a_0,a_1) \;=\; E\big([Y-(a_0+a_1X)]^2\big),\quad \forall\, a_0,a_1\in\mathbb{R}, \tag{7.2}$$

is called the linear quasi-regression of Y on X. The composition of X and f is denoted by $Q_{lin}(Y|X)$, i.e.,

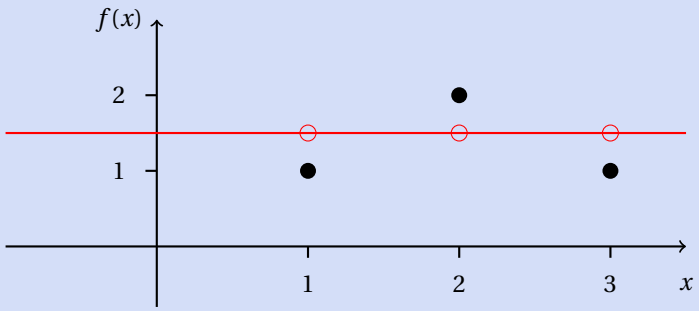


Figure 7.1. Linear quasi-regression

$$Q_{lin}(Y|X) \;=\; f(X) \;=\; \alpha_0 + \alpha_1 X. \tag{7.3}$$

Remark 7.3 (Distinguishing Between f and $f(X)$) Note that the linear quasi-regression $f\colon \mathbb{R} \to \mathbb{R}$ is a function assigning a real number to all real numbers. This applies even if X only takes on two different real values. In contrast, the number of different values of the composition $Q_{lin}(Y|X) = f(X)\colon \Omega \to \mathbb{R}$ is smaller or equal to the number of values of X. ◁

Example 7.4 (Discrete Regressor With Three Different Values) Let X and Y be real-valued random variables on $(\Omega, \mathscr{A}, P)$ with values $1,2,3$ and $1,2$, respectively. Furthermore, assume that their distribution is specified by

$$P(X{=}1, Y{=}1) = .25, \quad P(X{=}2, Y{=}2) = .5, \quad P(X{=}3, Y{=}1) = .25.$$

Then the linear quasi-regression $f\colon \mathbb{R} \to \mathbb{R}$ is specified by

$$f(x) \;=\; \alpha_0 + \alpha_1 \cdot x = 1.5 + 0 \cdot x = 1.5, \quad \forall\, x \in \mathbb{R},$$

and the composition of X and f is

$$Q_{lin}(Y|X) \;=\; \alpha_0 + \alpha_1 \cdot X = 1.5 + 0 \cdot X = 1.5$$

(see Exercise 7-3). The black points in Figure 7.1 represent the three pairs of values of X and Y. All values of the the linear quasi-regression are on the red line, which, in this example, is parallel to x-axis because its slope is 0. The red circles represent those values of the linear quasi-regression that have a nonzero probability $P_X(\{x\}) > 0$ with respect to the distribution of X. In contrast, in this example, $P_X(\{x\}) = 0$, for all $x \in \mathbb{R} \setminus \{1,2,3\}$. ◁

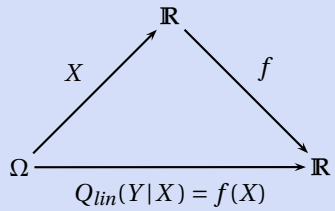


Figure 7.2. The regressor X, the linear quasi-regression f and their composition $Q_{lin}(Y|X) = f(X)$

Remark 7.5 (Linear Quasi-Regression vs. Regression) As the term 'linear quasi-regression' suggests, there is also a 'genuine' regression of Y on X (see Def. 10.25) and the two terms are not necessarily identical. As will be explained in more detail in Remark 10.27, the 'genuine' regression is a function $g\colon \mathbb{R} \to \mathbb{R}$ such that the composition $g(X)$ is X-measurable and minimizes the mean squared error $E\big([Y-g(X)]^2\big)$. In contrast to the conditional expectation $E(Y|X)$ [see Eq. (10.1)], the composition $Q_{lin}(Y|X) = f(X)$ minimizes the function *MSE* specified in Equation (7.2). Hence, f has to be a *linear* function even in those cases in which there are no $a_0, a_1 \in \mathbb{R}$, such that $E(Y|X) \underset{P}{=} a_0 + a_1 X$ (see Example 7.4). ◁

Remark 7.6 (Intercept and Slope) Note that the composition $Q_{lin}(Y|X) = f(X) = \alpha_0 + \alpha_1 X$ is a random variable on $(\Omega, \mathscr{A}, P)$ (see Fig. 7.2). The coefficient α_0 is called the *intercept* and α_1 the *slope* of (the linear quasi-regression) f (see Fig. 7.3). Obviously,

$$f(0) = \alpha_0 + \alpha_1 \cdot 0 = \alpha_0 . \tag{7.4}$$

Furthermore, if $x_1, x_2 \in \mathbb{R}$ and $x_2 > x_1$, then

$$\alpha_1 = \frac{1}{x_2 - x_1} \cdot \big[f(x_2) - f(x_1)\big] \tag{7.5}$$

(see Exercise 7-4). Equation (7.5) yields

$$\alpha_1 = f(x_2) - f(x_1), \quad \text{if } x_2 - x_1 = 1. \tag{7.6}$$

These equations justify calling α_0 the *intercept* and α_1 the *slope* of the linear quasi-regression f (see Fig. 7.3). Note that these equations also apply if $P(X{=}0) = P(X{=}x_1) = P(X{=}x_2) = 0$. They even apply if $0, x_1, x_2 \notin X(\Omega)$, because, by definition, f is a function on $\mathbb{R}$.

Figure 7.3 illustrates the intercept and the slope of a linear function such as the linear quasi-regression f. In this figure, $\alpha_0 = .5$ and $\alpha_1 = .85$. If X is discrete, then $Q_{lin}(Y|X) = f(X)$ is discrete as well. More precisely, the number of different values of $Q_{lin}(Y|X)$ is always smaller than or equal to the number of different values of X. In contrast, the linear quasi-regression $f\colon \mathbb{R} \to \mathbb{R}$ takes on uncountably many values unless its slope is 0. In this case its sole value is α_0. ◁

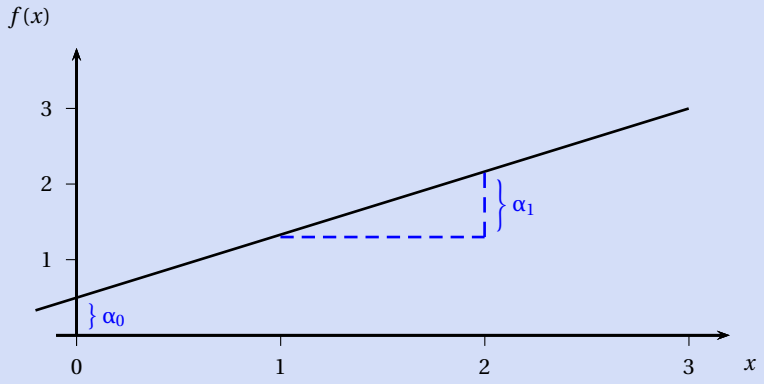


Figure 7.3. Intercept and slope of a linear function $f\colon \mathbb{R} \to \mathbb{R}$

7.2 Covariance

While the variance quantifies the variability of a numerical random variable, the covariance quantifies the degree of co-variation of two numerical random variables, i. e., the degree to which the two variables vary together in the following sense: If one variable takes on a large value (i. e., large positive deviation from its expectation), then the other one tends to take on a large value as well. Furthermore, if one variable takes on a small value (i. e., large negative deviation from its expectation), then the other one tends to take on a small value, too. In this case the covariance will be positive. However, the covariance may also be a negative real number. In this case, the two random variables co-vary in the following sense: If one variable takes on a large value, then the other one tends to take on a small value. Furthermore, if one variable takes on a small value, then the other one tends to take on a large value.

Definition 7.7 (Covariance)
Let $X, Y\colon (\Omega, \mathcal{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ be two numerical random variables with $E(X^2)$, $E(Y^2) < \infty$. Then the covariance of X and Y is defined by

$$Cov(X,Y) := E\big([X - E(X)] \cdot [Y - E(Y)]\big). \tag{7.7}$$

Comparing Equations (7.7) and (6.27) to each other shows that the variance is the covariance of a numerical random variable with itself.

Remark 7.8 (Correlated Numerical Random Variables) According to this definition, the *covariance* of X and Y is the expectation of the product of the mean

centered variables $X - E(X)$ and $Y - E(Y)$. Hence, a covariance can be negative, zero, or positive. If the covariance is different from zero, then we say that X and Y are *correlated*; otherwise, we say that they are *uncorrelated*. ◁

Remark 7.9 (Rules of Computation) The most important rules of computation for covariances are summarized in Box 7.1. Proofs are provided in Exercise 7-5. Rule (i) immediately implies

$$E(X \cdot Y) \;=\; E(X) \cdot E(Y) \;+\; Cov(X,Y)\,. \tag{7.8}$$

Hence, X and Y are uncorrelated if and only if $E(X \cdot Y) = E(X) \cdot E(Y)$. Furthermore, this equation and Theorem 6.24 imply that X and Y are uncorrelated if X and Y are independent.

Symmetry of the covariance [see Box 7.1 (v)] yields an alternative way to write Rule (viii) of Box 7.1:

$$Var\Big(\sum_{i=1}^{n} \alpha_i\, Y_i\Big) \;=\; \sum_{i=1}^{n} \alpha_i^2\, Var(Y_i) \;+\; 2 \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \alpha_i\, \alpha_j\, Cov(Y_i, Y_j). \tag{7.9}$$

This equation simplifies to

$$Var\Big(\sum_{i=1}^{n} \alpha_i\, Y_i\Big) \;=\; \sum_{i=1}^{n} \alpha_i^2\, Var(Y_i), \tag{7.10}$$

if $Y_1, \ldots, Y_n$ is a sequence of pairwise uncorrelated random variables. Note that independence of $Y_1, \ldots, Y_n$ implies $Cov(Y_i, Y_j) = 0,\ i \neq j$ [see Rule (vi)].

For $n = 2$, Rule (viii) simplifies to

$$\begin{aligned} &Var(\alpha_1\, Y_1 + \alpha_2\, Y_2) \\ &= \alpha_1^2\, Var(Y_1) + \alpha_2^2\, Var(Y_2) + 2\,\alpha_1\,\alpha_2 \cdot Cov(Y_1, Y_2)\,. \end{aligned} \tag{7.11}$$

Similarly, for $n = m = 2$, Rule (ix) simplifies to

$$\begin{aligned} &Cov(\alpha_1 X_1 + \alpha_2 X_2,\ \beta_1\, Y_1 + \beta_2\, Y_2) \\ &= \alpha_1 \beta_1\, Cov(X_1, Y_1) \;+\; \alpha_1 \beta_2 Cov(X_1, Y_2) \\ &\quad + \alpha_2 \beta_1\, Cov(X_2, Y_1) \;+\; \alpha_2 \beta_2 Cov(X_2, Y_2). \end{aligned} \tag{7.12}$$

◁

Remark 7.10 (Covariance of Indicators) For $A, B \in \mathscr{A}$, Rule (i) of Box 7.1 and Equations (1.32) and (6.4) yield

$$Cov(1_A, 1_B) \;=\; E(1_A \cdot 1_B) - E(1_A) \cdot E(1_B) \tag{7.13}$$

$$\;=\; P(A \cap B) - P(A) \cdot P(B). \tag{7.14}$$

◁

The following theorem helps to understand the relationship between the covariance and the variances of two numerical random variables X and Y.

Box 7.1 Rules of Computation for Covariances

Let X, Y be numerical random variables on the probability space $(\Omega, \mathscr{A}, P)$ with $E(X^2)$, $E(Y^2) < \infty$. Furthermore, let $\alpha, \beta \in \mathbb{R}$. Then:

$$Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y). \tag{i}$$

$$Cov(\alpha + X,\ \beta + Y) = Cov(X,Y). \tag{ii}$$

$$Cov(\alpha X,\ \beta Y) = \alpha\beta\, Cov(X,Y). \tag{iii}$$

$$Cov(X,X) = Var(X). \tag{iv}$$

$$Cov(X,Y) = Cov(Y,X). \tag{v}$$

$$X \perp\!\!\!\perp_P Y \quad \Rightarrow \quad Cov(X,Y) = 0. \tag{vi}$$

$$\exists \alpha \in \mathbb{R}\colon X \overset{=}{}_P \alpha \quad \Rightarrow \quad Cov(X,Y) = 0. \tag{vii}$$

If Y_i are real-valued random variables on the probability space $(\Omega, \mathscr{A}, P)$ with $E(Y_i^2) < \infty$ and $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$, then

$$Var\Big(\sum_{i=1}^{n} \alpha_i Y_i\Big) = \sum_{i=1}^{n} \alpha_i^2\, Var(Y_i) + \sum_{i=1}^{n} \sum_{j=1, i\neq j}^{n} \alpha_i \alpha_j\, Cov(Y_i, Y_j). \tag{viii}$$

If X_i, Y_j are real-valued random variables on the probability space $(\Omega, \mathscr{A}, P)$ with $E(X_i^2)$, $E(Y_j^2) < \infty$, and $\alpha_i, \beta_j \in \mathbb{R}$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$, then

$$Cov\Big(\sum_{i=1}^{n} \alpha_i X_i, \sum_{j=1}^{m} \beta_j Y_j\Big) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j\, Cov(X_i, Y_j). \tag{ix}$$

If $X_1 \overset{=}{}_P X_2$ and $E(Y^2)$, $E(X_1^2)$, $E(X_2^2) < \infty$, then

$$Cov(Y, X_1) = Cov(Y, X_2). \tag{x}$$

Theorem 7.11 (Cauchy-Schwarz Inequality)
If $X, Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ are random variables with $E(X^2)$, $E(Y^2) < \infty$, then

$$Cov(X,Y)^2 \leq Var(X) \cdot Var(Y). \tag{7.15}$$

Furthermore, if $Cov(X,Y) \neq 0$, then

$$Cov(X,Y)^2 = Var(X) \cdot Var(Y) \quad \Leftrightarrow \quad \exists\, a, b \in \mathbb{R}\colon Y \overset{=}{}_P a + b \cdot X. \tag{7.16}$$

(Proof p. 228)

Remark 7.12 (Squared Weigthed Sum of Random Variables) If $X, Y: (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ are random variables with $E(X^2)$, $E(Y^2) < \infty$ and $\alpha, \beta \in \mathbb{R}$, then $E[(\alpha X + \beta Y)^2] < \infty$ (see Exercise 7-2). ◁

In the following theorem we revisit the linear quasi-regression, studying three equivalent propositions. The first of these propositions deals with the residual variable $\epsilon := Y - f(X)$, where f is the linear quasi-regression of Y! on X. Note that this residual is not necessarily identical to the residual with respect to a conditional expectation that will be treated in chapters 9 to 11.

Theorem 7.13 (Three Characterizations of the Linear Quasi-Regression) *Let $X, Y: (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be two real-valued random variables with $E(X^2)$, $E(Y^2) < \infty$, and $Var(X) > 0$. Furthermore, let α_0, $\alpha_1 \in \mathbb{R}$, $f(X) = \alpha_0 + \alpha_1 X$ be the composition of X and $f: \mathbb{R} \to \mathbb{R}$, and define $\epsilon := Y - f(X)$. Then the following three propositions are equivalent to each other:*

(i) $E(\epsilon) = Cov(X, \epsilon) = 0$.

(ii) $\alpha_0 = E(Y) - \alpha_1 E(X)$ *and* $\alpha_1 = \dfrac{Cov(X,Y)}{Var(X)}$.

(iii) $f(X) = Q_{lin}(Y|X)$, *i. e.,* α_0, α_1 *minimize the function* $MSE(a_0, a_1)$ *defined by Equation* (7.2).

(Proof p. 229)

Remark 7.14 (Uniqueness) Suppose that the assumptions of Theorem 7.13 hold and $f(X) = Q_{lin}(Y|X)$. Then proposition (ii) of this theorem implies that the coefficients α_0 and α_1 are uniquely defined. Because $Q_{lin}(Y|X) = \alpha_0 + \alpha_1 X$, the linear quasi-regression $f: \mathbb{R} \to \mathbb{R}$ is uniquely defined as well. ◁

Remark 7.15 (Relationship Between Slope and Covariance) According to proposition (ii), a *zero covariance* between X and Y implies that the slopes of the linear quasi-regressions of Y on X and of X on Y are zero. A *negative covariance* implies that the slopes of the linear quasi-regressions of Y on X and of X on Y are negative, and a *positive covariance* implies that the slopes of the linear quasi-regressions of Y on X and of X on Y are positive. ◁

Example 7.16 (Discrete Regressor With Three Different Values – continued) In Example 7.4 we specified the distribution of (X,Y). Now we use the equations in Theorem 7.13 (ii) in order to compute the coefficients α_0 and α_1. For this purpose we have to compute the expectations of X and Y, the variance of X, and the covariance of X and Y:

$$E(X) = \sum_{i=1}^{n} x_i \cdot P(X{=}x_i) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} = 2,$$

$$E(Y) = \sum_{i=1}^{m} y_i \cdot P(Y{=}y_i) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2},$$

$$\begin{aligned}
Var(X) &= E(X^2) - E(X)^2 = \sum_{i=1}^{n} x_i^2 \cdot P(X{=}x_i) - E(X)^2 \quad \text{[Box 6.2 (i), (6.19)]} \\
&= 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{2} + + 3^2 \cdot \frac{1}{4} - 2^2 = \frac{1}{2}, \\
Cov(X,Y) &= E(X \cdot Y) - E(X) \cdot E(Y) \quad \text{[Box 7.1 (i)]} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} x_i \cdot y_j \cdot P\big((X,Y){=}(x_i,y_i)\big) - E(X) \cdot E(Y) \quad \text{[(6.3)]} \\
&= 1 \cdot 1 \cdot \frac{1}{4} + 2 \cdot 2 \cdot \frac{1}{2} + 3 \cdot 1 \cdot \frac{1}{4} - 2 \cdot \frac{3}{2} \\
&= \frac{1}{4} + \frac{8}{4} + \frac{3}{4} - \frac{12}{4} = 0.
\end{aligned}$$

Using the equations in Theorem 7.13 (ii) yields $\alpha_1 = Cov(X,Y)/Var(X) = \frac{0}{1/2} = 0$ and $\alpha_0 = E(Y) - \alpha_1 E(X) = \frac{3}{2} - 0 \cdot 2 = 1.5$, the same result as obtained in Exercise 7-3, in which we minimize the function $MSE(a_0, a_1)$. ◁

7.3 Correlation

As mentioned before, the covariance between two numerical random variables quantifies the strength of the dependence that can be described by a linear quasi-regression. However, the covariance is not invariant under multiplication with constants [scale transformations; see Box 7.1 (iii)] of the random variables involved. In contrast, the correlation, which quantifies the strength of the same kind of dependence *is invariant* under scale transformations (see Rem. 7.21).

Definition 7.17 (Correlation)
Let $X, Y : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be two numerical random variables with $E(X^2)$, $E(Y^2) < \infty$. Then the correlation of X and Y is defined by

$$Corr(X,Y) := \begin{cases} \dfrac{Cov(X,Y)}{SD(X) \cdot SD(Y)}, & \text{if } SD(X), SD(Y) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7.17}$$

Remark 7.18 (Correlation of a Random Variable With Itself) Because $Cov(X,X) = Var(X) = SD(X) \cdot SD(X)$, Equation (7.17) implies that $Corr(X,X) = 1$. Similarly, because $Cov(X,-X) = -Var(X) = -SD(X) \cdot SD(X)$, Equation (7.17) implies that $Corr(X,-X) = -1$. ◁

Remark 7.19 (Range of the Correlation) Note that

$$-1 \le Corr(X,Y) \le 1, \tag{7.18}$$

provided that $Corr(X,Y)$ exists, i. e., provided that the assumptions hold under which the correlation is defined. Equation (7.18) is an immediate implication of Equation (7.15). ◁

Remark 7.20 (Correlation and Z-Transformed Variables) If the standard deviations of X and Y are positive, then the correlation is also the expectation of the product of the Z-transformed variables [see Eq. (6.33)], i. e.,

$$Corr(X,Y) = E\left(\frac{X-E(X)}{SD(X)} \cdot \frac{Y-E(Y)}{SD(Y)}\right) \tag{7.19}$$

(see Exercise 7-6). ◁

Remark 7.21 (An Invariance Property of the Correlation) The correlation of linear transformations of X and Y is

$$Corr(a_0 + a_1 X,\ b_0 + b_1 Y) = \begin{cases} Corr(X,Y), & \text{if } a_1 \cdot b_1 > 0 \\ -Corr(X,Y), & \text{if } a_1 \cdot b_1 < 0, \end{cases} \tag{7.20}$$

where $a_0, a_1, b_0, b_1 \in \mathbb{R}$ (see Exercise 7-7). This equation means that the correlation is invariant (up to change of signs) under linear transformations, which include *translations* ($a_1 = 1$ and $b_1 = 1$) and *scale transformations* ($a_0 = b_0 = 0$ and $a_1, b_1 \neq 0$). ◁

In the special case $Y \underset{P}{=} a_0 + a_1 X$, $a_0, a_1 \in \mathbb{R}$, Equation (7.20) and Remark 7.18 imply that $Corr(X,Y) = 1$ if $a_1 > 0$ and $Corr(X,Y) = -1$ if $a_1 < 0$. Theorem 7.11 implies that the other direction of implication holds as well.

Corollary 7.22 (Perfect Correlation of Two Random Variables)
Let the assumptions of Definition 7.2 be satisfied and let $Var(X), Var(Y) > 0$. Then $|Corr(X,Y)| = 1$ if and only if there are $a_0, a_1 \in \mathbb{R}$, $a_1 \neq 0$, such that $Y \underset{P}{=} a_0 + a_1 X$. In this case

$$Corr(X,Y) = \begin{cases} 1, & \text{if } a_1 > 0 \\ -1, & \text{if } a_1 < 0. \end{cases} \tag{7.21}$$

Remark 7.23 (Covariance and Standard Deviations) Let $Y \underset{P}{=} a_0 + a_1 X$, $a_0, a_1 \in \mathbb{R}$. If $a_1 \geq 0$, then $Cov(X,Y) = SD(X) \cdot SD(Y)$. If $a_1 < 0$ then $Cov(X,Y) = -SD(X) \cdot SD(Y)$ (see Exercise 7-8). ◁

Remark 7.24 (Slope of a Linear Quasi-Regression and Correlation) If α_1 is the slope of $Q_{lin}(Y|X)$ (see Def. 7.2), then

$$\alpha_1 = \frac{Cov(X,Y)}{Var(X)} = Corr(X,Y) \cdot \frac{SD(Y)}{SD(X)} \tag{7.22}$$

[see proposition (ii) of Th. 7.13]. This equation shows that the slope α_1 of the linear quasi-regression has the same sign as the covariance and the correlation. The size of the absolute value of α_1 depends on the ratio $SD(Y)/SD(X)$ of the standard deviations. The smaller the standard deviation of X compared to the standard deviation of Y, the larger the absolute value of α_1. Furthermore, given a fixed variance $Var(X)$, this equation also shows that α_1 is proportional to $Cov(X,Y)$ and $Corr(X,Y)$. In this sense, all three parameters α_1, $Cov(X,Y)$, and $Corr(X,Y)$ quantify the strength of the dependence of Y on X described by a linear quasi-regression. Note, however, that α_1 and $Cov(X,Y)$ are not invariant under scale transformations of X and Y. This can be seen in the following equation for the slope α_1^* of the linear quasi-regression of bY on aX, $a,b \in \mathbb{R}$, $a,b \neq 0$:

$$\alpha_1^* \;=\; \frac{Cov(aX,bY)}{Var(aX)} \;=\; Corr(aX,bY)\cdot\frac{SD(bY)}{SD(aX)} \;=\; \frac{b}{a}\cdot\alpha_1. \tag{7.23}$$

Hence, the slope of the linear quasi-regression of bY on aX is identical to the slope of the linear quasi-regression of Y on X multiplied by $\frac{b}{a}$. In contrast, the slope of the linear quasi-regression is invariant under translations $c+X$, $d+Y$, $c,d \in \mathbb{R}$ (see Exercise 7-9). ◁

Example 7.25 (Joe and Ann With Random Assignment – continued) Consider the example presented in Table 2.2 (p. 54). In this example, the covariance of X and Y is most easily computed using

$$\begin{aligned} Cov(X,Y) \;&=\; E(X\cdot Y) - E(X)\cdot E(Y) && \text{[Box 7.1 (i)]}\\ &=\; \sum_{(x,y)} (x\cdot y)\cdot P(X{=}x, Y{=}y) \;-\; P(X{=}1)\cdot P(Y{=}1) && \text{[(6.15)]}\\ &=\; P(X{=}1, Y{=}1) \;-\; P(X{=}1)\cdot P(Y{=}1)\\ &=\; (.16+.08) \;-\; .40\cdot .51 \;=\; .036, \end{aligned}$$

where $P(X{=}1) = E(X) = .40$ and $P(Y{=}1) = E(Y) = .51$ have been computed in Example 6.7. Note that $\sum_{(x,y)}$ is the sum over all pairs (x,y) of values of X and Y. In this example, there are four such pairs, only one of which, namely $(1,1)$, yields a product $x\cdot y \neq 0$. Using the results of Example 6.31 on the variances of X and Y yields the correlation

$$Corr(X,Y) \;=\; \frac{Cov(X,Y)}{SD(X)\cdot SD(Y)} \;=\; \frac{.036}{\sqrt{.24}\cdot\sqrt{.2499}} \;\approx\; .147$$

Hence, treatment and outcome variables have a positive correlation. This is in accordance with comparing the conditional probability of success given treatment, $P(C\,|\,B) = .60$, to the conditional probability of success given no treatment, $P(C\,|\,B^c) = .45$ (see Example 4.15).

In this example,

$$Q_{lin}(Y\,|\,X) \;=\; \alpha_0 \;+\; \alpha_1\cdot X$$

$$
\begin{aligned}
&= \left(E(Y) - \frac{Cov(X,Y)}{Var(X)} \cdot E(X)\right) + \frac{Cov(X,Y)}{Var(X)} \cdot X \qquad \text{[Th. 7.13 (ii)]} \\
&= \left(.51 - \frac{.036}{.24} \cdot .40\right) + \frac{.036}{.24} \cdot X \\
&= .45 + .15 \cdot X,
\end{aligned}
$$

and the linear quasi-regression $f : \mathbb{R} \to \mathbb{R}$ of Y on X is specified by

$$f(x) = .45 + .15 \cdot x, \quad \forall\, x \in \mathbb{R}.$$

◁

7.4 Expectation Vector and Covariance Matrix

7.4.1 Random Vector and Random Matrix

Let $X = (X_1, \ldots, X_n)$ be an n-variate numerical random variable on a probability space $(\Omega, \mathscr{A}, P)$. In order to utilize matrix algebra, we consider the *column random vector*

$$\boldsymbol{x} := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},$$

i. e., of a column vector of the random variables $X_1, \ldots, X_n$. Correspondingly, we consider the *row random vector* $\boldsymbol{x}' := [X_1, \ldots, X_n]$, the *transpose* of $\boldsymbol{x}$.

In this section, we also consider a *random matrix*, i. e., a matrix

$$\boldsymbol{X} := \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1m} \\ X_{21} & X_{22} & \ldots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nm} \end{bmatrix} \tag{7.24}$$

of type $n \times m$ of numerical random variables X_{ij} on a probability space $(\Omega, \mathscr{A}, P)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Such a random matrix is an $n \times m$-array of an $(n \cdot m)$-variate random variable (see section 5.3).

7.4.2 Expectations of a Random Vector and a Random Matrix

The *expectation of a (row or column) random vector* is defined as the (row or column) vector of the expectations of its components, i. e.,

$$E(\boldsymbol{x}') := \big[E(X_1), \ldots, E(X_n)\big] \tag{7.25}$$

and $E(\boldsymbol{x}) := \big[E(X_1), \ldots, E(X_n)\big]'$, provided that the expectations exist. Hence,

$$E(\boldsymbol{x}') = \big(E(\boldsymbol{x})\big)'. \tag{7.26}$$

Analogously to Equation (7.25), the *expectation of an $n \times m$-random matrix* is defined as the $n \times m$-matrix of the expectations of its components, i. e.,

$$E\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix} := \begin{bmatrix} E(X_{11}) & E(X_{12}) & \dots & E(X_{1m}) \\ E(X_{21}) & E(X_{22}) & \dots & E(X_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \dots & E(X_{nm}) \end{bmatrix}, \tag{7.27}$$

provided that the expectations exist. Obviously, if $\boldsymbol{X}'$ denotes the transpose of the matrix $\boldsymbol{X}$, then

$$E(\boldsymbol{X}') \;=\; \big(E(\boldsymbol{X})\big)'. \tag{7.28}$$

In Box 7.2 we present some rules of computation for the expectations of random vectors and random matrices (for proofs see Exercise 7-10). In this box we use

$$\mathbf{a}'\boldsymbol{x} := \sum_{i=1}^{n} a_i \cdot X_i, \tag{7.29}$$

the *inner product* of an n-vector $\mathbf{a} = [a_1, \dots, a_n]'$ of real numbers and the random vector $\boldsymbol{x}$. Correspondingly,

$$\mathbf{A}\boldsymbol{x} := \begin{bmatrix} \mathbf{a}_1'\boldsymbol{x} \\ \vdots \\ \mathbf{a}_k'\boldsymbol{x} \end{bmatrix}, \tag{7.30}$$

where $\mathbf{a}_l'$, $l = 1, \dots, k$, denotes the lth row of the $(k \times n)$-matrix $\mathbf{A}$ of real numbers.

7.4.3 Covariance Matrix of two Multivariate Random Variables

Now we consider two multivariate numerical random variables $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$ on a probability space $(\Omega, \mathscr{A}, P)$. In particular, we assume that the second moments of all these random variables are finite and focus on their covariance matrix, again utilizing the representation of X and Y as row or column vectors that has been introduced at the beginning of section 7.4.1.

Note that $[\boldsymbol{x} - E(\boldsymbol{x})]\,[\boldsymbol{y} - E(\boldsymbol{y})]'$ is an $n \times m$-matrix of the random variables

$$[X_i - E(X_i)] \cdot [Y_j - E(Y_j)], \quad i = 1, \dots, n,\ j = 1, \dots, m.$$

Therefore, using (7.27), the *covariance matrix* $\boldsymbol{\Sigma}_{\boldsymbol{xy}}$ is defined by

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} \;:=\; E\big([\boldsymbol{x} - E(\boldsymbol{x})]\,[\boldsymbol{y} - E(\boldsymbol{y})]'\big). \tag{7.31}$$

In other words, the covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$ is the matrix of covariances, i. e.,

Box 7.2 Rules of Computation for Expectations of Random Matrices

Let $X = (X_1,\dots,X_n)$ be an n-variate and and $Y = (Y_1,\dots,Y_m)$ an m-variate real-valued random variable on a probability space $(\Omega,\mathscr{A},P)$ such that the expectations of X_i and Y_j are finite, for all $i = 1,\dots,n$, $j = 1,\dots,m$. Furthermore, let $\boldsymbol{x} = [X_1,\dots,X_n]'$ and $\boldsymbol{y} = [Y_1,\dots,Y_m]'$ denote column vectors, $\mathbf{a} = [a_1,\dots,a_n]'$ a column vector of real numbers, and let $\mathbf{A}$ and $\mathbf{B}$ be matrices of types $k\times n$ and $k\times m$, respectively, each of their components being a real number. Furthermore, let $\mathbf{C}$ and $\mathbf{D}$ be matrices of real numbers of types $l\times n$ and $r\times m$, respectively. Finally, let $\boldsymbol{X}$ be an $(n\times k)$-matrix of real-valued random variables on $(\Omega,\mathscr{A},P)$, all with finite expectations. Then

$$\boldsymbol{x} \underset{P}{=} \mathbf{a} \;\Rightarrow\; E(\boldsymbol{x}) = \mathbf{a}. \tag{i}$$

$$E(\mathbf{a}+\boldsymbol{x}) = \mathbf{a} + E(\boldsymbol{x}). \tag{ii}$$

$$E(\mathbf{a}'\boldsymbol{x}) = \mathbf{a}'E(\boldsymbol{x}) = E(\boldsymbol{x})'\mathbf{a} = E(\boldsymbol{x}'\mathbf{a}). \tag{iii}$$

$$E(\mathbf{A}\boldsymbol{x}) = \mathbf{A}E(\boldsymbol{x}). \tag{iv}$$

$$E(\mathbf{A}\boldsymbol{x}+\mathbf{B}\boldsymbol{y}) = \mathbf{A}E(\boldsymbol{x}) + \mathbf{B}E(\boldsymbol{y}). \tag{v}$$

Let $\boldsymbol{X}$ be an $(n\times k)$-matrix and $\boldsymbol{Y}$ an $(m\times k)$-matrix of real-valued random variables on $(\Omega,\mathscr{A},P)$, all with finite second moments. Then

$$\boldsymbol{X} \underset{P}{=} \mathbf{A}' \;\Rightarrow\; E(\boldsymbol{X}) = \mathbf{A}'. \tag{vi}$$

$$E(\mathbf{A}'+\boldsymbol{X}) = \mathbf{A}' + E(\boldsymbol{X}). \tag{vii}$$

$$E(\mathbf{C}\boldsymbol{X}) = \mathbf{C}E(\boldsymbol{X}). \tag{viii}$$

$$E(\mathbf{C}\boldsymbol{X}\boldsymbol{Y}'\mathbf{D}') = \mathbf{C}E(\boldsymbol{X}\boldsymbol{Y}')\mathbf{D}'. \tag{ix}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \begin{bmatrix} \sigma_{X_1Y_1} & \sigma_{X_1Y_2} & \dots & \sigma_{X_1Y_m} \\ \sigma_{X_2Y_1} & \sigma_{X_2Y_2} & \dots & \sigma_{X_2Y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_nY_1} & \sigma_{X_nY_2} & \dots & \sigma_{X_nY_m} \end{bmatrix}, \tag{7.32}$$

where $\sigma_{X_iY_j} := Cov(X_i,Y_j) = E\big([X_i - E(X_i)]\cdot[Y_j - E(Y_j)]\big)$, $i = 1,\dots,n,\ j = 1,\dots,m$. If we assume that the second moments of the random variables X_i and Y_j are finite, then all covariances $Cov(X_i,Y_j)$ are finite as well, and we say that $\boldsymbol{\Sigma}_{\boldsymbol{xy}}$ *exists*.

If we consider a univariate random variable Y, then $\boldsymbol{y} = [Y]$ is also a vector consisting of a single component, the random variable Y. In this special case, $\boldsymbol{\Sigma}_{\boldsymbol{xy}}$ is a matrix of type $n\times 1$, the column vector

Box 7.3 Rules of Computation for Covariance Matrices

Let $X = (X_1, \ldots, X_n)$ be an n-variate and and $Y = (Y_1, \ldots, Y_m)$ an m-variate real-valued random variable on a probability space $(\Omega, \mathscr{A}, P)$ such that the second moments of X_i and Y_j are finite, for all $i = 1, \ldots, n$, $j = 1, \ldots, m$. Furthermore, let $\boldsymbol{x} = [X_1, \ldots, X_n]'$ and $\boldsymbol{y} = [Y_1, \ldots, Y_m]'$ denote column vectors, $\mathbf{a} = [a_1, \ldots, a_n]'$ and $\mathbf{b} = [b_1, \ldots, b_m]'$ column vectors of real numbers, and let $\mathbf{A}$ and $\mathbf{B}$ be matrices of types $k \times n$ and $k \times m$, respectively, each of their components being a real number. Finally, let $\mathbf{O}$ denote the $(n \times m)$-null matrix. Then:

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} = E(\boldsymbol{x}\,\boldsymbol{y}') - E(\boldsymbol{x})\,E(\boldsymbol{y}'). \tag{i}$$

$$\boldsymbol{\Sigma}_{\mathbf{a}+\boldsymbol{x},\mathbf{b}+\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{xy}}. \tag{ii}$$

$$\boldsymbol{\Sigma}_{\mathbf{A}\boldsymbol{x},\mathbf{B}\boldsymbol{y}} = \mathbf{A}\,\boldsymbol{\Sigma}_{\boldsymbol{xy}}\,\mathbf{B}'. \tag{iii}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \boldsymbol{\Sigma}'_{\boldsymbol{yx}}. \tag{iv}$$

$$X \underset{P}{\perp\!\!\!\perp} Y \;\Rightarrow\; \boldsymbol{\Sigma}_{\boldsymbol{xy}} = \mathbf{O}. \tag{v}$$

$$\boldsymbol{x} \underset{P}{=} \mathbf{a} \;\Rightarrow\; \boldsymbol{\Sigma}_{\boldsymbol{xy}} = \mathbf{O}. \tag{vi}$$

Additionally, let $\boldsymbol{w} = [W_1, \ldots, W_r]'$ and $\boldsymbol{z} = [Z_1, \ldots, Z_s]'$ be real-valued random column vectors on $(\Omega, \mathscr{A}, P)$ such that all their components have finite second moments. Furthermore, let $\mathbf{C}$ and $\mathbf{D}$ be matrices of real numbers of type $l \times r$ and $l \times s$. Then:

$$\boldsymbol{\Sigma}_{\mathbf{A}\boldsymbol{x}+\mathbf{B}\boldsymbol{y},\mathbf{C}\boldsymbol{w}+\mathbf{D}\boldsymbol{z}} = \mathbf{A}\,\boldsymbol{\Sigma}_{\boldsymbol{xw}}\,\mathbf{C}' + \mathbf{A}\,\boldsymbol{\Sigma}_{\boldsymbol{xz}}\,\mathbf{D}' + \mathbf{B}\,\boldsymbol{\Sigma}_{\boldsymbol{yw}}\,\mathbf{C}' + \mathbf{B}\,\boldsymbol{\Sigma}_{\boldsymbol{yz}}\,\mathbf{D}'. \tag{vii}$$

$$n = s \text{ and } \boldsymbol{x} \underset{P}{=} \boldsymbol{z} \;\Rightarrow\; \boldsymbol{\Sigma}_{\boldsymbol{xy}} = \boldsymbol{\Sigma}_{\boldsymbol{zy}}. \tag{viii}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \begin{bmatrix} \sigma_{X_1 Y} \\ \vdots \\ \sigma_{X_n Y} \end{bmatrix}.$$

Another special case is $\boldsymbol{x} = \boldsymbol{y}$. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$ of $\boldsymbol{x}$ and $\boldsymbol{x}$ is called the *variance-covariance matrix of* $\boldsymbol{x}$ (or of X). Hence,

$$\boldsymbol{\Sigma}_{\boldsymbol{xx}} := E\big([\boldsymbol{x} - E(\boldsymbol{x})]\,[\boldsymbol{x} - E(\boldsymbol{x})]'\big) \tag{7.33}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{xx}} = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma^2_{X_n} \end{bmatrix}. \tag{7.34}$$

The diagonal components of the matrix $\mathbf{\Sigma}_{\boldsymbol{xx}}$ are the variances of the variables $X_1,\ldots,X_n$, because $\sigma_{X_iX_i} := Cov(X_i,X_i) = Var(X_i) = \sigma^2_{X_i},\ i = 1,\ldots,n$.

In Box 7.3 we present some rules of computation for covariance matrices. They are proved in Exercise 7-11.

7.5 Multiple Linear Quasi-Regression

In the following definition we generalize the concept of a linear quasi-regression considering a multivariate regressor $X = (X_1,\ldots,X_n)$. We use the notation $\boldsymbol{x} = [X_1,\ldots,X_n]'$ to denote the column vector of X, $\boldsymbol{\beta}' = [\beta_1,\ldots,\beta_n]$ for the row vector of the real numbers $\beta_1,\ldots,\beta_n$, and $\mathbf{b}' = [b_1,\ldots,b_n]$ for the row vector of the real numbers $b_1,\ldots,b_n$.

Definition 7.26 (Multiple Linear Quasi-Regression)
Let $X_i\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$, $i = 1,\ldots,n$, and $Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be real-valued random variables, define $X := (X_1,\ldots,X_n)$, assume $E(X_i^2), E(Y^2) < \infty$, $i = 1,\ldots,n$, and that the inverse $\mathbf{\Sigma}^{-1}_{\boldsymbol{xx}}$ exists. Define the function $f\colon\mathbb{R}^n\to\mathbb{R}$ by

$$f(x) \;=\; \beta_0 + \sum_{i=1}^{n}\beta_i\, x_i, \quad \forall\, x = (x_1,\ldots,x_n)\in\mathbb{R}^n, \tag{7.35}$$

where $\beta_0, \boldsymbol{\beta} = [\beta_1,\ldots,\beta_n]'$ minimize the function $MSE\colon\mathbb{R}^{n+1}\to\mathbb{R}$ with

$$MSE(b_0,b_1,\ldots,b_n) \;=\; E\big([Y-(b_0+\mathbf{b}'\boldsymbol{x})]^2\big), \quad \forall (b_0,b_1,\ldots,b_n)\in\mathbb{R}^{n+1}. \tag{7.36}$$

Then f is called the linear quasi-regression of Y on $X_1,\ldots,X_n$. The composition of X and f is denoted by $Q_{lin}(Y|X)$ or $Q_{lin}(Y|X_1,\ldots,X_n)$, i. e.,

$$Q_{lin}(Y|X) \;:=\; f(X) \;=\; \beta_0+\boldsymbol{\beta}'\boldsymbol{x} \;=\; \beta_0+\sum_{i=1}^{n}\beta_i\, X_i. \tag{7.37}$$

Remark 7.27 (Coefficient of Determination) Let the assumptions of definition 7.26 hold assume $Var(Y) > 0$. Then

$$Q^2_{Y|X} \;:=\; \frac{Var[Q_{lin}(Y|X)]}{Var(Y)} \tag{7.38}$$

is called the *coefficient of determination* of the linear quasi-regression $Q_{lin}(Y|X)$. ◁

In the following theorem, we generalize Theorem 7.13 considering a multivariate real-valued regressor $X = (X_1,\ldots,X_n)$. In this theorem $\mathbf{\Sigma}_{\boldsymbol{x}\epsilon}$ denotes the covariance vector of $\boldsymbol{x}$ and ϵ, which is defined by

$$\epsilon := Y - Q_{lin}(Y|X_1,\ldots,X_n) \tag{7.39}$$

and called the *residual of Y with respect to its linear quasi-regression on X*.

Theorem 7.28 (Charactizations of the Multiple Linear Quasi-Regression) *Let $X_1,\ldots,X_n,Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be real-valued random variables, assume $E(X_i^2), E(Y^2)<\infty$ for all $i=1,\ldots,n$, and that the inverse $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ exists. Furthermore, let*

$$f(X_1,\ldots,X_n) := \beta_0+\boldsymbol{\beta}'\boldsymbol{x},\quad \beta_0\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^n, \tag{7.40}$$

be the composition of a function $f\colon\mathbb{R}^n\to\mathbb{R}$, and define $\epsilon := Y-f(X_1,\ldots,X_n)$. Then the following three propositions are equivalent to each other:

(i) $E(\epsilon)=0$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{x}\epsilon}=\boldsymbol{0}$.

(ii) $\beta_0 = E(Y)-\boldsymbol{\beta}'E(\boldsymbol{x})$ *and* $\boldsymbol{\beta}=\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xy}}$.

(iii) $f(X_1,\ldots,X_n)=Q_{lin}(Y|X_1,\ldots,X_n)$.

(Proof p. 230)

Remark 7.29 (Uniqueness) Suppose that the assumptions of Theorem 7.28 hold and $f(X_1,\ldots,X_n)=Q_{lin}(Y|X)$. Then proposition (ii) of this theorem implies that the coefficients $\beta_0,\beta_1,\ldots,\beta_n$ are uniquely defined. Because $f(X_1,\ldots,X_n)=\beta_0+\boldsymbol{\beta}'\boldsymbol{x}$ $=Q_{lin}(Y|X)$, the linear quasi-regression f and $Q_{lin}(Y|X)$ are uniquely defined as well. ◁

7.6 Proofs

Proof of Theorem 7.11

Suppose $Var(Y)=0$. Then rules (iv) of Box 6.2 and (vii) of Box 7.1 imply that $Cov(X,Y)=0$. This shows that the Inequality (7.15) holds if $Var(Y)=0$. Now suppose $Var(Y)>0$. Then

$$\begin{aligned}
0 &\le Var\left(X-\frac{Cov(X,Y)}{Var(Y)}\cdot Y\right)\cdot Var(Y)\\
&= \left(Var(X)-2\cdot\frac{Cov(X,Y)}{Var(Y)}Cov(X,Y)+\frac{Cov(X,Y)^2}{Var(Y)^2}Var(Y)\right)\cdot Var(Y) \qquad [(7.11)]\\
&= Var(X)\cdot Var(Y) - Cov(X,Y)^2.
\end{aligned}$$

According to the first part of the proof,

$$Cov(X,Y)^2 = Var(X)\cdot Var(Y) \ \Leftrightarrow\ Var(Y)=0 \text{ or } Var\left(X-\frac{Cov(X,Y)}{Var(Y)}\cdot Y\right)=0.$$

Rule (iv) of Box 6.2 implies that this is equivalent to

$$\exists\, a \in \mathbb{R} : Y \underset{P}{=} a \quad \text{or} \quad \exists\, c \in \mathbb{R} : X - \frac{Cov(X,Y)}{Var(Y)} \cdot Y \underset{P}{=} c.$$

Because $Cov(X,Y) \neq 0$, this is equivalent to

$$\exists\, a \in \mathbb{R} : Y \underset{P}{=} a + 0 \cdot X \quad \text{or} \quad \exists\, c \in \mathbb{R} : Y \underset{P}{=} \left(-\frac{c \cdot Var(Y)}{Cov(X,Y)}\right) + \left(\frac{Var(Y)}{Cov(X,Y)}\right) \cdot X.$$

Obviously, in both cases, there is a linear function of X that is P-equivalent to Y. Furthermore, if there are $a, b \in \mathbb{R}$ with $Y \underset{P}{=} a + bX$, then

$$\begin{aligned} Cov(X,Y)^2 &= Cov(X, a+bX)^2 && \text{[Box 7.1 (x)]} \\ &= b^2 Cov(X,X)^2 && \text{[Box 7.1 (ii), (iii)]} \\ &= b^2 Var(X)^2 && \text{[Box 7.1 (ii), (iv)]} \\ &= Var(X) \cdot Var(Y). && \text{[Box 6.2 (ii), (iii)]} \end{aligned}$$

Proof of Theorem 7.13

The proof is organized as follows: (iii) ⇒ (ii) ⇒ (i) ⇒ (iii), which will prove that (i), (ii), and (iii) are equivalent.

(iii) ⇒ (ii). The first partial derivatives of

$$\begin{aligned} MSE(a_0, a_1) &= E\left([Y - (a_0 + a_1 X)]^2\right) \\ &= E(Y^2) + E[(a_0 + a_1 X)^2] - 2E[Y \cdot (a_0 + a_1 X)] && \text{[Box 6.1 (vi)]} \\ &= E(Y^2) + a_0^2 + a_1^2 E(X^2) + 2a_0 a_1 E(X) - 2a_0 E(Y) - 2a_1 E(X \cdot Y) \end{aligned}$$

with respect to a_0 and a_1 are

$$\frac{\partial MSE(a_0, a_1)}{\partial a_0} = 2a_0 + 2a_1 E(X) - 2E(Y)$$

and

$$\frac{\partial MSE(a_0, a_1)}{\partial a_1} = 2a_1 E(X^2) + 2a_0 E(X) - 2E(X \cdot Y).$$

If $f(X) = \alpha_0 + \alpha_1 X = Q_{lin}(Y \mid X)$, then

$$2\alpha_0 + 2\alpha_1 E(X) - 2E(Y) = 0$$

and

$$2\alpha_1 E(X^2) + 2\alpha_0 E(X) - 2E(X \cdot Y) = 0.$$

Solving the first equation for α_0 yields

$$\alpha_0 = E(Y) - \alpha_1 E(X).$$

Inserting this result into the second equation yields

$$\alpha_1 E(X^2) + E(Y) \cdot E(X) - \alpha_1 E(X)^2 - E(X \cdot Y) = 0.$$

Using $Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(X)$ and $Var(X) = E(X^2) - E(X)^2$ we receive

$$\alpha_1 = \frac{Cov(X,Y)}{Var(X)}.$$

(ii) $\Rightarrow$ (i).

$$\begin{aligned}
E(\epsilon) &= E[Y - f(X)] \\
&= E[Y - (\alpha_0 + \alpha_1 X)] && \text{[def. of } f(X)\text{]} \\
&= E\big(Y - [E(Y) - \alpha_1 E(X) + \alpha_1 X]\big) && \text{[(ii)]} \\
&= E(Y) - E(Y) + \alpha_1 E(X) - \alpha_1 \cdot E(X) && \text{[Box 6.1 (vi)]} \\
&= 0. \\
Cov(X,\epsilon) &= Cov\big(X, [Y - (\alpha_0 + \alpha_1 X)]\big) && \text{[def. of } \epsilon\text{]} \\
&= Cov(X,Y) - \alpha_1 Var(X) && \text{[Box 7.1 (ii), (iii)]} \\
&= Cov(X,Y) - \frac{Cov(X,Y)}{Var(X)} \cdot Var(X) && \text{[(ii)]} \\
&= 0.
\end{aligned}$$

(i) $\Rightarrow$ (iii). Let $f(X) := \alpha_0 + \alpha_1 X$, $\alpha_0, \alpha_1 \in \mathbb{R}$, be a linear function of X with $E(\epsilon) = 0$ and $Cov(X,\epsilon) = 0$, where $\epsilon := Y - f(X)$. Then, for any linear function $h(X) := a_0 + a_1 X$, $a_0, a_1 \in \mathbb{R}$,

$$\begin{aligned}
E\big((Y - f(X))[f(X) - h(X)]\big) &= E\big(\epsilon \cdot [f(X) - h(X)]\big) && \text{[def. of } \epsilon\text{]} \\
&= E\big(\epsilon \cdot [(\alpha_0 + \alpha_1 X) - (a_0 + a_1 X)]\big) && \text{[defs. of } f(X), h(X)\text{]} \\
&= E\big(\epsilon \cdot [(\alpha_0 - a_0) + (\alpha_1 - a_1) X]\big) \\
&= (\alpha_0 - a_0) \cdot E(\epsilon) + (\alpha_1 - a_1) E(\epsilon \cdot X) && \text{[Box 6.1 (vi)]} \\
&= 0. && \text{[(i)]}
\end{aligned}$$

Using this result and considering

$$\begin{aligned}
E\big([Y - h(X)]^2\big) &= E\big[\big([Y - f(X)] + [f(X) - h(X)]\big)^2\big] \\
&= E\big([Y - f(X)]^2\big) + E\big([f(X) - h(X)]^2\big) + 2 \cdot E\big([Y - f(X)][f(X) - h(X)]\big) \\
&= E\big([Y - f(X)]^2\big) + E\big([f(X) - h(X)]^2\big) \\
&\geq E\big([Y - f(X)]^2\big),
\end{aligned}$$

and '=' holds if $f(X) \underset{P}{=} h(X)$.

Proof of Theorem 7.28

The proof is organized as follows: (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) $\Rightarrow$ (iii), which will prove that (i), (ii), and (iii) are equivalent.

(iii) $\Rightarrow$ (ii). The first partial derivative of

$$\begin{aligned}
&MSE(b_0, b_1, \ldots, b_n) \\
&= E\big([Y - (b_0 + \mathbf{b}'\boldsymbol{x})]^2\big) && \text{[(7.36)]} \\
&= E(Y^2) + E\big((b_0 + \mathbf{b}'\boldsymbol{x})^2\big) - 2E\big(Y \cdot (b_0 + \mathbf{b}'\boldsymbol{x})\big) && \text{[Box 6.1 (vi)]} \\
&= E(Y^2) + b_0^2 + E\big((\mathbf{b}'\boldsymbol{x})^2\big) + 2b_0 \cdot E(\mathbf{b}'\boldsymbol{x}) - 2b_0 E(Y) - 2E(\mathbf{b}'\boldsymbol{x} \cdot Y) && \text{[Box 6.1 (i), (iii)]} \\
&= E(Y^2) + b_0^2 + E\big((\mathbf{b}'\boldsymbol{x})^2\big) + 2b_0 \cdot \mathbf{b}' E(\boldsymbol{x}) - 2b_0 E(Y) - 2\mathbf{b}' E(\boldsymbol{x} \cdot Y) && \text{[Box 7.2 (iii)]}
\end{aligned}$$

with respect to b_0 is

$$\frac{\partial MSE(b_0, b_1, \ldots, b_n)}{\partial b_0} \;=\; 2b_0 + 2\mathbf{b}' E(\boldsymbol{x}) - 2E(Y).$$

If $f(X_1, \ldots, X_n) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x} = Q_{lin}(Y \mid X_1, \ldots, X_n)$, then

$$2\beta_0 + 2\boldsymbol{\beta}' E(\boldsymbol{x}) - 2E(Y) \;=\; 0.$$

Dividing both sides by 2, and solving for β_0 yields

$$\beta_0 \;=\; E(Y) - \boldsymbol{\beta}' E(\boldsymbol{x}).$$

Gathering the first partial derivatives of $MSE(b_0, b_1, \ldots, b_n)$ with respect to $b_1, \ldots, b_n$ in a vector yields

$$2E(\boldsymbol{x}\boldsymbol{x}')\,\mathbf{b} + 2b_0 E(\boldsymbol{x}) - 2E(\boldsymbol{x} \cdot Y).$$

If $f(X_1, \ldots, X_n) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x} = Q_{lin}(Y \mid X_1, \ldots, X_n)$, then $2E(\boldsymbol{x}\boldsymbol{x}')\,\boldsymbol{\beta} + 2\beta_0 E(\boldsymbol{x}) - 2E(\boldsymbol{x} \cdot Y) = \mathbf{0}$, and dividing both sides by 2 yields

$$E(\boldsymbol{x}\boldsymbol{x}')\,\boldsymbol{\beta} + \beta_0 E(\boldsymbol{x}) - E(\boldsymbol{x} \cdot Y) \;=\; \mathbf{0}.$$

Inserting our result $\beta_0 = E(Y) - \boldsymbol{\beta}' E(\boldsymbol{x})$, using $\boldsymbol{\Sigma}_{\boldsymbol{x}y} = E(\boldsymbol{x} \cdot Y) - E(\boldsymbol{x}) \cdot E(Y)$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} = E(\boldsymbol{x}\,\boldsymbol{x}') - E(\boldsymbol{x})\,E(\boldsymbol{x}')$ [see Box 7.3 (i)] yields

$$\begin{aligned}
&E(\boldsymbol{x}\boldsymbol{x}')\,\boldsymbol{\beta} + \big(E(Y) - \boldsymbol{\beta}' E(\boldsymbol{x})\big) E(\boldsymbol{x}) - E(\boldsymbol{x} \cdot Y) \\
&= E(\boldsymbol{x}\boldsymbol{x}')\,\boldsymbol{\beta} + E(Y) \cdot E(\boldsymbol{x}) - E(\boldsymbol{x}) E(\boldsymbol{x}')\,\boldsymbol{\beta} - E(\boldsymbol{x} \cdot Y) \\
&= \big(E(\boldsymbol{x}\boldsymbol{x}') - E(\boldsymbol{x}) E(\boldsymbol{x}')\big)\,\boldsymbol{\beta} - \big(E(\boldsymbol{x} \cdot Y) - E(\boldsymbol{x}) \cdot E(Y)\big) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\,\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\boldsymbol{x}y} \;=\; \mathbf{0}.
\end{aligned}$$

Adding $\boldsymbol{\Sigma}_{\boldsymbol{x}y}$ on both sides yields $\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\,\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}y}$, and multiplying both sides from the left by $\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1}$ we receive

$$\boldsymbol{\beta} \;=\; \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\,\boldsymbol{\Sigma}_{\boldsymbol{x}y}.$$

(ii) $\Rightarrow$ (i).

$$\begin{aligned}
E(\epsilon) \;&=\; E\big(Y - f(X)\big) \\
&=\; E\big(Y - (\beta_0 + \boldsymbol{\beta}'\boldsymbol{x})\big) && \text{[def. of } f(X)\text{]} \\
&=\; E\big(Y - (E(Y) - \boldsymbol{\beta}' E(\boldsymbol{x}) + \boldsymbol{\beta}'\boldsymbol{x})\big) && \text{[(ii)]} \\
&=\; E(Y) - E(Y) + \boldsymbol{\beta}' E(\boldsymbol{x}) - \boldsymbol{\beta}' E(\boldsymbol{x}) && \text{[Box 6.1 (vi), (i), Box 7.2 (iii)]} \\
&=\; 0. \\
\boldsymbol{\Sigma}_{\boldsymbol{x}\epsilon} \;&=\; \boldsymbol{\Sigma}_{\boldsymbol{x}, Y - (\beta_0 + \boldsymbol{\beta}'\boldsymbol{x})} && \text{[def. of } \epsilon\text{]} \\
&=\; \boldsymbol{\Sigma}_{\boldsymbol{x}y} - \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\,\boldsymbol{\beta} && \text{[Box 7.3 (ii), (iii)]} \\
&=\; \boldsymbol{\Sigma}_{\boldsymbol{x}y} - \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\,\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\,\boldsymbol{\Sigma}_{\boldsymbol{x}y} && \text{[(ii)]} \\
&=\; \mathbf{0}.
\end{aligned}$$

(i) $\Rightarrow$ (iii). Let $f(X) := \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}$, $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^n$ such that $E(\epsilon) = 0$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}\epsilon} = \mathbf{0}$, where $\epsilon := Y - f(X)$. Then, for any linear function $h(X) := b_0 + \mathbf{b}'\boldsymbol{x}$, $b_0 \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^n$,

$$
\begin{aligned}
& E\big((Y - f(X))[f(X) - h(X)]\big) \\
&= E\big(\epsilon \cdot [f(X) - h(X)]\big) && \text{[def. of } \epsilon\text{]} \\
&= E\big(\epsilon \cdot [(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}) - (b_0 + \mathbf{b}'\boldsymbol{x})]\big) && \text{[defs. of } f(X), h(X)\text{]} \\
&= E\big(\epsilon \cdot \big[(\beta_0 - b_0) + (\boldsymbol{\beta} - \mathbf{b})'\boldsymbol{x}\big]\big) \\
&= (\beta_0 - b_0) \cdot E(\epsilon) + (\boldsymbol{\beta} - \mathbf{b})' E(\epsilon \cdot \boldsymbol{x}) && \text{[Box 6.1 (vi), Box 7.3 (iii)]} \\
&= (\boldsymbol{\beta} - \mathbf{b})' \boldsymbol{\Sigma}_{\boldsymbol{x}\epsilon} && \text{[(i), Box 7.3 (i)]} \\
&= 0. && \text{[(i)]}
\end{aligned}
$$

Using this result and considering

$$
\begin{aligned}
E\big([Y - h(X)]^2\big) &= E\big[\big([Y - f(X)] + [f(X) - h(X)]\big)^2\big] \\
&= E\big([Y - f(X)]^2\big) + E\big([f(X) - h(X)]^2\big) + 2 \cdot E\big([Y - f(X)][f(X) - h(X)]\big) \\
&= E\big([Y - f(X)]^2\big) + E\big([f(X) - h(X)]^2\big) \\
&\geq E\big([Y - f(X)]^2\big).
\end{aligned}
$$

Hence, $f(X) = Q_{lin}(Y|X)$.

7.7 Exercises

▷ **Exercise 7-1** Show that $Corr(X,Y) = E\left(\dfrac{X - E(X)}{SD(X)} \cdot \dfrac{Y - E(Y)}{SD(Y)}\right)$.

▷ **Exercise 7-2** Prove the proposition of Remark 7.12.

▷ **Exercise 7-3** Use Definition 7.2 in order to determine the coefficients α_0 and α_1 of the linear quasi-regression in Example 7.4.

▷ **Exercise 7-4** Consider the linear quasi-regression f with $f(x) = \alpha_0 + \alpha_1 x$, $x \in \mathbb{R}$. Prove: If $x_1, x_2 \in \mathbb{R}$, then $\alpha_0 = f(0)$ and $\alpha_1 = \frac{1}{x_2 - x_1}[f(x_2) - f(x_1)]$.

▷ **Exercise 7-5** Prove the propositions of Box 7.1. Note that this also proves the rules of computation summarized in Box 6.2, because $Var(X) = Cov(X,X)$.

▷ **Exercise 7-6** Prove the proposition of Remark 7.20.

▷ **Exercise 7-7** Show that $Corr(\alpha_0 + \alpha_1 X,\ \beta_0 + \beta_1 Y) = Corr(X,Y)$, where $\alpha_0, \alpha_1, \beta_0, \alpha_1 \in \mathbb{R}$ and $\alpha_1, \beta_1 > 0$.

▷ **Exercise 7-8** Show

$$
Cov(X,Y) = \begin{cases} SD(X) \cdot SD(Y), & \text{if } \alpha_1 > 0 \\ -SD(X) \cdot SD(Y), & \text{if } \alpha_1 < 0, \end{cases} \tag{7.41}
$$

provided that there are $\alpha_0, \alpha_1 \in \mathbb{R}$ with $Y \underset{P}{=} \alpha_0 + \alpha_1 X$, $\alpha_1 \neq 0$.

▷ **Exercise 7-9** Prove Equation (7.23) and that the slope is invariant under translations $c + X$, $d + Y$, $c, d \in \mathbb{R}$.

▷ **Exercise 7-10** Prove the rules of computation of Box 7.2.

▷ **Exercise 7-11** Prove the rules of computation of Box 7.3.

Solutions

▷ **Solution 7-1**

$$\begin{aligned} Corr(X,Y) &= \frac{Cov(X,Y)}{SD(X)\cdot SD(Y)} \\ &= \frac{E\big([X-E(X)]\cdot[Y-E(Y)]\big)}{SD(X)\cdot SD(Y)} && \text{[def. of } Cov(X,Y)\text{]} \\ &= E\left(\frac{[X-E(X)]\cdot[Y-E(Y)]}{SD(X)\cdot SD(Y)}\right) && \text{[Box 6.1 (iii)]} \\ &= E\left(\frac{X-E(X)}{SD(X)}\cdot\frac{Y-E(Y)}{SD(Y)}\right). \end{aligned}$$

▷ **Solution 7-2** If $X, Y : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ are random variables, then

$$(\alpha X + \beta Y)^2 = \alpha^2 X^2 + \beta^2 Y^2 + 2\alpha\beta X\cdot Y$$

is also a random variable on $(\Omega, \mathscr{A}, P)$ (see Def. 5.1, Example 2.61, and Th. 2.57), and this implies

$$\begin{aligned} E[(\alpha X+\beta Y)^2] &= E(\alpha^2 X^2 + \beta^2 Y^2 + 2\alpha\beta X\cdot Y) \\ &= \alpha^2 E(X^2) + \beta^2 E(Y^2) + 2\alpha\beta E(X\cdot Y). && \text{[Box 6.1 (vii)]} \end{aligned}$$

The terms $E(X^2)$ and $E(Y^2)$ are finite by assumption and, according to Remark 7.1, $E(X\cdot Y)$ is finite as well. This implies $E[(\alpha X+\beta Y)^2] < \infty$.

▷ **Solution 7-3**

$$\begin{aligned} MSE(a_0,a_1) &= E\big([Y-(a_0+a_1X)]^2\big) && \text{[(7.36)]} \\ &= \frac{1}{4}\cdot(1-a_0-a_1)^2 + \frac{1}{2}\cdot(2-a_0-2a_1)^2 + \frac{1}{4}\cdot(1-a_0-3a_1)^2 && \text{[(6.3)]} \\ &= \frac{1}{4}\cdot(1+a_0^2+a_1^2-2a_0-2a_1+2a_0a_1) \\ &\quad + \frac{1}{2}\cdot(4+a_0^2+4a_1^2-4a_0-8a_1+4a_0a_1) \\ &\quad + \frac{1}{4}\cdot(1+a_0^2+9a_1^2-2a_0-6a_1+6a_0a_1) \\ &= \frac{5}{2} + a_0^2 + \frac{9}{2}\cdot a_1^2 - 3a_0 - 6a_1 + 4a_0a_1. \end{aligned}$$

The partial derivatives are

$$\frac{\partial MSE(a_0,a_1)}{\partial a_0} = 2a_0 - 3 + 4a_1 \qquad \text{and} \qquad \frac{\partial MSE(a_0,a_1)}{\partial a_1} = 9a_1 - 6 + 4a_0.$$

Fixing the partial derivatives to 0 and denoting the solutions by α_0 and α_1 yields

$$2\alpha_0 - 3 + 4\alpha_1 = 0 \qquad \text{and} \qquad 9\alpha_1 - 6 + 4\alpha_0 = 0.$$

The first equation implies $\alpha_0 = -2\alpha_1 + \frac{3}{2}$. Inserting this result into the second equation yields $9\alpha_1 - 6 - 8\alpha_1 + 6 = 0$, which implies $\alpha_1 = 0$ and $\alpha_0 = \frac{3}{2}$.

▷ **Solution 7-4** The equation $f(x) = \alpha_0 + \alpha_1 x$, $x \in \mathbb{R}$, yields $f(0) = \alpha_0$,

$$f(x_1) \;=\; \alpha_0 + \alpha_1 x_1, \qquad \text{and} \qquad f(x_2) \;=\; \alpha_0 + \alpha_1 x_2.$$

Hence,

$$f(x_2) - f(x_1) \;=\; \alpha_0 + \alpha_1 x_2 - (\alpha_0 + \alpha_1 x_1) \;=\; \alpha_1 (x_2 - x_1).$$

Multiplying both sides by $x_2 - x_1$ yields $\alpha_1 = \frac{1}{x_2 - x_1}\,[f(x_2) - f(x_1)]$.

▷ **Solution 7-5** (i)

$$\begin{aligned}
Cov(X,Y) \;&=\; E\big([X - E(X)] \cdot [Y - E(Y)]\big) && [(7.7)]\\
&=\; E[X \cdot Y - X \cdot E(Y) - E(X) \cdot Y + E(X) \cdot E(Y)]\\
&=\; E(X \cdot Y) \;-\; E[X \cdot E(Y)] \;-\; E[E(X) \cdot Y] \;+\; E(X) \cdot E(Y) && [\text{Box 6.1 (vi), (i)}]\\
&=\; E(X \cdot Y) \;-\; 2E(X) \cdot E(Y) \;+\; E(X) \cdot E(Y) && [\text{Box 6.1 (iii)}]\\
&=\; E(X \cdot Y) \;-\; E(X) \cdot E(Y).
\end{aligned}$$

(ii)

$$\begin{aligned}
Cov(\alpha + X, \beta + Y) \;&=\; E\big([\alpha + X - E(\alpha + X)] \cdot [\beta + Y - E(\beta + Y)]\big) && [(7.7)]\\
&=\; E\big([\alpha + X - \alpha - E(X)] \cdot [\beta + Y - \beta - E(Y)]\big) && [\text{Box 6.1 (ii)}]\\
&=\; E\big([X - E(X)] \cdot [Y - E(Y)]\big)\\
&=\; Cov(X,Y). && [(7.7)]
\end{aligned}$$

(iii)

$$\begin{aligned}
Cov(\alpha X, \beta Y) \;&=\; E(\alpha X \cdot \beta Y) - E(\alpha X) \cdot E(\beta Y) && [\text{Box 7.1 (i)}]\\
&=\; \alpha\beta E(X \cdot Y) - \alpha\beta E(X) \cdot E(Y) && [\text{Box 6.1 (iii)}]\\
&=\; \alpha\beta\,[E(X \cdot Y) - E(X) \cdot E(Y)]\\
&=\; \alpha\beta\, Cov(X,Y). && [\text{Box 7.1 (i)}]
\end{aligned}$$

(iv) This rule immediately follows from Equations (6.27) and (7.7).

(v) This rule immediately follows from Equation (7.7).

(vi) Independence of X and Y implies $E(X \cdot Y) = E(X) \cdot E(Y)$ (see Th. 6.24) and $Cov(X,Y) = 0$ [see Rule (i)].

(vii) According to Lemma 5.47, X and Y are independent if $X \underset{P}{=} \alpha$. This implies that $Cov(X,Y) = 0$ [see Rule (i) of Box 7.1].

(viii)

$$\begin{aligned}
&Var\Big(\sum_{i=1}^{n} \alpha_i Y_i\Big)\\
&= E\left(\left[\sum_{i=1}^{n} \alpha_i Y_i - E\left(\sum_{i=1}^{n} \alpha_i Y_i\right)\right]^2\right) && [(6.27)]\\
&= E\left[\left(\sum_{i=1}^{n} \alpha_i Y_i - \alpha_i \sum_{i=1}^{n} E(Y_i)\right)^2\right] && [\text{Box 6.1 (vii)}]\\
&= E\left[\sum_{i=1}^{n} \alpha_i \,(Y_i - E(Y_i))^2\right]
\end{aligned}$$

$$= E\left[\sum_{i=1}^{n} \alpha_i^2 \big(Y_i - E(Y_i)\big)^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \alpha_i \alpha_j \big(Y_i - E(Y_i)\big)\big(Y_j - E(Y_j)\big)\right]$$

$$= \sum_{i=1}^{n} \alpha_i^2 E(Y_i - E(Y_i))^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \alpha_i \alpha_j E\big[\big(Y_i - E(Y_i)\big)\big(Y_j - E(Y_j)\big)\big] \quad \text{[Box 6.1 (vii)]}$$

$$= \sum_{i=1}^{n} \alpha_i^2 \, Var(Y_i) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \alpha_i \alpha_j Cov(Y_i, Y_j). \quad \text{[(6.27), (7.7)]}$$

(ix)

$$Cov\Big(\sum_{i=1}^{n} \alpha_i X_i, \sum_{j=1}^{m} \beta_j Y_j\Big)$$

$$= E\left[\left(\sum_{i=1}^{n} \alpha_i X_i - E\Big(\sum_{i=1}^{n} \alpha_i X_i\Big)\right)\left(\sum_{j=1}^{m} \beta_i Y_j - E\Big(\sum_{j=1}^{m} \beta_j Y_j\Big)\right)\right] \quad \text{[(7.7)]}$$

$$= E\left[\left(\sum_{i=1}^{n} \alpha_i \big[X_i - E(X_i)\big]\right)\left(\sum_{j=1}^{m} \beta_j \big[Y_j - E(Y_j)\big]\right)\right]$$

$$= E\left[\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \big[X_i - E(X_i)\big]\big[Y_j - E(Y_j)\big]\right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j Cov(X_i, Y_j). \quad \text{[Box 6.1 (vii), (7.7)]}$$

(x) If $X_1 \underset{P}{=} X_2$, then $[Y - E(Y)] \cdot [X_1 - E(X_1)] \underset{P}{=} [Y - E(Y)] \cdot [X_2 - E(X_2)]$. According to Corollary 5.20, these two product variables have the same distribution and according to Corollary 6.17 the same expectation. However, the expectations of these product variables are the covariances. Hence, $Cov(Y, X_1) = Cov(Y, X_2)$.

▷ **Solution 7-6** If $SD(X), SD(Y) > 0$, then

$$\begin{aligned} Corr(X,Y) &= \frac{Cov(X,Y)}{SD(X) \cdot SD(Y)} && \text{[(7.17)]} \\ &= \frac{E\big([X - E(X)] \cdot [Y - E(Y)]\big)}{SD(X) \cdot SD(Y)} && \text{[(7.7)]} \\ &= E\left(\frac{[X - E(X)] \cdot [Y - E(Y)]}{SD(X) \cdot SD(Y)}\right) && \text{[Box 6.1 (iii)]} \\ &= E\left(\frac{X - E(X)}{SD(X)} \cdot \frac{Y - E(Y)}{SD(Y)}\right). \end{aligned}$$

▷ **Solution 7-7** According to Rules (ii) and (iii) of Box 7.1, $Cov(\alpha_0 + \alpha_1 X, \beta_0 + \beta_1 Y) = \alpha_1 \beta_1 Cov(X, Y)$. Similarly, according to Rules (iv) and (ii) of Box 6.2, $Var(\alpha_0 + \alpha_1 X) = \alpha_1^2 Var(X)$, which implies $SD(\alpha_0 + \alpha_1 X) = \alpha_1 SD(X)$. Hence,

$$\begin{aligned} Corr(\alpha_0 + \alpha_1 X, \beta_0 + \beta_1 Y) &= \frac{Cov(\alpha_0 + \alpha_1 X, \beta_0 + \beta_1 Y)}{SD(\alpha_0 + \alpha_1 X) \cdot SD(\beta_0 + \beta_1 Y)} \\ &= \frac{\alpha_1 \beta_1 Cov(X,Y)}{\alpha_1 SD(X) \cdot \beta_1 SD(Y)} \\ &= \frac{Cov(X,Y)}{SD(X) \cdot SD(Y)} \\ &= Corr(X,Y). \end{aligned}$$

▷ **Solution 7-8** According to Box 7.1 (x), $Cov(X,Y) = Cov(X,\alpha_0+\alpha_1 X)$ if $Y \underset{P}{=} \alpha_0+\alpha_1 X$. Box 6.2 (v), (ii), (iii) imply $Var(Y) = \alpha_1^2 Var(X)$. Therefore,

$$Corr(X,Y) = \frac{Cov(X,\alpha_0+\alpha_1 X)}{SD(X)\cdot\sqrt{\alpha_1^2 Var(X)}} = \frac{\alpha_1 Var(X)}{SD(X)\cdot|\alpha_1|\cdot SD(X)} = \frac{\alpha_1}{|\alpha_1|}.$$

Hence, $Corr(X,Y) = 1$ if $\alpha_1 > 0$ and $Corr(X,Y) = -1$ if $\alpha_1 < 0$. Therefore,

$$Corr(X,Y) := \frac{Cov(X,Y)}{SD(X)\cdot SD(Y)}$$

yields (7.41).

▷ **Solution 7-9**

$$\begin{aligned}
\alpha_1^* &= \frac{Cov(aX,bY)}{Var(aX)} && \text{[Th. 7.13 (ii)]}\\
&= \frac{ab\cdot Cov(X,Y)}{a^2\cdot Var(X)} && \text{[Box 7.1 (iii), Box 6.2 (iii)]}\\
&= \frac{b}{a}\cdot\alpha_1. && \text{[Th. 7.13 (ii)]}
\end{aligned}$$

According to Theorem 7.13 (ii), the slope of the linear quasi-regression of $c+Y$ on $d+X$, is

$$\begin{aligned}
\frac{Cov(c+X,d+Y)}{Var(c+X)} &= \frac{Cov(X,Y)}{Var(X)} && \text{[Box 7.1 (ii), Box 6.2 (ii)]}\\
&= \alpha_1. && \text{[Th. 7.13 (ii)]}
\end{aligned}$$

▷ **Solution 7-10** (i). Equation (7.25) and Rule (i) of Box 6.1 imply

$$E(\boldsymbol{x}) = \big[E(X_1),\ldots,E(X_n)\big]' = \big[a_1,\ldots,a_n\big]' = \mathbf{a}.$$

(ii). Equation (7.25) and Rule (ii) of Box 6.1 imply

$$E(\mathbf{a}+\boldsymbol{x}) = \begin{bmatrix} E(a_1+X_1)\\ \vdots\\ E(a_n+X_n)\end{bmatrix} = \begin{bmatrix} a_1+E(X_1)\\ \vdots\\ a_n+E(X_n)\end{bmatrix} = \mathbf{a}+E(\boldsymbol{x}).$$

(iii). Equation (7.29) and Rule (vi) of Box 6.1 imply

$$E(\mathbf{a}'\boldsymbol{x}) = E\Big(\sum_{i=1}^n a_i X_i\Big) = \sum_{i=1}^n a_i\cdot E(X_i) = \mathbf{a}'E(\boldsymbol{x}).$$

The other equations summarized in (iii) follow from the fact that $\mathbf{a}'\boldsymbol{x}$ is a one-dimensional random variable (see Example 2.61) and $\mathbf{a}'\boldsymbol{x} = \boldsymbol{x}'\mathbf{a}$.

(v). Let $\mathbf{a}_l'$ and $\mathbf{b}_l'$, $l = 1,\ldots,k$, denote the row vectors of $\mathbf{A}$ and $\mathbf{B}$, respectively. Applying Equation (7.25), Rule (vi) of Box 6.1, and Rule (iii) to the terms $\mathbf{a}_l'\boldsymbol{x}$ and $\mathbf{b}_l'\boldsymbol{y}$, $l = 1,2\ldots,k$, respectively, yields

$$\begin{aligned}
E(\mathbf{A}\boldsymbol{x}+\mathbf{B}\boldsymbol{y}) &= E\begin{bmatrix}\mathbf{a}_1'\boldsymbol{x}+\mathbf{b}_1'\boldsymbol{y}\\ \vdots\\ \mathbf{a}_k'\boldsymbol{x}+\mathbf{b}_k'\boldsymbol{y}\end{bmatrix} = \begin{bmatrix}E(\mathbf{a}_1'\boldsymbol{x}+\mathbf{b}_1'\boldsymbol{y})\\ \vdots\\ E(\mathbf{a}_k'\boldsymbol{x}+\mathbf{b}_k'\boldsymbol{y})\end{bmatrix} = \begin{bmatrix}E(\mathbf{a}_1'\boldsymbol{x})+E(\mathbf{b}_1'\boldsymbol{y})\\ \vdots\\ E(\mathbf{a}_k'\boldsymbol{x})+E(\mathbf{b}_k'\boldsymbol{y})\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{a}_1'E(\boldsymbol{x})+\mathbf{b}_1'E(\boldsymbol{y})\\ \vdots\\ \mathbf{a}_k'E(\boldsymbol{x})+\mathbf{b}_k'E(\boldsymbol{y})\end{bmatrix} = \mathbf{A}E(\boldsymbol{x})+\mathbf{B}E(\boldsymbol{y}).
\end{aligned}$$

(iv). This rule is a special case of Rule (v) with $\mathbf{B} = \mathbf{0}$.

(vi). Let $\mathbf{a}_l$ and $\boldsymbol{x}_l$, $l = 1,\dots,k$, denote the column vectors of $\mathbf{A}'$ and $\boldsymbol{X}$, respectively. Then Equation (7.27) and Rule (i) imply

$$E(\boldsymbol{X}) = \big[E(\boldsymbol{x}_1),\dots,E(\boldsymbol{x}_k)\big] = \big[E(\mathbf{a}_1),\dots,E(\mathbf{a}_k)\big] = \big[\mathbf{a}_1,\dots,\mathbf{a}_k\big] = \mathbf{A}'.$$

(vii). Let $\mathbf{a}_l$ and $\boldsymbol{x}_l$, $l = 1,\dots,k$, denote the column vectors of $\mathbf{A}'$ and $\boldsymbol{X}$, respectively. Then Equation (7.25) and Rule (ii) imply

$$\begin{aligned} E(\mathbf{A}'+\boldsymbol{X}) &= E\big(\mathbf{a}_1+\boldsymbol{x}_1,\dots,\mathbf{a}_k+\boldsymbol{x}_k\big) = \big[E(\mathbf{a}_1+\boldsymbol{x}_1),\dots,E(\mathbf{a}_k+\boldsymbol{x}_k)\big] \\ &= \big[\mathbf{a}_1+E(\boldsymbol{x}_1),\dots,\mathbf{a}_k+E(\boldsymbol{x}_k)\big] = \mathbf{A}'+E(\boldsymbol{X}). \end{aligned}$$

(viii). Let $\mathbf{c}_i'$, $i = 1,\dots,l$, denote the row vectors of $\mathbf{C}$. Then Equation (7.27) and Rule (iii) imply

$$E(\mathbf{C}\boldsymbol{X}) = E\begin{bmatrix}\mathbf{c}_1'\boldsymbol{X}\\ \vdots \\ \mathbf{c}_l'\boldsymbol{X}\end{bmatrix} = \begin{bmatrix}E(\mathbf{c}_1'\boldsymbol{X})\\ \vdots \\ E(\mathbf{c}_l'\boldsymbol{X})\end{bmatrix} = \begin{bmatrix}\mathbf{c}_1'E(\boldsymbol{X})\\ \vdots \\ \mathbf{c}_l'E(\boldsymbol{X})\end{bmatrix} = \mathbf{C}E(\boldsymbol{X}).$$

(ix). Rule (viii), Equation (7.28), and the rules for the transpose of a matrix yield

$$\begin{aligned} E(\mathbf{C}\boldsymbol{X}\boldsymbol{Y}'\mathbf{D}') &= \mathbf{C}\,E(\boldsymbol{X}\boldsymbol{Y}'\mathbf{D}') && [(\text{viii})] \\ &= \mathbf{C}\,\big(E(\mathbf{D}\boldsymbol{Y}\boldsymbol{X}')\big)' && [(7.28)] \\ &= \mathbf{C}\,\big(\mathbf{D}E(\boldsymbol{Y}\boldsymbol{X}')\big)' && [(\text{viii})] \\ &= \mathbf{C}\,E(\boldsymbol{Y}\boldsymbol{X}')'\mathbf{D}' && [(7.28)] \\ &= \mathbf{C}E(\boldsymbol{X}\boldsymbol{Y}')\mathbf{D}'. && [(7.28)] \end{aligned}$$

▷ **Solution 7-11** (i).

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}} &= E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\big) && [(7.31)] \\ &= E\big(\boldsymbol{x}\,\boldsymbol{y}' - \boldsymbol{x}\,E(\boldsymbol{y})' - E(\boldsymbol{x})\,\boldsymbol{y}' + E(\boldsymbol{x})\,E(\boldsymbol{y})'\big) \\ &= E(\boldsymbol{x}\,\boldsymbol{y}') - E(\boldsymbol{x})\,E(\boldsymbol{y})' - E(\boldsymbol{x})\,E(\boldsymbol{y}') + E(\boldsymbol{x})\,E(\boldsymbol{y})' && [\text{Box 7.2 (iii)}] \\ &= E(\boldsymbol{x}\,\boldsymbol{y}') - E(\boldsymbol{x})\,E(\boldsymbol{y}'). \end{aligned}$$

(ii).

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{a}+\boldsymbol{x},\mathbf{b}+\boldsymbol{y}} &= E\big([\mathbf{a}+\boldsymbol{x}-E(\mathbf{a}+\boldsymbol{x})]\,[\mathbf{b}+\boldsymbol{y}-E(\mathbf{b}+\boldsymbol{y})]'\big) && [(7.31)] \\ &= E\big([\mathbf{a}+\boldsymbol{x}-\mathbf{a}+E(\boldsymbol{x})]\,[\mathbf{b}+\boldsymbol{y}-\mathbf{b}+E(\boldsymbol{y})]'\big) && [\text{Box 7.2 (ii)}] \\ &= E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\big) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}. && [(7.31)] \end{aligned}$$

(iii).

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{A}\boldsymbol{x},\mathbf{B}\boldsymbol{y}} &= E\big([\mathbf{A}\boldsymbol{x}-E(\mathbf{A}\boldsymbol{x})]\,[\mathbf{B}\boldsymbol{y}-E(\mathbf{B}\boldsymbol{y})]'\big) && [(7.31)] \\ &= E\big([\mathbf{A}\boldsymbol{x}-\mathbf{A}E(\boldsymbol{x})]\,[\mathbf{B}\boldsymbol{y}-\mathbf{B}E(\boldsymbol{y})]'\big) && [\text{Box 7.2 (iv)}] \\ &= E\big[\mathbf{A}[\boldsymbol{x}-E(\boldsymbol{x})]\,\big(\mathbf{B}\,[\boldsymbol{y}-E(\boldsymbol{y})]\big)'\big] \\ &= E\big(\mathbf{A}[\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\mathbf{B}'\big) \\ &= \mathbf{A}E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\big)\mathbf{B}' && [\text{Box 7.2 (ix)}] \\ &= \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}\,\mathbf{B}'. && [(7.31)] \end{aligned}$$

(iv).

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{xy}} &= E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\big) && [(7.31)]\\
&= \Big(E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\big)'\Big)' && [(\mathbf{A}')'=\mathbf{A}]\\
&= E\big(\big[[\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{y}-E(\boldsymbol{y})]'\big]'\big)' && [(7.28)]\\
&= E\big([\boldsymbol{y}-E(\boldsymbol{y})]\,[\boldsymbol{x}-E(\boldsymbol{x})]'\big)' && [(\mathbf{ab}')'=\mathbf{ba}']\\
&= \boldsymbol{\Sigma}'_{\boldsymbol{yx}}. && [(7.31)]
\end{aligned}$$

(v). Independence of the multivariate random variables X and Y implies $X_i \underset{P}{\perp\!\!\!\perp} Y_j$, for all $i = 1,\ldots,n$ and $j = 1,\ldots,m$. Therefore, Rule (vi) of Box 7.1 implies $Cov(X_i, Y_j) = 0$, for all $i = 1,\ldots,n$ and $j = 1,\ldots,m$. Equation (7.32) then implies $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \mathbf{0}$.

(vi). If $\boldsymbol{x} \underset{P}{=} \mathbf{a}$, then Rule (vii) of Box (7.1) yields $Cov(X_i, Y_j) = 0$, for all $i = 1,\ldots,n$, $j = 1,\ldots,m$. Equation (7.32) then implies $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \mathbf{0}$.

(vii).

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{A}\boldsymbol{x}+\mathbf{B}\boldsymbol{y},\,\mathbf{C}\boldsymbol{w}+\mathbf{D}\boldsymbol{z}}\\
&= E\big([\mathbf{A}\boldsymbol{x}+\mathbf{B}\,\boldsymbol{y}-E(\mathbf{A}\,\boldsymbol{x}+\mathbf{B}\,\boldsymbol{y})]\,[\mathbf{C}\,\boldsymbol{w}+\mathbf{D}\,\boldsymbol{z}-E(\mathbf{C}\,\boldsymbol{w}+\mathbf{D}\,\boldsymbol{z})]'\big) && [(7.31)]\\
&= E\big([\mathbf{A}\boldsymbol{x}-E(\mathbf{A}\,\boldsymbol{x})+\mathbf{B}\,\boldsymbol{y}-E(\mathbf{B}\,\boldsymbol{y})]\,[\mathbf{C}\,\boldsymbol{w}-E(\mathbf{C}\,\boldsymbol{w})+\mathbf{D}\,\boldsymbol{z}-E(\mathbf{D}\,\boldsymbol{z})]'\big)\\
&= E\big([\mathbf{A}\boldsymbol{x}-\mathbf{A}E(\boldsymbol{x})]\,[\mathbf{C}\,\boldsymbol{w}-\mathbf{C}\,E(\boldsymbol{w})]' + [\mathbf{A}\boldsymbol{x}-\mathbf{A}E(\boldsymbol{x})]\,[\mathbf{D}\,\boldsymbol{z}-\mathbf{D}\,E(\boldsymbol{z})]'\\
&\quad + [\mathbf{B}\,\boldsymbol{y}-\mathbf{B}\,E(\boldsymbol{y})]\,[\mathbf{C}\,\boldsymbol{w}-\mathbf{C}\,E(\boldsymbol{w})]' + [\mathbf{B}\,\boldsymbol{y}-\mathbf{B}\,E(\boldsymbol{y})]\,[\mathbf{D}\,\boldsymbol{z}-\mathbf{D}\,E(\boldsymbol{z})]'\big) && [\text{Box 7.2 (iv)}]\\
&= E\big(\mathbf{A}\,[\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{w}-E(\boldsymbol{w})]'\mathbf{C}'\big) + E\big(\mathbf{A}\,[\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{z}-E(\boldsymbol{z})]'\mathbf{D}'\big)\\
&\quad + E\big(\mathbf{B}\,[\boldsymbol{y}-E(\boldsymbol{y})]\,[\boldsymbol{w}-E(\boldsymbol{w})]'\mathbf{C}'\big) + E\big(\mathbf{B}\,[\boldsymbol{y}-E(\boldsymbol{y})]\,[\boldsymbol{z}-E(\boldsymbol{z})]'\mathbf{D}'\big)\\
&= \mathbf{A}E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{w}-E(\boldsymbol{w})]'\big)\mathbf{C}' + \mathbf{A}E\big([\boldsymbol{x}-E(\boldsymbol{x})]\,[\boldsymbol{z}-E(\boldsymbol{z})]'\big)\mathbf{D}'\\
&\quad + \mathbf{B}E\big([\boldsymbol{y}-E(\boldsymbol{y})]\,[\boldsymbol{w}-E(\boldsymbol{w})]'\big)\mathbf{C}' + \mathbf{B}E\big([\boldsymbol{y}-E(\boldsymbol{y})]\,[\boldsymbol{z}-E(\boldsymbol{z})]'\big)\mathbf{D}' && [\text{Box 7.2 (ix)}]\\
&= \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{xw}}\,\mathbf{C}' + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{xz}}\,\mathbf{D}' + \mathbf{B}\,\boldsymbol{\Sigma}_{\boldsymbol{yw}}\,\mathbf{C}' + \mathbf{B}\,\boldsymbol{\Sigma}_{\boldsymbol{yz}}\mathbf{D}'. && [(7.31)]
\end{aligned}$$

(viii). If $\boldsymbol{x} \underset{P}{=} \boldsymbol{z}$, then Rule (x) of Box 7.1 implies that $Cov(X_i, Y_j) = Cov(Z_i, Y_j)$, for all $i = 1,\ldots,n$, $j = 1,\ldots,m$. Equation (7.32) then implies $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \boldsymbol{\Sigma}_{\boldsymbol{zy}}$.

Part III
Conditional Expectation and Regression

Chapter 9
Conditional Expectation Value and Discrete Conditional Expectation

In chapter 6 we introduced the concepts covariance and correlation, which quantify the strength of the kind of dependence that can be described by a linear quasi-regression. In the next four chapters we introduce the concept of a *conditional expectation* and a 'genuine' regression. These concepts can be used to describe how the $(X{=}x)$-conditional expectation values of a numerical random variable Y depend on the values of a (numerical, non-numerical, multivariate) random variable X. In this chapter we start with the concepts *$(X{=}x)$-conditional expectation value* and *discrete conditional expectation*, presuming that X is a discrete random variable. In this case, the conditional expectation $E(Y|X)$ is easily defined as that random variable whose values are the conditional expectation values $E(Y|X{=}x)$. In chapter 10 we introduce the general concept of a conditional expectation, dropping the assumption that X is discrete. Chapters 12 and 13 deal with parametrizations of a conditional expectation, and Chapter 11 is devoted to the concepts *residual with respect to a conditional expectation*, *conditional variance*, *conditional covariance*, and partial correlation.

9.1 Conditional Expectation Value

Remember, the expectation of a numerical random variable $Y: (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is defined by $E(Y) = \int Y\, dP$, using the probability measure P. Now we choose an event $B \in \mathscr{A}$ with $P(B) > 0$ and, instead of P, we consider the *B-conditional-probability measure* $P^B: \mathscr{A} \to [0,1]$ defined by

$$P^B(A) \;:=\; P(A|B), \quad \forall\, A \in \mathscr{A}, \tag{9.1}$$

(see Def. 4.24). Referring to this measure,

$$E^B(Y) \;:=\; \int Y\, d\, P^B, \tag{9.2}$$

defines the *P^B-expectation of Y*, i. e., the expectation of Y with respect to the measure P^B. Reading the following definition, also remember that

$$\{X{=}x\} \;=\; X^{-1}(\{x\}) \;=\; \{\omega \in \Omega\colon X(\omega) = x\}$$

denotes the event that X takes on the value x and that we defined $P(X{=}x) := P(\{X{=}x\})$ (see Rem. 5.4). Assuming $P(X{=}x) > 0$ and $A \in \mathscr{A}$, we define

$$P(A|X{=}x) \;:=\; P(A|\{X{=}x\}) \tag{9.3}$$

and, analogously to P^B, we define the $(X{=}x)$-*conditional probability measure* $P^{X=x}\colon \mathscr{A} \to [0,1]$ by

$$P^{X=x}(A) \;:=\; P(A|X{=}x), \quad \forall\, A \in \mathscr{A}. \tag{9.4}$$

Remark 9.1 (A First Property of $P^{X=x}$) If $X\colon (\Omega,\mathscr{A},P) \to (\Omega'_X,\mathscr{A}'_X)$ is a random variable, $x \in \Omega'_X$, and $P(X{=}x) > 0$, then $f(X) \underset{P^{X=x}}{=} f(x)$ (see Exercise 9-1). ◁

Definition 9.2 (Conditional Expectation Value)
Let $Y\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a random variable.

(i) If $B \in \mathscr{A}$ with $P(B) > 0$ and Y is quasi-integrable with respect to P^B, then we define

$$E(Y|B) \;:=\; E^B(Y) \;=\; \int Y\, dP^B, \tag{9.5}$$

call it the conditional expectation value of Y given the event B (or the B-conditional expectation value of Y) and say that it exists.

(ii) If $X\colon (\Omega,\mathscr{A},P) \to (\Omega'_X,\mathscr{A}'_X)$ is a random variable, $x \in \Omega'_X$ with $\{x\} \in \mathscr{A}'_X$ and $P(X{=}x) > 0$, and Y is quasi-integrable with respect to $P^{X=x}$, then we define

$$E(Y|X{=}x) \;:=\; E(Y|\{X{=}x\}) \tag{9.6}$$

and call it the conditional expectation value of Y given $X{=}x$ (or the $(X{=}x)$-conditional expectation value of Y) and say that it exists.

Note that $E(Y|B)$ can be infinite. The only restriction is $B \in \mathscr{A}$ with $P(B) > 0$ and that Y is quasi-integrable with respect to P^B. Otherwise the integral $\int Y\, dP^B$ is not defined. Note that the random variable X in Definition 9.2 (ii) can be numerical, i. e., $\Omega'_X \subset \overline{\mathbb{R}}$, multivariate (see section 5.3), or nonnumerical.

The following theorem addresses the relationship between the B-conditional expectation value of Y and the expectation of $1_B \cdot Y$ with respect to P.

Theorem 9.3 (B-Conditional Expectation Value and the P-Expectation)
Let the assumptions of Definition 9.2 (i) hold. Then:

(i) $E(Y)$ *exists* $\;\Rightarrow\;$ $E(Y|B)$ *exists.*
(ii) $E(Y) < \infty \;\Rightarrow\; E(Y|B) < \infty$.
(iii) Furthermore,

$$E(Y|B) \;=\; \frac{1}{P(B)} \cdot \int 1_B \cdot Y\, dP \;=\; \frac{1}{P(B)} \cdot E(1_B \cdot Y) \tag{9.7}$$

$$= \frac{1}{P(B)} \cdot Cov(Y, 1_B) + E(Y). \qquad (9.8)$$

(Proof p. 284)

For convenience, we use the following notation:

$$1_{X=x} := 1_{\{X=x\}}, \qquad P(X=x) := P(\{X=x\}), \qquad (9.9)$$

and

$$E^{X=x}(Y) := E^{\{X=x\}}(Y). \qquad (9.10)$$

Using this notation, Equations (9.5) and (9.7) yield the following corollary.

Corollary 9.4 ($(X=x)$-Conditional Expectation Value and P-Expectation) *If the assumptions of Definition 9.2 (ii) hold, then*

$$\begin{aligned} E(Y|X=x) &= E^{X=x}(Y) = \int Y \, dP^{X=x} \\ &= \frac{1}{P(X=x)} \cdot \int 1_{X=x} \cdot Y \, dP = \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot Y). \end{aligned} \qquad (9.11)$$

Remark 9.5 (B-Conditional Probability) If $A \in \mathscr{A}$ and $P(B) > 0$, then

$$\begin{aligned} E(1_A|B) &= \frac{1}{P(B)} \cdot E(1_B \cdot 1_A) && [(9.7)] \\ &= \frac{1}{P(B)} \cdot P(A \cap B) && [\text{Box 6.1 (iv)}] \\ &= P(A|B) = P^B(A). && [(4.2), (9.1)] \end{aligned} \qquad (9.12)$$

Because P^B is a probability measure, these equations imply $0 \le E(1_A|B) \le 1$. ◁

Remark 9.6 ($(X=x)$-Conditional Probability) For $B = \{X=x\}$, Equation (9.12) implies

$$\begin{aligned} P(A|X=x) &= E(1_A|X=x) = \frac{1}{P(X=x)} \cdot E(1_{X=x} \cdot 1_A) \\ &= \frac{1}{P(X=x)} \cdot P(A \cap \{X=x\}) \\ &= P(A|\{X=x\}) = P^{X=x}(A), \end{aligned} \qquad (9.13)$$

provided that $A \in \mathscr{A}$ and $P(X=x) > 0$. The term $P(A|X=x)$ is also called the *conditional probability of A given $X=x$* or the *$(X=x)$-conditional probability of A*. Equations (9.13) show that $E(1_A|X=x) = P(A|X=x)$ is identical to the conditional probability $P(A|\{X=x\})$ of A given the event $\{X=x\}$ (see Def. 4.12). ◁

Remark 9.7 ($(X=x)$-Conditional Probability of $\{Y=y\}$) If $Y\colon (\Omega,\mathscr{A},P)\to(\Omega'_Y,\mathscr{A}'_Y)$ is a random variable and the assumptions of Definition 9.2 (ii) hold, then we also use the notation

$$P(Y=y\,|\,X=x) \;:=\; P(\{Y=y\}\,|\,X=x) \;=\; P(1_{Y=y}=1\,|\,X=x) \;=\; E(1_{Y=y}\,|\,X=x) \tag{9.14}$$

and call it the $(X=x)$-*conditional probability of* $\{Y=y\}$ [see Eqs. (9.9)]. ◁

9.2 Transformation Theorem

If $P_Y^{X=x}\colon \mathscr{A}'_Y\to[0,1]$ denotes the distribution of Y with respect to the $(X=x)$-conditional-probability measure $P^{X=x}$ and $E_Y^{X=x}(g)$ the expectation of g with respect to the distribution $P_Y^{X=x}$, then the transformation theorem (cf. Th. 6.13) for the conditional expectation value $E(Y\,|\,X=x)$ can be formulated as follows:

Theorem 9.8 (Transformation Theorem for $E(Y\,|\,X=x)$)
Let $Y\colon(\Omega,\mathscr{A},P)\to(\Omega'_Y,\mathscr{A}'_Y)$, $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be random variables, and $g\colon(\Omega'_Y,\mathscr{A}'_Y)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a measurable function. Furthermore, let $x\in\Omega'_X$ with $\{x\}\in\mathscr{A}'_X$ and $P(X=x)>0$.

(i) If g is nonnegative or with finite expectation $E_Y^{X=x}(g)$, then

$$\begin{aligned} E_Y^{X=x}(g) \;&=\; \int g\,dP_Y^{X=x} \;=\; \int g(y)\,P_Y^{X=x}(dy) \\ &=\; \int g(Y)\,dP^{X=x} \;=\; E^{X=x}[g(Y)] \;=\; E[g(Y)\,|\,X=x]. \end{aligned} \tag{9.15}$$

(ii) $E_Y^{X=x}(g)$ is finite if and only if $E^{X=x}[g(Y)]$ is finite.

There are two important points in Equations (9.15). *First*, these equations show the relationship between integrals of the composition $g(Y)$ with respect to the conditional-probability measure $P^{X=x}$ on $(\Omega,\mathscr{A})$ on one side, and the distribution $P_Y^{X=x}$ of Y with respect to $P^{X=x}$ on the other side. *Second*, $E[g(Y)\,|\,X=x]$ is identical to the expectation of g with respect to the distribution $P_Y^{X=x}$, i. e., the distribution of Y with respect to the probability measure $P^{X=x}$. Thus, using the distribution $P_{g(Y)}^{X=x}$ of $g(Y)$ with respect to $P^{X=x}$ is not necessary.

Remark 9.9 ($(X=x)$-Conditional Expectation Value of $g(Y)$ and P-Expectation)
Equations (9.15) and Equations (9.11) imply

$$E[g(Y)\,|\,X=x] \;=\; \frac{1}{P(X=x)}\cdot\int 1_{X=x}\cdot g(Y)\;dP \;=\; \frac{1}{P(X=x)}\cdot E[1_{X=x}\cdot g(Y)]. \tag{9.16}$$

◁

Remark 9.10 (A Special Case of the Transformation Theorem) Let $(\Omega'_Y, \mathscr{A}'_Y) = (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ and g be the identity function $id\colon \overline{\mathbb{R}} \to \overline{\mathbb{R}}$, defined by $id(y) = y$ for all $y \in \overline{\mathbb{R}}$, which implies $id(Y) = Y$. If we assume that $x \in \Omega'_X$ with $\{x\} \in \mathscr{A}'_X$ and $P(X{=}x) > 0$, and that Y is nonnegative or with finite expectation $E^{X=x}(Y)$, then Equations (9.15) yield

$$E(Y \,|\, X{=}x) \;=\; E^{X=x}(Y) \;=\; \int Y \; dP^{X=x} \;=\; \int y \; P_Y^{X=x}(dy) \;=\; \int id \; dP_Y^{X=x}. \tag{9.17}$$

◁

Using the notation introduced in Equation (9.14), Equations (9.15), Theorem 9.8 (i), Equations (6.3) and (6.6) imply the following corollary.

Corollary 9.11 (Y Discrete, $g(Y)$ Real-Valued)
Let the assumptions of Theorem 9.8 (i) hold and let $g\colon (\Omega'_Y, \mathscr{A}'_Y) \to (\mathbb{R}, \mathscr{B})$.

(i) If Y is discrete and we assume that there is a finite set $\{y_1, \ldots, y_n\} \in \mathscr{A}'_Y$ with $P_Y(\{y_1, \ldots, y_n\}) = 1$, then

$$E[g(Y) \,|\, X{=}x] \;=\; \sum_{i=1}^{n} g(y_i) \cdot P(Y{=}y_i \,|\, X{=}x). \tag{9.18}$$

(ii) If Y is discrete and we assume that there is a countably infinite set $\{y_1, y_2, \ldots\} \in \mathscr{A}'_Y$ with $P_Y(\{y_1, y_2, \ldots\}) = 1$, then

$$E[g(Y) \,|\, X{=}x] \;=\; \sum_{i=1}^{\infty} g(y_i) \cdot P(Y{=}y_i \,|\, X{=}x). \tag{9.19}$$

Note that, in this corollary, Y does not have to be real-valued or numerical. We only assume that $g(Y)$ is real-valued. In contrast, in the following theorem we have to assume that Y itself is real-valued.

Corollary 9.12 (Discrete and Real-Valued Y)
Let the assumptions of Definition 9.2 (i) hold.

(i) If Y is discrete and there is a finite set $\{y_1, \ldots, y_n\} \in \mathscr{A}'_Y = \mathscr{B}$ of real numbers with $P_Y(\{y_1, \ldots, y_n\}) = 1$. Then

$$E(Y \,|\, X{=}x) \;=\; \sum_{i=1}^{n} y_i \cdot P(Y{=}y_i \,|\, X{=}x). \tag{9.20}$$

(ii) If Y is discrete and there is a countably infinite set $\{y_1, y_2, \ldots\} \in \mathscr{A}'_Y = \mathscr{B}$ of real numbers with $P_Y(\{y_1, y_2, \ldots\}) = 1$, then

$$E(Y \,|\, X{=}x) \;=\; \sum_{i=1}^{\infty} y_i \cdot P(Y{=}y_i \,|\, X{=}x). \tag{9.21}$$

Box 9.1 Rules of Computation for B-Conditional Expectation Values

Let $Y\colon (\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a random variable, let $A,B,C\in\mathscr{A}$ with $P(B)>0$, let the conditional expectation value $E(Y|B)$ exist, and let $\alpha\in\mathbb{R}$. Then:

$$Y \underset{P^B}{=} \alpha \quad\Rightarrow\quad E(Y|B)=\alpha. \tag{i}$$

$$E(\alpha+Y|B) \;=\; \alpha \;+\; E(Y|B)\,. \tag{ii}$$

$$E(\alpha\cdot Y|B) \;=\; \alpha\cdot E(Y|B)\,. \tag{iii}$$

$$E(1_A\cdot 1_C|B) \;=\; P(A\cap C|B). \tag{iv}$$

For $i=1,\dots,n$, let $Y_i\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be random variables with finite B-conditional expectation values $E(Y_i|B)$ and $\alpha_i\in\mathbb{R}$. Then

$$E\Big(\sum_{i=1}^{n}\alpha_i\cdot Y_i\,\Big|\,B\Big) \;=\; \sum_{i=1}^{n}\alpha_i\cdot E(Y_i|B). \tag{v}$$

Let $X,Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be random variables that are nonnegative or with finite B-conditional expectation values. Then:

$$X \underset{P^B}{=} Y \quad\Rightarrow\quad E(X|B) \;=\; E(Y|B)\,. \tag{vi}$$

$$X \underset{P^B}{=} Y \quad\Leftrightarrow\quad \forall A\in\mathscr{A}\colon\; E(1_A X|B) \;=\; E(1_A Y|B). \tag{vii}$$

$$E(X^2),E(Y^2)<\infty,\; X \underset{P^B}{\perp\!\!\!\perp} Y \quad\Rightarrow\quad E(X\cdot Y|B) \;=\; E(X|B)\cdot E(Y|B)\,. \tag{viii}$$

9.3 Other Properties

Because $E(Y|B)$ is defined as the expectation $E^B(Y)$ of Y with respect to the probability measure P^B, all properties of the expectation with respect to P can be translated to $E(Y|B)$, simply replacing P by P^B and $E(Y)$ by $E^B(Y)=E(Y|B)$. Box 9.1 is such a translation of Box 6.1. Note that, according to Theorem 9.3 (i), the expectation $E(Y|B)$ exists if $E(Y)$ exists, provided that $P(B)>0$. Of course, the rules for $E(Y|B)$ also apply to the $(X{=}x)$-conditional expectation value $E(Y|X{=}x)$ [see Def. 9.2 (ii)].

However, there are additional properties dealing with the relationship between the expectation and the conditional expectation value. Some of these have already been formulated in Theorem 9.3. Other additional properties are summarized in Box 9.2 and proved in Exercise 9-2.

Rule (ii) shows how the $(X{=}x)$-conditional expectation values $E(Y|X{=}x)$ can be computed from the conditional expectation values $E(Y|X{=}x,Z{=}z_i)$ and the conditional probabilities $P(Z{=}z_i|X{=}x)$. Hence, considering Equation (9.20) and Rule (ii) in Box 9.2 shows that we have two different equations for computing the

Box 9.2 Rules of Computation for $(X{=}x)$-Conditional Expectation Values

Let $Y\colon (\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a numerical random variable. Furthermore, let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable, and let $x\in\Omega'_X$ with $\{x\}\in\mathscr{A}'_X$ and $P(X{=}x)>0$. If $E(Y|X{=}x)$ exists and $f\colon(\Omega'_X,\mathscr{A}'_X)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ is a measurable function and $E(Y^2)$, $E[f(X)^2]<\infty$, then

$$E[f(X)\cdot Y|X{=}x] \;=\; f(x)\cdot E(Y|X{=}x) \;=\; E[f(x)\cdot Y|X{=}x]. \tag{i}$$

If $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ is a random variable and $z_1,\ldots,z_m\in\Omega'_Z$ such that $P_Z(\{z_1,\ldots,z_m\})=1$ and, for all $i=1,\ldots,m$, $\{z_i\}\in\mathscr{A}'_Z$ and $P(X{=}x,Z{=}z_i)>0$, then

$$E(Y|X{=}x) \;=\; \sum_{i=1}^{m} E(Y|X{=}x,Z{=}z_i)\cdot P(Z{=}z_i|X{=}x). \tag{ii}$$

Correspondingly, if $z_1,z_2,\ldots\in\Omega'_Z$ such that $P_Z(\{z_1,z_2,\ldots\})=1$ and, for all $i=1,2\ldots$, $\{z_i\}\in\mathscr{A}'_Z$ and $P(X{=}x,Z{=}z_i)>0$, then

$$E(Y|X{=}x) \;=\; \sum_{i=1}^{\infty} E(Y|X{=}x,Z{=}z_i)\cdot P(Z{=}z_i|X{=}x). \tag{iii}$$

conditional expectation value $E(Y|X{=}x)$. Finally, note that a special case of Rule (ii) is

$$E(Y) \;=\; \sum_{i=1}^{m} E(Y|Z{=}z_i)\cdot P(Z{=}z_i) \tag{9.22}$$

(see Exercise 9-3). According to this equation, we can also compute the expectation of Y from the conditional expectations $E(Y|Z{=}z_i)$ and the probabilities $P(Z{=}z_i)$.

9.4 Discrete Conditional Expectation

The *discrete conditional expectation* $E(Y|X)$ of a numerical random variable $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ given a random variable $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ is now defined as that random variable on $(\Omega,\mathscr{A},P)$ whose values are identical to the conditional expectation values $E(Y|X{=}x)$. In this definition we have to assume that X is *discrete*, i. e., we assume that there is a finite or countable set $\Omega'_0\subset\Omega'_X$ such that $P_X(\Omega'_0)=1$ and $P(X{=}x)>0$ for all $x\in\Omega'_0$ (see Def. 5.52). In chapter 10 this limitation is dropped.

Definition 9.13 (Discrete Conditional Expectation)
Let $Y\colon (\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a numerical random variable that is nonnegative or has a finite expectation and let the random variable $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be discrete.

(i) If $x_1,\dots,x_m\in\Omega'_X$ such that $P_X(\{x_1,\dots,x_m\})=1$ and, for all $i=1,\dots,m$ and $P(X{=}x_i)>0$, then the discrete conditional expectation of Y given X is defined by

$$E(Y|X) := \sum_{i=1}^{m} E(Y|X{=}x_i)\cdot 1_{X=x_i}\,. \tag{9.23}$$

(ii) If $x_1,x_2,\ldots\in\Omega'_X$ such that $P_X(\{x_1,x_2,\ldots\})=1$ and, for all $i=1,2,\ldots$, $\{x_i\}\in\mathscr{A}'_X$ and $P(X{=}x_i)>0$, then the discrete conditional expectation of Y given X is defined by

$$E(Y|X) := \sum_{i=1}^{\infty} E(Y|X{=}x_i)\cdot 1_{X=x_i}\,. \tag{9.24}$$

Hence, in contrast to a conditional expectation value $E(Y|X{=}x)$, a discrete conditional expectation $E(Y|X)$ is a discrete *random variable* (see Def. 5.52) on $(\Omega,\mathscr{A},P)$.

Remark 9.14 (X-Conditional Probability) If $A\in\mathscr{A}$, then we use the notation

$$P(A|X) := E(1_A|X) \tag{9.25}$$

and call it the *X-conditional probability of A*. If Y is dichotomous with values 0 and 1, we also use the notation $P(Y{=}1|X)$ for the *X-conditional probability of the event* $\{Y{=}1\}$. If Y is dichotomous with values 0 and 1, Equations (9.20) and (9.25) then yield

$$P(Y{=}1|X) = E(Y|X). \tag{9.26}$$

◁

Remark 9.15 (Uniqueness and Values of the Conditional Expectation) An alternative way to write Equations (9.23) and (9.24) is

$$E(Y|X)(\omega) = \begin{cases} E(Y|X{=}x), & \text{if } \omega\in X^{-1}(\{x\}), \quad \forall x\in\Omega'_0,\\ 0, & \text{otherwise.}\end{cases} \tag{9.27}$$

Hence, the values of the conditional expectation $E(Y|X)$ are uniquely defined by Equations (9.23) and (9.24) for all $\omega\in\Omega$ (see Example 9.21). Assigning the value $E(Y|X)(\omega)=0$ if $\omega\in\Omega\setminus X^{-1}(\Omega'_0)$ is arbitrary, but note that $P\big(\Omega\setminus X^{-1}(\Omega'_0)\big)=0$. According to Equation (9.27) and Definition 5.52 this arbitrary assignment does not occur if $P(X{=}x)>0$ for all $x\in X(\Omega)$, i. e., if $\Omega'_0=X(\Omega)$. ◁

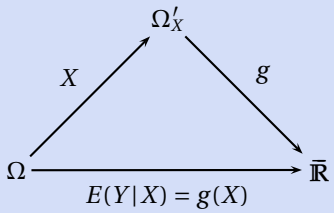


Figure 9.1. The conditional expectation $E(Y|X)$ as the composition of X and its factorization

9.5 Discrete Regression

Remark 9.16 (Measurability and Factorization) Definition 9.13 implies that the discrete conditional expectation $E(Y|X)$ is a random variable on $(\Omega, \mathscr{A}, P)$ that is measurable with respect to X. In more formal terms, $E(Y|X)\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ and $\sigma[E(Y|X)] \subset \sigma(X)$. The reason is that there is a measurable function $g : (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ that is defined by

$$g(x) = \begin{cases} E(Y|X{=}x), & \forall\, x \in \Omega'_0, \\ 0, & \text{otherwise.} \end{cases} \tag{9.28}$$

Hence, $E(Y|X) = g(X)$ (see Fig. 9.1) and Lemma 2.52 implies that $E(Y|X)$ is measurable with respect to X. The function g is called the *factorization of* $E(Y|X)$ or the *discrete regression of* Y *on* X. ◁

Definition 9.17 (Discrete Regression)
Under the assumptions specified in Definition 9.13, the function $g\colon \Omega'_X \to \mathbb{R}$ *defined by Equation* (9.28) *is called the* discrete regression of Y on X.

Remark 9.18 (Regressand and Regressor) Considering the conditinal expectation $E(Y|X)$, we call Y the *regressand* and X the *regressor*. Whereas the regressand Y has to be numerical, the regressor X can be *any* random variable on $(\Omega, \mathscr{A}, P)$ as long as it is discrete and satisfies $P(X{=}x) > 0$ for all $x \in \Omega'_0$. ◁

Remark 9.19 (Multivariate Numerical or Qualitative Regressors) Note that the codomain Ω'_X of X can be *any* set as long as there is a subset $\Omega'_0 \subset \Omega'_X$ such that Ω'_0 is finite, or countable with $P(X \in \Omega'_0) = 1$ and $P(X{=}x) > 0$ for all $x \in \Omega'_0$. Hence, the *regressor* X can be uni- or multivariate (see Exercises 9-4 and 9-5). If $X = (X_1, \ldots, X_n)$ is a discrete multivariate random variable, then we also use the notation $E(Y|X_1, \ldots, X_n)$ instead of $E(Y|X)$ for the conditional expectation of Y given X. ◁

Table 9.1. Joe and Ann With Random Assignment: Conditional Expectations

Elements of Ω			Random variables			Conditional expectations		
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$P(X=1\|U)$
(*Joe*,	*no*,	−)	.09	*Joe*	0	0	.70	.45	.40
(*Joe*,	*no*,	+)	.21	*Joe*	0	1	.70	.45	.40
(*Joe*,	*yes*,	−)	.04	*Joe*	1	0	.80	.60	.40
(*Joe*,	*yes*,	+)	.16	*Joe*	1	1	.80	.60	.40
(*Ann*,	*no*,	−)	.24	*Ann*	0	0	.20	.45	.40
(*Ann*,	*no*,	+)	.06	*Ann*	0	1	.20	.45	.40
(*Ann*,	*yes*,	−)	.12	*Ann*	1	0	.40	.60	.40
(*Ann*,	*yes*,	+)	.08	*Ann*	1	1	.40	.60	.40

9.6 Examples

We treat two examples in some detail. Example 9.20 is straightforward, whereas Example 9.21 exemplifies that the values of a conditional expectation are uniquely defined by Equation (9.23) for all $\omega \in \Omega$.

Example 9.20 (Joe and Ann With Random Assignment – continued) Table 9.1 contains three conditional expectations we may consider in the example already used in chapter 1 (see p. 7). All of them are random variables taking a numerical value for each $\omega \in \Omega$. This means that we might look at their expectations, their variances, their covariances and correlations with other random variables. According to Remark 9.15, the values of $E(Y|X)$ are the conditional expectation values $E(Y|X{=}x)$ for all $x \in \Omega'_0$, and they are 0 for all $x \in \Omega'_X \setminus \Omega'_0$.

We start illustrating the conditional expectation of Y given X. Both random variables, X and Y, are specified in Table 9.1. We consider the random variable $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ with $(\Omega'_X, \mathscr{A}'_X) = (\mathbb{R}, \mathscr{B})$. In this case, $\Omega'_X = \mathbb{R}$ and $\Omega'_0 = \{0,1\}$. Because X takes on a value in $\Omega'_0 = \{0,1\}$ for *all* $\omega \in \Omega$, the conditional expectation $E(Y|X)$ either takes on the value $E(Y|X{=}0)$ or the value $E(Y|X{=}1)$. It does not take on the value 0, because $\{X{=}x\} = \emptyset$ for all $x \in \mathbb{R} \setminus \{0,1\}$ [see Eq. (9.27)].

Because Y is an indicator, $E(Y|X{=}x) = P(Y{=}1|X{=}x)$ [see Eqs. (9.14) and (9.26)]. Hence, if we want to compute the values of $E(Y|X) = P(Y{=}1|X)$, we have to compute the conditional probabilities $P(Y{=}1|X{=}x)$. For $x = 0$ we receive

$$P(Y{=}1|X{=}0) = \frac{P(Y{=}1, X{=}0)}{P(X{=}0)} = \frac{.21 + .06}{.09 + .21 + .24 + .06} = \frac{.27}{.60} = .45,$$

and for $x=1$,

$$P(Y=1\,|\,X=1) \;=\; \frac{P(Y=1,X=1)}{P(X=1)} \;=\; \frac{.16+.08}{.04+.16+.12+.08} \;=\; \frac{.24}{.40} \;=\; .60.$$

Now we consider the conditional expectation $E(Y\,|\,X,U)$. The regressor is the random variable $(X,U)\colon(\Omega,\mathscr{A},P)\to[\mathbb{R}\times\Omega_U,\mathscr{B}\otimes\mathscr{P}(\Omega_U)]$, where $\Omega_U=\{Joe,\,Ann\}$, and $\Omega_0'=\{0,1\}\times\Omega_U$. Because,

$$\forall\,(x,u)\;\in\;(\mathbb{R}\times\Omega_U)\setminus(\{0,1\}\times\Omega_U)\colon\;\{(X{=}x,U{=}u)\}=\emptyset,$$

$E(Y\,|\,X,U)$ takes on a value in Ω_0' for *all* $\omega\in\Omega$. Furthermore, because Y is an indicator, $E(Y\,|\,X,U)=P(Y{=}1\,|\,X,U)$, and this conditional expectation has only four values: the conditional probabilities $P(Y{=}1\,|\,X{=}x,U{=}u)$. For $x=0$, $u{=}Joe$ we receive

$$P(Y=1\,|\,X=0,U{=}Joe) \;=\; \frac{P(Y=1,X=0,U{=}Joe)}{P(X=0,U{=}Joe)} \;=\; \frac{.21}{.09+.21} \;=\; .70,$$

for $x=1$, $u{=}Joe$,

$$P(Y=1\,|\,X=1,U{=}Joe) \;=\; \frac{P(Y=1,X=1,U{=}Joe)}{P(X=1,U{=}Joe)} \;=\; \frac{.16}{.04+.16} \;=\; .80,$$

for $x=0$, $u{=}Ann$,

$$P(Y=1\,|\,X=0,U{=}Ann) \;=\; \frac{P(Y=1,X=0,U{=}Ann)}{P(X=0,U{=}Ann)} \;=\; \frac{.06}{.24+.06} \;=\; .20,$$

and for $x=1$, $u{=}Ann$,

$$P(Y=1\,|\,X=1,U{=}Ann) \;=\; \frac{P(Y=1,X=1,U{=}Ann)}{P(X=1,U{=}Ann)} \;=\; \frac{.08}{.12+.08} \;=\; .40.$$

◁

Example 9.21 (No Treatment for Joe) Let us use a second example in order to illustrate the concepts introduced above. Again the random experiment consists of sampling a person, observing whether or not the sampled person receives a treatment ($x=1$ vs. $x=0$), and observing whether or not a success criterion is reached some time after treatment. In this new example, we fixed new probabilities of the elementary events. For instance, now the probability that Joe receives treatment is zero. This is useful to illustrate some general properties of conditional expectations. Also note that the probabilities of the other elementary events have been changed as well. The only restriction on the probabilities of the elementary events in such a hypothetical example is that they sum up to one.

Using the probabilities displayed in Table 9.2, Equation (9.20) yields:

$$\begin{aligned} E(Y\,|\,X=1) \;&=\; \sum_{i=1}^{2} y_i\cdot P(Y=y_i\,|\,X=1) \\ &=\; 0\cdot P(Y=0\,|\,X=1)\;+\;1\cdot P(Y=1\,|\,X=1) \;=\; P(Y=1\,|\,X=1) \\ &=\; \frac{P(Y=1,X=1)}{P(X=1)} \;=\; \frac{0+.152}{0+0+.228+.152} \;=\; .40 \end{aligned}$$

Table 9.2. No Treatment for Joe With Conditional Expectations

Elements of Ω		Observables			Conditional expectations		
Unit, Treatment, Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\mid X,U)$	$E(Y\mid X)$	$P(X=1\mid U)$
(*Joe*, *no*, −)	.152	*Joe*	0	0	.696	.60	0
(*Joe*, *no*, +)	.348	*Joe*	0	1	.696	.60	0
(*Joe*, *yes*, −)	0	*Joe*	1	0	0	.40	0
(*Joe*, *yes*, +)	0	*Joe*	1	1	0	.40	0
(*Ann*, *no*, −)	.096	*Ann*	0	0	.20	.60	.76
(*Ann*, *no*, +)	.024	*Ann*	0	1	.20	.60	.76
(*Ann*, *yes*, −)	.228	*Ann*	1	0	.40	.40	.76
(*Ann*, *yes*, +)	.152	*Ann*	1	1	.40	.40	.76

for the treatment condition. Applying the corresponding formula to the control condition yields $E(Y\mid X=0) = (.348+.024)\,/\,(.152+.348+.096+.024) = .60$. Note that the conditional probabilities $P(Y=1\mid X=1)$ and $P(Y=1\mid X=0)$ do *not* necessarily add up to 1. In contrast, the sum of $P(Y=1\mid X=1)$ and $P(Y=0\mid X=1)$ and the sum of $P(Y=1\mid X=0)$ and $P(Y=0\mid X=0)$ is always equal to 1, provided that Y is dichotomous with values 0 and 1.

Table 9.2 also displays the conditional probability $P(X=1\mid U)$, whose values are the treatment probabilities of Joe and Ann. For Joe, the treatment probability is $P(X=1\mid U=\mathit{Joe}) = 0$, and for Ann it is

$$P(X=1\mid U=\mathit{Ann}) \;=\; (.228+.152)\,/\,(.096+.024+.228+.152) \;=\; .76.$$

Finally, we compute the conditional expectations $E(Y\mid X=x,U=u)$:

$$E(Y\mid X=0,U=\mathit{Joe}) \;=\; .348\,/\,(.152+.348) \;=\; .696,$$

$$E(Y\mid X=0,U=\mathit{Ann}) \;=\; .024\,/\,(.096+.024) \;=\; .20,$$

and

$$E(Y\mid X=1,U=\mathit{Ann}) \;=\; .152\,/\,(.228+.152) \;=\; .40.$$

Note that $E(Y\mid X=1,U=\mathit{Joe})$ is not defined, because $P(X=1,U=\mathit{Joe})=0$. ◁

Example 9.22 (No Treatment for Joe – continued) Using the results obtained in Example 9.21, Equation (9.23) yields:

$$\begin{aligned} E(Y|X) &= \sum_x E(Y|X=x)\cdot 1_{X=x} \\ &= E(Y|X=0)\cdot 1_{X=0} + E(Y|X=1)\cdot 1_{X=1} \\ &= .60\cdot 1_{X=0} + .40\cdot 1_{X=1}. \end{aligned}$$

Hence, the values $E(Y|X)(\omega)$ of the X-conditional expectation of Y are

$$E(Y|X)(\omega) = .60\cdot 1_{X=0}(\omega) + .40\cdot 1_{X=1}(\omega) = .60, \quad \text{if } X(\omega)=0,$$

and

$$E(Y|X)(\omega) = .60\cdot 1_{X=0}(\omega) + .40\cdot 1_{X=1}(\omega) = .40, \quad \text{if } X(\omega)=1.$$

These are the only two values that $E(Y|X)$ takes on depending on the outcome $\omega\in\Omega$ of the random experiment considered (see the first column of Table 9.2). This example illustrates that $E(Y|X)$ is a random variable on $(\Omega,\mathscr{A},P)$ just like X, Y, and U. Note again that the two values of $E(Y|X)=P(Y=1|X)$ add up to 1 only by coincidence.

Table 9.2 shows two additional conditional expectations, $E(X|U)=P(X=1|U)$ as well as $E(Y|X,U)=P(Y=1|X,U)$. Again using Equation (9.23) and the results obtained in Example 9.21, the conditional expectation $P(X=1|U)$ is

$$\begin{aligned} P(X=1|U) &= \sum_u P(X=1|U=u)\cdot 1_{U=u} \\ &= P(X=1|U=Joe)\cdot 1_{U=Joe} + P(X=1|U=Ann)\cdot 1_{U=Ann} \\ &= .00\cdot 1_{U=Joe} + .76\cdot 1_{U=Ann}. \end{aligned}$$

Hence, the values $P(X=1|U)(\omega)$ of the U-conditional expectation of X are

$$P(X=1|U)(\omega) = .00\cdot 1_{U=Joe}(\omega) + .76\cdot 1_{U=Ann}(\omega) = .00, \quad \text{if } U(\omega)=Joe$$

and

$$P(X=1|U)(\omega) = .00\cdot 1_{U=Joe}(\omega) + .76\cdot 1_{U=Ann}(\omega) = .76, \quad \text{if } U(\omega)=Ann.$$

These are the only two values that $P(X=1|U)$ takes on. Again, Table 9.2 shows how the values of $P(X=1|U)$ are assigned to the outcomes $\omega\in\Omega$.

Finally, let us turn to the conditional expectation $E(Y|X,U)$. Using the results obtained in Example 9.21, its values are

$$E(Y|X,U)(\omega) = E(Y|X=0,U=Joe) = .696, \quad \text{if } X(\omega)=0 \text{ and } U(\omega)=Joe$$

$$E(Y|X,U)(\omega) = E(Y|X=0,U=Ann) = .20, \quad \text{if } X(\omega)=0 \text{ and } U(\omega)=Ann$$

and

$$E(Y|X,U)(\omega) = E(Y|X=1,U=Ann) = .40, \quad \text{if } X(\omega)=1 \text{ and } U(\omega)=Ann,$$

whereas

$$E(Y|X,U)(\omega) = 0, \quad \text{if } X(\omega)=1 \text{ and } U(\omega)=Joe.$$

Note that the value $E(Y|X,U)(\omega)$ *is* defined for $\omega \in \{X=1, U=Joe\}$ although the conditional expectation value $E(Y|X=1, U=Joe)$ *is not* defined. Also note that in this case the value $E(Y|X,U)(\omega) = 0$ is arbitrarily fixed and has no substantive meaning. However, because $P(X=1, U=Joe) = 0$, this has no disadvantageous consequences. [In chapter 10, we show that the values of a conditional expectation can arbitrarily be fixed for all elements ω of a subset A of Ω for which $P(A) = 0$.]

According to Equation (9.23), the conditional expectation $E(Y|X,U)$ is

$$\begin{aligned} E(Y|X,U) &= \sum_{(x,u)} E(Y|X=x, U=u) \cdot 1_{X=x, U=u} \\ &= .696 \cdot 1_{X=0, U=Joe} + .20 \cdot 1_{X=0, U=Ann} + .40 \cdot 1_{X=1, U=Ann}. \end{aligned} \tag{9.29}$$

The pair (1, *Joe*) is not an element of the set Ω_0' (see Def. 9.13), and therefore the corresponding indicator $1_{X=1, U=Joe}$ does not occur in this sum. Hence, if

$$\omega \in \{X=1, U=Joe\} = \{(Joe, yes, -), (Joe, yes, +)\},$$

then all three indicators occurring in Equation (9.29) take on the value 0, implying that $E(Y|X,U)(\omega) = 0$ for these two elements of Ω. ◁

9.7 Proofs

Proof of Theorem 9.3

According to Lemma 4.27 (ii), Equation (9.5) can also be written

$$\begin{aligned} E(Y|B) = E^B(Y) &= \int Y \, dP^B && [(9.5)] \\ &= \int Y \, d\left(\frac{1}{P(B)} \cdot 1_B\right) \odot P && [\text{Lem. 4.27 (ii)}] \\ &= \frac{1}{P(B)} \cdot \int 1_B \cdot Y \, dP && [(3.72), (3.32)] \\ &= \frac{1}{P(B)} \cdot E(1_B \cdot Y), && [(6.1)] \end{aligned}$$

which yields Equation (9.7). Using Lemma 3.33 (i) yields Theorem 9.3 (i) and Lemma 3.33 (ii) implies Theorem 9.3 (ii).

According to Equation (9.7),

$$\begin{aligned} E(Y|B) &= \frac{1}{P(B)} \cdot E(1_B \cdot Y) && [(9.7)] \\ &= \frac{1}{P(B)} \cdot [Cov(1_B, Y) + E(1_B) \cdot E(Y)] && [\text{Box 7.1 (i)}] \\ &= \frac{1}{P(B)} \cdot Cov(1_B, Y) + E(Y). && [E(1_B) = P(B)] \end{aligned}$$

9.8 Exercises

▷ **Exercise 9-1** Prove the proposition of Remark 9.1

▷ **Exercise 9-2** Prove the rules of computation of conditional expectation values displayed in Box 9.2.

▷ **Exercise 9-3** Show that Equation (9.22) is a special case of Rule (ii) of Box 9.2.

▷ **Exercise 9-4** Why is the conditional expectation value $E(Y|X=1, U=Joe)$ not defined in the example presented in Table 9.2 (p. 282)?

▷ **Exercise 9-5** Compute the values of the conditional expectation $E(Y|X,U)$ in the example presented in Table 9.2 (p. 282).

Solutions

▷ **Solution 9-1** It is sufficient to prove $P^{X=x}(\{f(X)=f(x)\}^c)=0$, which is equivalent to $P^{X=x}(\{f(X)=f(x)\})=1$.

$$\begin{aligned} P^{X=x}[f(X)=f(x)] &= \frac{P\big(f(X)=f(x), X=x\big)}{P(X=x)} && [(9.4), (4.2)] \\ &= \frac{P(X=x)}{P(X=x)} && [\{X=x\}\subset\{f(X)=f(x)\}, (9.9)] \\ &= 1. \end{aligned}$$

▷ **Solution 9-2**

(i)
$$\begin{aligned} E[f(X)\cdot Y\,|X=x] &= \frac{1}{P(X=x)}\cdot E[1_{X=x}\cdot f(X)\cdot Y] && [(9.7)] \\ &= \frac{1}{P(X=x)}\cdot E[1_{X=x}\cdot f(x)\cdot Y] && [1_{X=x}\cdot f(X) = 1_{X=x}\cdot f(x)] \\ &= f(x)\cdot\frac{1}{P(X=x)}\cdot E(1_{X=x}\cdot Y) && [\text{Box 6.1 (iii)}] \\ &= f(x)\cdot E(Y\,|X=x) && [(9.7)] \\ &= E[f(x)\cdot Y\,|X=x]. && [\text{Box 9.1 (iii)}] \end{aligned}$$

(ii)
$$\begin{aligned} E(Y|X=x) &= \frac{1}{P(X=x)}\cdot E(1_{X=x}\cdot Y) && [(9.11)] \\ &= \frac{1}{P(X=x)}\cdot E\Big(1_{X=x}\cdot\sum_{i=1}^{m} 1_{Z=z_i}\cdot Y\Big) && \Big[1 \underset{P}{=} \sum_{i=1}^{m} 1_{Z=z_i}, \text{Rem. 6.27}\Big] \\ &= \sum_{i=1}^{m}\frac{1}{P(X=x)}\cdot E(1_{X=x}\cdot 1_{Z=z_i}\cdot Y) && [\text{Box 6.1 (ii), (iii)}] \\ &= \sum_{i=1}^{m}\frac{P(X=x, Z=z_i)}{P(X=x)}\cdot\frac{1}{P(X=x, Z=z_i)}\cdot E(1_{X=x, Z=z_i}\cdot Y) \end{aligned}$$

$$= \sum_{i=1}^{m} E(Y \mid X{=}x, Z{=}z_i) \cdot P(Z{=}z_i \mid X{=}x). \qquad [(9.11)]$$

(iii) $$\sum_{i=1}^{\infty} E(Y \mid X{=}x, Z{=}z_i) \cdot P(Z{=}z_i \mid X{=}x)$$

$$= \sum_{i=1}^{\infty} P(Z{=}z_i \mid X{=}x) \cdot \frac{1}{P(X{=}x, Z{=}z_i)} \cdot E(1_{X=x, Z=z_i} \cdot Y) \qquad [(9.11)]$$

$$= \sum_{i=1}^{\infty} \frac{1}{P(X{=}x)} \cdot E(1_{X=x, Z=z_i} \cdot Y) \qquad [(4.2)]$$

$$= \frac{1}{P(X{=}x)} \sum_{i=1}^{\infty} \int 1_{X=x, Z=z_i} \cdot Y \, dP \qquad [\text{Def. 6.1}]$$

$$= \frac{1}{P(X{=}x)} \sum_{i=1}^{\infty} \left[\int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^+ \, dP - \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^- \, dP \right] \qquad [\text{Def. 3.28}]$$

$$= \frac{1}{P(X{=}x)} \left[\sum_{i=1}^{\infty} \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^+ \, dP - \sum_{i=1}^{\infty} \int 1_{Z=z_i} \cdot 1_{X=x} \cdot Y^- \, dP \right]$$

[$1_{X=x} \cdot Y$ is quasi-integrable]

$$= \frac{1}{P(X{=}x)} \left[\int 1_{X=x} \cdot Y^+ \, dP - \int 1_{X=x} \cdot Y^- \, dP \right] \qquad [(3.65),\ 1 \underset{P}{=} \sum_{i=1}^{\infty} 1_{Z=z_i},\ (3.44)]$$

$$= \frac{1}{P(X{=}x)} \int 1_{X=x} \cdot Y \, dP \qquad [\text{Def. 3.28}]$$

$$= E(Y \mid X{=}x). \qquad [(9.11)]$$

▷ **Solution 9-3** This is easily seen considering the special case $X{=}\alpha$, $\alpha \in \Omega'_X$, i. e., the case in which X is a constant. Then $\{X{=}\alpha\} = \Omega$, and Equation (9.11) yields $E(Y \mid X{=}\alpha) = E(Y)$, and $E(Y \mid X{=}\alpha, Z{=}z_i) = E(Y \mid Z{=}z_i)$. Hence, Rule (ii) of Box 9.2 yields

$$E(Y) = \sum_{i=1}^{m} E(Y \mid Z{=}z_i) \cdot P(Z{=}z_i).$$

▷ **Solution 9-4** In this example, $P(X{=}1, U{=}Joe) = 0$. This implies that the conditional probabilities $P(Y{=}y \mid X{=}1, U{=}Joe)$ that are used in the definition of $E(Y \mid X{=}1, U{=}Joe)$ (see Def. 9.2) are not defined.

▷ **Solution 9-5** The values of the conditional expectation $E(Y \mid X, U)$ are the four conditional expectation values $E(Y|X{=}x, U{=}u)$. Because $E(Y \mid X{=}x, U{=}u) = P(Y{=}1 \mid X{=}x, U{=}u)$, they can be computed as follows:

$$P(Y{=}1 \mid X{=}0, U{=}Joe) = \frac{P(Y{=}1, X{=}0, U{=}Joe)}{P(X{=}0, U{=}Joe)} = \frac{.348}{.152 + .348} = .696,$$

$$P(Y{=}1 \mid X{=}0, U{=}Ann) = \frac{P(Y{=}1, X{=}0, U{=}Ann)}{P(X{=}0, U{=}Ann)} = \frac{.024}{.096 + .024} = .20,$$

$$P(Y{=}1 \mid X{=}1, U{=}Ann) = \frac{P(Y{=}1, X{=}1, U{=}Ann)}{P(X{=}1, U{=}Ann)} = \frac{.152}{.228 + .152} = .40.$$

The conditional expectation value $E(Y|X=1, U=Joe) = P(Y=1|X=1, U=Joe)$ is undefined, because $P(X=1, U=Joe) = 0$. Choosing 0 as a value of $E(Y|X,U)$ for $\omega_3 = (Joe, yes, -)$ and $\omega_4 = (Joe, yes, +)$ uniquely defines $E(Y|X,U)$.

Chapter 10
Conditional Expectation

In chapter 9 we treated the conditional expectation value given an event and the conditional expectation $E(Y|X)$ assuming that X is discrete taking on each of its values x with a positive probability $P(X{=}x) > 0$. In this chapter we introduce the general concept of a conditional expectation given a σ-algebra $\mathscr{C}$. The price of this generalization is that a $\mathscr{C}$-conditional expectation is uniquely defined only up to P-equivalence, i. e., if there are two versions of such a $\mathscr{C}$-conditional expectation of a numerical random variable Y, then they are not necessarily identical, but they are P-equivalent. Furthermore, if $\mathscr{C}$ is generated by a random variable X, then a $\mathscr{C}$-conditional expectation is also called a *conditional expectation of Y given X*. This definition also applies if X is continuous. Hence, it also holds if $P(X{=}x) = 0$ for all values x of X. In this chapter, we also define the general concept of a regression as a factorization g of a conditional expectation $E(Y|X) = g(X)$, and an $(X{=}x)$-conditional expectation value $E(Y|X{=}x)$ as a value $g(x)$ of g. This means that $E(Y|X{=}x)$ is defined even if $P(X{=}x) = 0$. However, $E(Y|X{=}x)$ is not uniquely defined. Nevertheless, we can formulate propositions about the conditional expectation values $E(Y|X{=}x)$ for P_X-almost all values x of X. Finally, we introduce the concepts of *mean independence*, and study their relationship to stochastic independence and correlational independence.

10.1 Assumptions and Definitions

Throughout this chapter, we will make the following assumptions and use the following notation.

Notation and Assumptions 10.1
$Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a numerical random variable that is nonnegative or has a finite expectation $E(Y)$. Furthermore, $\mathscr{C} \subset \mathscr{A}$ is a σ-algebra, and $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ is a random variable.

The definition of a conditional expectation given a σ-algebra is already found in Kolmogorov (1933/1977) (see also Kolmogorov, 1956). Reading the following definition, remember that, for a random variable $V\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, we use $\sigma(V) = V^{-1}(\overline{\mathscr{B}})$ to denote the σ-algebra generated by V, and $\sigma(V) \subset \mathscr{C}$ means that V is $(\mathscr{C}, \overline{\mathscr{B}})$-measurable (see Def. 2.26 and Cor. 2.28).

Definition 10.2 (Conditional Expectation Given a σ-Algebra)
Let the assumptions 10.1 hold. A random variable $V\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ is called a (version of the) $\mathscr{C}$-conditional expectation of Y with respect to P, if the following two conditions hold:

(a) $\sigma(V) \subset \mathscr{C}$.
(b) $E(1_C \cdot V) = E(1_C \cdot Y), \quad \forall\, C \in \mathscr{C}$.

If V satisfies (a) and (b), then we also use the notation $E(Y|\mathscr{C}) := V$.

Remark 10.3 (X-Conditional Expectation) If the assumptions 10.1 hold, then we define

$$E(Y|X) := E[Y|\sigma(X)] \tag{10.1}$$

and call it a version of the *X-conditional expectation* of Y with respect to P. If there is no ambiguity, we will omit the reference to the measure P. Otherwise, we also use the terms *X-conditional P-expectation of Y*. ◁

Remark 10.4 (Conditional Probability Given a σ-Algebra) Let the assumptions 10.1 hold and let $A \in \mathscr{A}$. Then we define

$$P(A|\mathscr{C}) := E(1_A|\mathscr{C}) \tag{10.2}$$

and call it a version of the *$\mathscr{C}$-conditional probability* of A with respect to P. Similarly, we define

$$P(A|X) := E[1_A|\sigma(X)] \tag{10.3}$$

and call it a version of the *X-conditional probability* of A with respect to P.

Furthermore, considering the event $\{Y=y\}$, we also use the notation

$$P(Y=y|X) := P(\{Y=y\}|X) = E(1_{Y=y}|X). \tag{10.4}$$

◁

Remark 10.5 (Conditioning on the Smallest σ-Algebra) If $\mathscr{C} = \{\Omega,\emptyset\}$, then Definition 10.2 (a) implies that $E(Y|\mathscr{C})$ is a constant (see Example 2.14), and 10.2 (b) implies

$$E(Y|\mathscr{C}) = E(Y),$$

because $E[1_\Omega \cdot E(Y)] = E(1_\Omega \cdot Y) = E(Y)$ and $E[1_\emptyset \cdot E(Y)] = E(1_\emptyset \cdot Y) = 0$. In fact, if $\mathscr{C} = \{\Omega,\emptyset\}$, then $E(Y)$ is the only version of the $\mathscr{C}$-conditional expectation of Y. Correspondingly, if X is a constant, i. e., if $X = \alpha$, $\alpha \in \Omega'_X$, then

$$E(Y|X) = E(Y).$$

◁

Remark 10.6 (C-Conditional Expectation of a $\mathscr{C}$-Conditional Expectation) Consider an event $C \in \mathscr{C}$ with $P(C) > 0$ and a version V of the C-conditional expectation of Y defined in Equation (9.5). Then

$$\begin{aligned} E(V|C) &= \frac{1}{P(C)} \cdot E(1_C \cdot V) && [(9.7)] \\ &= \frac{1}{P(C)} \cdot E(1_C \cdot Y) && [\text{Def. 10.2 (b)}] \\ &= E(Y|C). && [(9.7)] \end{aligned} \tag{10.5}$$

Inserting $E(Y|\mathscr{C}) = V$ into this equation shows that with condition (b) of Definition 10.2 we also require

$$E[E(Y|\mathscr{C})|C] = E(Y|C), \quad \forall C \in \mathscr{C} \text{ with } P(C) > 0. \tag{10.6}$$

◁

Remark 10.7 (Multivariate X) If $X = (X_1, \ldots, X_n)$ is an n-variate random variable on $(\Omega, \mathscr{A}, P)$ (see section 5.3), then $E(Y|X)$ is also denoted by $E(Y|X_1, \ldots, X_n)$. ◁

Remark 10.8 ($\mathscr{C}$-Conditional and X-Conditional Expectation) If $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ is a *nonnegative* real-valued random variable, then $\sigma(X^2) = \sigma(X)$ [see Example 2.56 (i)]. Therefore, in this case $E(Y|X)$ and $E(Y|X^2)$ are just two different notations of the conditional expectation $E[Y|\sigma(X)]$. If X takes on also negative real numbers, then $\sigma(X^2) \subset \sigma(X)$, but $\sigma(X^2) = \sigma(X)$ does not necessarily hold [see Example 2.56 (ii)]. ◁

10.2 Existence and Uniqueness

By its definition it is not obvious that a conditional expectation exists and that it is well-defined. These issues are addressed in the following theorem.

Theorem 10.9 (Existence and Uniqueness of a $\mathscr{C}$-Conditional Expectation) *Let the assumptions 10.1 hold. Then:*

(i) There is a $\mathscr{C}$-measurable random variable $V\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ that is nonnegative (if Y is nonnegative) or has a finite expectation $E(V)$ (if $E(Y)$ is finite) satisfying

$$E(1_C \cdot V) = E(1_C \cdot Y), \quad \forall C \in \mathscr{C}. \tag{10.7}$$

(ii) If $V, V^*\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ satisfy (10.7) and $\sigma(V), \sigma(V^*) \subset \mathscr{C}$, then $V \underset{P}{=} V^*$.

For a proof see Bauer (1996, Theorem 15.1, p. 111). Using the term 'version of a conditional expectation' already hints at the fact that a conditional expectation,

even if it exists, is not necessarily uniquely defined. However, according to Theorem 10.9 (ii), different versions of a conditional expectations are P-equivalent (see Remark 5.13 and Exercise 10-1).

Remark 10.10 (The Sets $\mathscr{E}(Y|\mathscr{C})$ and $\mathscr{E}(Y|X)$) We define $\mathscr{E}(Y|\mathscr{C})$ to be the set of all $\mathscr{C}$-measurable random variables satisfying Equation (10.7). Hence, $\mathscr{E}(Y|\mathscr{C})$ is the set of all versions of the $\mathscr{C}$-conditional expectation of Y with respect to the measure P. Similarly, $\mathscr{E}(Y|X)$ denotes the set of all versions of the X-conditional expectation of Y. The sets $\mathscr{P}(A|\mathscr{C})$ and $\mathscr{P}(A|X)$ are defined correspondingly. ◁

10.2.1 Uniqueness With Respect to a Probability Measure

Remark 10.11 (Uniqueness of $E(Y|\mathscr{C})$ With Respect to a Probability Measure) Let the assumptions 10.1 hold and let Q be a probability measure on $(\Omega,\mathscr{A})$. Then we define

$$E(Y|\mathscr{C}) \text{ is } Q\text{-unique} \quad :\Leftrightarrow \quad \forall V, V^* \in \mathscr{E}(Y|\mathscr{C}):\ V \underset{Q}{=} V^*. \tag{10.8}$$

This term is convenient not only for $Q = P$. According to the following remark $E(Y|\mathscr{C})$ is P-unique. ◁

Remark 10.12 ($\mathscr{E}(Y|\mathscr{C})$ is a P-Equivalence Class) Let $V \in \mathscr{E}(Y|\mathscr{C})$ and suppose that the random variable $V^*\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ is $\mathscr{C}$-measurable with $V \underset{P}{=} V^*$. Then, according to Theorem 3.68 (i), Equation (10.7) also holds for V^*, and this implies $V^* \in \mathscr{E}(Y|\mathscr{C})$. Hence, if $V \in \mathscr{E}(Y|\mathscr{C})$, then Theorem 10.9 (ii) implies:

$$V^* \in \mathscr{E}(Y|\mathscr{C}) \quad \Leftrightarrow \quad V^* \underset{P}{=} V \text{ and } V^* \text{ is } \mathscr{C}\text{-measurable.} \tag{10.9}$$

Therefore, $\mathscr{E}(Y|\mathscr{C})$ is the P-equivalence class of V in the set of all $\mathscr{C}$-measurable random variables (see Def. 2.74). ◁

Remark 10.13 (P-Equivalence and $\mathscr{C}$-Measurability) Suppose that V is a version of the $\mathscr{C}$-conditional expectation, i. e., $V \in \mathscr{E}(Y|\mathscr{C})$, and that $V^*\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ is a random variable. Then (10.9) implies

$$V^* \in \mathscr{E}(Y|\mathscr{C}) \quad \Rightarrow \quad V^* \underset{P}{=} V.$$

However, $V^* \underset{P}{=} E(Y|\mathscr{C})$ may be true and yet $V^* \notin \mathscr{E}(Y|\mathscr{C})$, because $V^* \underset{P}{=} E(Y|\mathscr{C})$ does not imply that V^* is $\mathscr{C}$-measurable. ◁

Remark 10.14 (Versions of $E(Y|\mathscr{C})$) For simplicity, we also say that V is a *version of* $E(Y|\mathscr{C})$, meaning $V \in \mathscr{E}(Y|\mathscr{C})$. ◁

Remark 10.15 (Consistency of Definitions) If X is discrete such that there is a finite or countable set $\Omega_0' \subset \Omega_X'$ with $P_X(\Omega_0') = 1$ and $P(X{=}x) > 0$ for all $x \in \Omega_0'$, then the discrete conditional expectation introduced in Definition 9.13 is a version of the conditional expectation of Y given X defined in Remark 10.3 (see Exercise 10-2). This implies

$$E(Y|X) \underset{P}{=} \sum_{x\in\Omega'_0} E(Y|X{=}x)\cdot 1_{X=x} \tag{10.10}$$

and

$$\forall x \in \Omega'_X\colon \quad E(Y|X)(\omega) \;=\; E(Y|X{=}x), \quad \text{if } \omega \in \{X{=}x\}. \tag{10.11}$$

This equation shows that the conditional expectation $E(Y|X)$ describes how the $(X{=}x)$-conditional expectation values of Y depend on the values $x \in \Omega'_0$. ◁

Remark 10.16 (Uniqueness of $E(Y|X)$) If we add $\Omega'_0 = X(\Omega)$ to the assumptions of Remark 10.15, then

$$E(Y|X) \;=\; \sum_{x\in X(\Omega)} E(Y|X{=}x)\cdot 1_{X=x} \tag{10.12}$$

Hence, under these assumptions, $V = V^*$ for all $V, V^* \in \mathscr{E}(Y|X)$, i. e., there is only one single version of the X-conditional expectation of Y. The corresponding results for $E(Y|\mathscr{C})$ are formulated in the following section. ◁

10.2.2 A Necessary and Sufficient Condition of Uniqueness

Now we present a necessary and sufficient condition for uniqueness of a conditional expectation. Reading this theorem, remember that a set is countable, if it is either finite or denumerable. This also applies to a partition of Ω, i. e., a set $\mathscr{E}$ of pairwise disjoint nonempty subsets of Ω with $\bigcup_{A_i\in\mathscr{E}} A_i = \Omega$.

Theorem 10.17 (Uniqueness of $E(Y|\mathscr{C})$)
Let the assumptions 10.1 hold and let $\mathscr{C} = \sigma(\mathscr{E})$, where $\mathscr{E} = \{A_1, A_2, \ldots\}$ is a countable partition of Ω. Then $V = V^*$ for all $V, V^* \in \mathscr{E}(Y|\mathscr{C})$ if and only if

$$P(A_i) \;>\; 0, \quad \forall\, A_i \in \mathscr{E}. \tag{10.13}$$

(Proof p. 311)

Remark 10.18 (Values of a $\mathscr{C}$-Conditional Expectation) According to Theorem 10.17, $E(Y|\mathscr{C})$ is uniquely defined if and only if (10.13) holds, and in this case we can write

$$E(Y|\mathscr{C}) \;=\; \sum_{A_i\in\mathscr{E}} E(Y|A_i)\cdot 1_{A_i} \tag{10.14}$$

and

$$\forall A_i \in \mathscr{E}\colon \quad E(Y|\mathscr{C})(\omega) \;=\; E(Y|A_i), \quad \text{if } \omega \in A_i. \tag{10.15}$$

This equation shows that the conditional expectation $E(Y|\mathscr{C})$ describes how the conditional expectation values $E(Y|A_i)$ depend on the events $A_i \in \mathscr{E}$ (see Exercise 10-3). ◁

10.2.3 Examples

Example 10.19 (No Treatment for Joe – continued) In Table 9.2 (p. 282), the conditional expectation $E(Y|X)$ of the outcome variable Y on the treatment variable X has only two different values, the conditional expectations

$$E(Y|X{=}0) \;=\; .60 \quad \text{and} \quad E(Y|X{=}1) \;=\; .40.$$

The last but one column of Table 9.2 shows how these values are assigned to the eight possible outcomes $\omega \in \Omega$. The values and $E(Y|X)$ itself are uniquely defined.

In contrast, the conditional expectation $E(Y|X,U)$ of Y on the treatment variable X and the person variable U has four different values, .696, .20, .40, and 0 (see Example 9.22). Note that these four values define only one element, say V, of $\mathscr{E}(Y|X,U)$. If, instead of $E(Y|X,U)(\omega) = 0$ for $\omega \in \{(Joe, yes, -), (Joe, yes, +)\}$, we define

$$E(Y|X,U)(\omega) = \alpha, \quad \alpha \neq 0, \quad \alpha \in \mathbb{R}, \quad \text{for } \omega \in \{(Joe, yes, -), (Joe, yes, +)\},$$

then we have a new element, say V^*, of $\mathscr{E}(Y|X,U)$. Because α can be any nonzero real number, in this example, the set $\mathscr{E}(Y|X,U)$ is uncountably infinite. However, because $P(X{=}1, U{=}Joe) = 0$, two elements V and V^* of $\mathscr{E}(Y|X,U)$ are always identical with probability 1, i. e., V and V^* are P-equivalent. Also note that $E(Y|X)$ and $E(Y|X,U)$ are random variables on the same probability space as the other random variables such as Y, X, and U. ◁

Example 10.20 (No Treatment for Joe – continued) In Example 9.22, we specified the conditional expectations $E(Y|X)$, $E(Y|X,U)$, and $P(X{=}1\,|\,U)$. Now we check, if $E(Y|X)$ satisfies conditions (a) and (b) of Definition 10.2. First of all,

$$\mathscr{C} \;=\; \{\Omega,\, \emptyset,\, \{X{=}0\},\, \{X{=}1\}\} \;=\; \sigma(X)$$

is the σ-algebra generated by X, where, in this example, $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ with $\Omega'_X = \{0,1\}$ and $\mathscr{A}'_X = \{\Omega'_X, \emptyset, \{0\}, \{1\}\}$. If $V \in \mathscr{E}(Y|X)$ and $\overline{\mathscr{B}}$ is the Borel σ-algebra on $\overline{\mathbb{R}}$, then

$$V^{-1}(B) \;=\; \begin{cases} \Omega, & \text{if } .40 \in B \text{ and } .60 \in B, \\ \emptyset, & \text{if } .40 \notin B \text{ and } .60 \notin B, \\ \{X{=}0\}, & \text{if } .40 \notin B \text{ and } .60 \in B, \\ \{X{=}1\}, & \text{if } .40 \in B \text{ and } .60 \notin B, \end{cases} \qquad \forall B \in \overline{\mathscr{B}}.$$

Hence, in this example, $V^{-1}(\overline{\mathscr{B}}) = \mathscr{C} = \sigma(X)$. Therefore, condition (a) of Definition 10.2 is satisfied for $V = E(Y|X)$ specified in Example 9.22.

Now we check condition (b) of Definition 10.2 . For $C = \Omega$ this condition requires $E(1_\Omega \cdot V) = E(1_\Omega \cdot Y)$. The expectation of $1_\Omega \cdot Y$ is

$$E(1_\Omega \cdot Y) \;=\; E(Y) \;=\; P(Y{=}1) \;=\; .348 + 0 + .024 + .152 \;=\; .524,$$

and the expectation of $1_\Omega \cdot V$ is:

$$\begin{aligned} E(1_\Omega \cdot V) &= E(V) = .60 \cdot P(X=0) + .40 \cdot P(X=1) \\ &= .60 \cdot (.152 + .348 + .096 + .024) + .40 \cdot (0 + 0 + .228 + .152) = .524 \end{aligned}$$

(see Table 9.2, p. 282). For $C = \emptyset$, condition (b) of Definition 10.2 requires that $E(1_\emptyset \cdot V) = E(1_\emptyset \cdot Y)$, which is always true [see Box 6.1 (v)]. For $C = \{X=0\}$, condition (b) of Definition 10.2 requires $E(1_{X=0} \cdot V) = E(1_{X=0} \cdot Y)$, and the expectations of $1_{X=0} \cdot Y$ and $1_{X=0} \cdot V$ are

$$E(1_{X=0} \cdot Y) = E(1_{X=0} \cdot 1_{Y=1}) = P(X=0, Y=1) = .348 + .024 = .372$$

and

$$\begin{aligned} E(1_{X=0} \cdot V) &= .60 \cdot P(1_{X=0}=1, X=0) + 0 \cdot P(1_{X=0}=1, X=1) \\ &= .60 \cdot P(X=0) = .60 \cdot (.152 + .348 + .096 + .024) = .372 \end{aligned}$$

(see Table 9.2 and Exercise 10-4). Finally, for $C = \{X=1\}$, condition (b) of Definition 10.2 requires that $E(1_{X=1} \cdot V) = E(1_{X=1} \cdot Y)$ and the expectation of $1_{X=1} \cdot Y$ and $1_{X=1} \cdot V$ are

$$E(1_{X=1} \cdot Y) = E(1_{X=1} \cdot 1_{Y=1}) = P(X=1, Y=1) = 0 + .152 = .152$$

and

$$\begin{aligned} E(1_{X=1} \cdot V) &= 0 \cdot P(1_{X=1}=1, X=0) + .40 \cdot P(1_{X=1}=1, X=1) \\ &= .40 \cdot P(X=1) = .40 \cdot (0 + 0 + .228 + .152) = .152 \end{aligned}$$

(see Table 9.2). Hence, $V = E(Y|X)$ satisfies conditions (a) and (b) of Definition 10.2. Therefore, $V = E(Y|X)$ is in fact an element of $\mathscr{E}(Y|X) = \mathscr{E}[Y|\sigma(X)]$. As mentioned before, in this example, $E(Y|X)$ is uniquely defined. This means that it is the only element of the set $\mathscr{E}(Y|X)$. ◁

10.3 Rules of Computation and Other Properties

10.3.1 Rules of Computation

Some rules of computation for $\mathscr{C}$-conditional expectations are presented in Box 10.1, some of which are analog to the rules for expectations (see Box 6.1) and to the rules for B-conditional expectation values (see Box 9.1). In Rule (iv), the term $E[E(Y|\mathscr{C})]$ denotes the expectation (with respect to the measure P) of a $\mathscr{C}$-conditional expectation of Y. Similarly, in Rule (v), the term $E[E(Y|\mathscr{C})|\mathscr{C}_0]$ denotes the $\mathscr{C}_0$-conditional expectation of the $\mathscr{C}$-conditional expectation of Y, where we presume $\mathscr{C}_0 \subset \mathscr{C}$. (For a proof of these rules see Exercise 10-5).

For convenience, in Box 10.2 we translate these rules to $(X=x)$-conditional expectations, i. e., to the case in which $\mathscr{C} = \sigma(X)$. Hence, these properties are special cases of those listed in Box 10.1, and they do not need proofs of their own.

In some of these rules we refer to the composition $f(X) = f \circ X$, where the function $f\colon (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is assumed to be $(\mathscr{A}'_X, \overline{\mathscr{B}})$-measurable. According to Equation (2.29),

$$\sigma[f(X)] = [f(X)]^{-1}(\overline{\mathscr{B}}) = X^{-1}[f^{-1}(\overline{\mathscr{B}})] \ \subset \ X^{-1}(\mathscr{A}'_X) = \sigma(X)\,.$$

In other words, we assume that the composition $f(X)$ is measurable with respect to the σ-algebra $\mathscr{C} = X^{-1}(\mathscr{A}'_X) = \sigma(X)$. Therefore, $\sigma[f(X)]$ can take the role of $\mathscr{C}_0$ in Rule (v) of Box 10.1. Furthermore, $f(X)$ takes the role of Y in Rule (vii) of Box 10.1, and the role of Y_1 in Rules (xiv) and (xiii) of Box 10.1. Remember, according Lemma 2.52, the composition $f(X)$ is measurable with respect to the σ-algebra $\mathscr{C} = X^{-1}(\mathscr{A}'_X) = \sigma(X)$ if $f\colon \Omega'_X \to \overline{\mathbb{R}}$ is $(\mathscr{A}'_X, \overline{\mathscr{B}})$-measurable. Furthermore, according Corollary 2.53, the composition $f(X)$ is measurable with respect to the σ-algebra $\sigma(X)$ if $f\colon \Omega'_X \to \Omega'$, where Ω' is finite or countable and f is $\big((\mathscr{A}'_X, \mathscr{P}(\Omega')\big)$-measurable.

10.3.2 Monotonicity

Box 10.3 displays some monotonicity properties that are proved in Exercise 10-6. Of course, if $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ is a random variable, then these properties also hold for $\mathscr{C} = \sigma(X)$ and $E(Y\,|\,X) = E[Y\,|\,\sigma(X)]$. For example, Rule (i) can then be written

$$Y \underset{P}{\leqq} Z \quad \Rightarrow \quad E(Y\,|\,X) \underset{P}{\leqq} E(Z\,|\,X) \tag{10.16}$$

and Rule (vi) yields

$$Y \underset{P}{\geqq} \alpha \quad \Rightarrow \quad \exists\, V \in \mathscr{E}(Y\,|\,X)\colon V \geq \alpha. \tag{10.17}$$

10.3.3 Convergence Theorems

Now we turn to convergence of $\mathscr{C}$-conditional expectations. Theorems 10.21 and 10.22 provide sufficient conditions that allow to exchange taking the limit and taking the conditional expectation, e. g.,

$$\lim_{i\to\infty} E(Y_i\,|\,\mathscr{C}) \ \underset{P}{=} \ E(\lim_{i\to\infty} Y_i\,|\,\mathscr{C}).$$

This is not only of technical interest for many proofs; instead, it also describes a continuity property of the conditional expectation: If Y_i is a good approximation to Y, then $E(Y_i\,|\,\mathscr{C})$ is a good approximation to $E(Y\,|\,\mathscr{C})$.

The first theorem deals with *monotone convergence*, the second with *dominated convergence*. Reading these theorems, note that $\lim_{i\to\infty} Y_i \underset{P}{=} Y$ means

$$P\big(\{\omega \in \Omega\colon \lim_{i\to\infty} Y_i(\omega) = Y(\omega)\}\big) \ = \ 1, \tag{10.18}$$

i. e., the sequence $(Y_i,\, i \in \mathbb{N})$, converges P-almost surely pointwise to Y.

Box 10.1 Rules of Computation for $\mathscr{C}$-Conditional Expectations

Let $Y, Y_1, Y_2 : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be numerical random variables that are nonnegative or have a finite expectation, $\mathscr{C} \subset \mathscr{A}$ a σ-algebra, and $\alpha \in \mathbb{R}$. Then:

$$E(\alpha|\mathscr{C}) \underset{P}{=} \alpha. \tag{i}$$

$$E(\alpha + Y|\mathscr{C}) \underset{P}{=} \alpha + E(Y|\mathscr{C}). \tag{ii}$$

$$E(\alpha \cdot Y|\mathscr{C}) \underset{P}{=} \alpha \cdot E(Y|\mathscr{C}). \tag{iii}$$

$$E[E(Y|\mathscr{C})] = E(Y). \tag{iv}$$

$$E\big[E(Y|\mathscr{C})\,\big|\,\mathscr{C}_0\big] \underset{P}{=} E(Y|\,\mathscr{C}_0), \quad \text{if } \mathscr{C}_0 \subset \mathscr{C} \text{ is a } \sigma\text{-algebra.} \tag{v}$$

$$E(Y|\mathscr{C}) \underset{P}{=} E(Y), \quad \text{if } Y \underset{P}{\perp\!\!\!\perp} \mathscr{C}. \tag{vi}$$

$$E(Y|\mathscr{C}) \underset{P}{=} Y, \quad \text{if } Y \text{ is } \mathscr{C}\text{-measurable.} \tag{vii}$$

$$E(Y|\mathscr{C}) \underset{P}{=} E\big[Y\,\big|\,E(Y|\mathscr{C})\big]. \tag{viii}$$

$$E(Y_1|\mathscr{C}) \underset{P}{=} E(Y_2|\mathscr{C}), \quad \text{if } Y_1 \underset{P}{=} Y_2. \tag{ix}$$

$$-\infty < E(Y) < \infty \quad \Rightarrow \quad \exists\, V \in \mathscr{E}(Y|\mathscr{C}) \colon V \text{ is real-valued.} \tag{x}$$

$$E(Y^2) < \infty \quad \Rightarrow \quad E[E(Y|\mathscr{C})^2] < \infty. \tag{xi}$$

$$Cov[Y, E(Y|\mathscr{C})] = Var[E(Y|\mathscr{C})], \quad \text{if } E(Y^2) < \infty. \tag{xii}$$

$$Cov[Y_1, E(Y_2|\mathscr{C})] = Cov(Y_1, Y_2), \quad \text{if } \sigma(Y_1) \subset \mathscr{C},\ E(Y_1^2), E(Y_2^2) < \infty. \tag{xiii}$$

Let $E(Y_1^2), E(Y_2^2) < \infty$ or Y_1, Y_2 nonnegative. Then $\sigma(Y_1) \subset \mathscr{C}$ implies:

$$E(Y_1 \cdot Y_2|\mathscr{C}) \underset{P}{=} Y_1 \cdot E(Y_2|\mathscr{C}). \tag{xiv}$$

If Y_1, Y_2 are nonnegative or real-valued with finite expectations, then there is a nonnegative or real-valued version $E(Y_1|\mathscr{C}) \in \mathscr{E}(Y_1|\mathscr{C})$ and a nonnegative or real-valued version $E(Y_2|\mathscr{C}) \in \mathscr{E}(Y_2|\mathscr{C})$ such that

$$E(Y_1 + Y_2|\mathscr{C}) \underset{P}{=} E(Y_1|\mathscr{C}) + E(Y_2|\mathscr{C}). \tag{xv}$$

If $Y_1, \ldots, Y_n$ are real-valued with finite expectations and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, then

$$E\Big(\sum_{i=1}^{n} \alpha_i \cdot Y_i \,\Big|\, \mathscr{C}\Big) \underset{P}{=} \sum_{i=1}^{n} \alpha_i \cdot E(Y_i|\mathscr{C}). \tag{xvi}$$

Box 10.2 Rules of Computation for X-Conditional Expectations

Let $Y, Y_1, Y_2 : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be numerical random variables that are nonnegative or have a finite expectation, let $X : (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable, let $f : (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be measurable, and let $\alpha \in \mathbb{R}$. Then:

$$E(\alpha \,|\, X) \underset{P}{=} \alpha. \tag{i}$$

$$E(\alpha + Y \,|\, X) \underset{P}{=} \alpha + E(Y \,|\, X). \tag{ii}$$

$$E(\alpha \cdot Y \,|\, X) \underset{P}{=} \alpha \cdot E(Y \,|\, X). \tag{iii}$$

$$E[E(Y \,|\, X)] = E(Y). \tag{iv}$$

$$E\big[E(Y \,|\, X) \,\big|\, f(X)\big] \underset{P}{=} E\big[Y \,|\, f(X)\big]. \tag{v}$$

$$E(Y \,|\, X) \underset{P}{=} E(Y), \quad \text{if } Y \underset{P}{\perp\!\!\!\perp} X. \tag{vi}$$

$$E[f(X) \,|\, X] \underset{P}{=} f(X), \quad \text{if } f(X) \geq 0 \text{ or } E[f(X)] < \infty. \tag{vii}$$

$$E(Y \,|\, X) \underset{P}{=} E\big[Y \,\big|\, E(Y \,|\, X)\big]. \tag{viii}$$

$$E(Y_1 \,|\, X) \underset{P}{=} E(Y_2 \,|\, X), \quad \text{if } Y_1 \underset{P}{=} Y_2. \tag{ix}$$

$$-\infty < E(Y) < \infty \quad \Rightarrow \quad \exists\, V \in \mathscr{E}(Y \,|\, X) : V \text{ is real-valued.} \tag{x}$$

$$E(Y^2) < \infty \quad \Rightarrow \quad E\big[E(Y \,|\, X)^2\big] < \infty. \tag{xi}$$

$$Cov[Y, E(Y \,|\, X)] = Var[E(Y \,|\, X)], \quad \text{if } E(Y^2) < \infty. \tag{xii}$$

$$Cov[f(X), E(Y \,|\, X)] = Cov[f(X), Y], \quad \text{if } E[f(X)^2], E(Y^2) < \infty. \tag{xiii}$$

Let $E(Y^2), E[f(X)^2] < \infty$ or $Y, f(X)$ nonnegative. Then:

$$E[f(X) \cdot Y \,|\, X] \underset{P}{=} f(X) \cdot E(Y \,|\, X). \tag{xiv}$$

If Y_1, Y_2 are nonnegative or real-valued with finite expectations, then there is a nonnegative or real-valued version $E(Y_1 \,|\, X) \in \mathscr{E}(Y_1 \,|\, X)$ and a nonnegative or real-valued version $E(Y_2 \,|\, X) \in \mathscr{E}(Y_2 \,|\, X)$ such that

$$E(Y_1 + Y_2 \,|\, X) \underset{P}{=} E(Y_1 \,|\, X) + E(Y_2 \,|\, X). \tag{xv}$$

If $Y_1, \ldots, Y_n$ are real-valued with finite expectations and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, then

$$E\Big(\sum_{i=1}^{n} \alpha_i \cdot Y_i \,\Big|\, X\Big) \underset{P}{=} \sum_{i=1}^{n} \alpha_i \cdot E(Y_i \,|\, X). \tag{xvi}$$

Box 10.3 Monotonicity of Conditional Expectations

Let $Y, Z\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be numerical random variables that are nonnegative or have a finite expectations, $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, and $\alpha \in \mathbb{R}$. Then:

$$Y \underset{P}{\leq} Z \quad \Rightarrow \quad E(Y|\mathscr{C}) \underset{P}{\leq} E(Z|\mathscr{C}). \tag{i}$$

$$Y \underset{P}{\geq} 0 \text{ and } E(Y) = 0 \quad \Rightarrow \quad E(Y|\mathscr{C}) \underset{P}{=} 0. \tag{ii}$$

$$Y \underset{P}{\geq} \alpha \quad \Rightarrow \quad E(Y|\mathscr{C}) \underset{P}{\geq} \alpha. \tag{iii}$$

$$Y \underset{P}{\leq} \alpha \quad \Rightarrow \quad E(Y|\mathscr{C}) \underset{P}{\leq} \alpha. \tag{iv}$$

$$Y \underset{P}{=} \alpha \quad \Rightarrow \quad E(Y|\mathscr{C}) \underset{P}{=} \alpha. \tag{v}$$

$$Y \underset{P}{\geq} \alpha \quad \Rightarrow \quad \exists\, V \in \mathscr{E}(Y|\mathscr{C})\colon V \geq \alpha. \tag{vi}$$

$$Y \underset{P}{=} \alpha \quad \Rightarrow \quad \exists\, V \in \mathscr{E}(Y|\mathscr{C})\colon V = \alpha. \tag{vii}$$

$$Y \underset{P}{\leq} \alpha \quad \Rightarrow \quad \exists\, V \in \mathscr{E}(Y|\mathscr{C})\colon V \leq \alpha. \tag{viii}$$

$$Y \underset{P}{=} \alpha \quad \Rightarrow \quad E(1_{Y=\alpha}|\mathscr{C}) \underset{P}{=} P(Y=\alpha|\mathscr{C}) \underset{P}{=} 1. \tag{ix}$$

If $A \in \mathscr{A}$ with $P(A) = 0$, then

$$E(1_A|\mathscr{C}) \underset{P}{=} P(A|\mathscr{C}) \underset{P}{=} 0. \tag{x}$$

$$E(1_{A^c}|\mathscr{C}) \underset{P}{=} 1 - E(1_A|\mathscr{C}) \underset{P}{=} P(A^c|\mathscr{C}) \underset{P}{=} 1. \tag{xi}$$

Theorem 10.21 (Monotone Convergence)
Let $Y, Y_i\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, $i \in \mathbb{N}$, be random variables.

(i) If the sequence $(Y_i, i \in \mathbb{N})$, is increasing with $\lim_{i\to\infty} Y_i \underset{P}{=} Y$, and Y_i is nonnegative $\forall\, i \in \mathbb{N}$, then

$$\lim_{i\to\infty} E(Y_i|\mathscr{C}) \underset{P}{=} E(Y|\mathscr{C}). \tag{10.19}$$

(ii) If $Y_i \geq 0$, $\forall\, i \in \mathbb{N}$, then

$$E\Big(\sum_{i=1}^{\infty} Y_i \,\Big|\, \mathscr{C}\Big) \underset{P}{=} \sum_{i=1}^{\infty} E(Y_i|\mathscr{C}). \tag{10.20}$$

(Proof p. 312)

If the sequence $(Y_i, i \in \mathbb{N})$ is not increasing, then an additional assumption is necessary in order to guarantee convergence of the conditional expectations. A

sufficient condition is that all $|Y_i|$ are dominated by the same P-integrable function Z.

Theorem 10.22 (Dominated Convergence)
Let $Y_i\colon (\Omega,\mathcal{A},P) \to (\overline{\mathbb{R}},\overline{\mathcal{B}})$, $i \in \mathbb{N}$, and $Z\colon (\Omega,\mathcal{A},P) \to (\overline{\mathbb{R}},\overline{\mathcal{B}})$ be random variables. If

(a) Z has a finite expectation and $|Y_i| \le Z$, $\forall\, i \in \mathbb{N}$,
(b) $Y\colon (\Omega,\mathcal{A},P) \to (\overline{\mathbb{R}},\overline{\mathcal{B}})$ is a random variable such that $\lim_{i\to\infty} Y_i \underset{P}{=} Y$,

then

$$\lim_{i\to\infty} E(Y_i|\mathscr{C}) \underset{P}{=} E(Y|\mathscr{C}). \tag{10.21}$$

For a proof see Klenke (2008, Theorem 8.14 (viii), p. 174).

10.4 Factorization, Regression, and Conditional Expectation Value

A concept closely associated with an X-conditional expectation $E(Y|X)$ of Y is a *factorization of $E(Y|X)$*, which is synonymously called a *regression of Y on X*. This concept is also used for a general definition of a conditional expectation value.

10.4.1 Existence of a Factorization

Lemma 2.52 can be applied to $E(Y|X)$, which, by definition, is measurable with respect to X. This immediately implies the following corollary:

Corollary 10.23 (Existence of a Factorization of a Conditional Expectation)
Let $X\colon (\Omega,\mathcal{A},P) \to (\Omega'_X,\mathcal{A}'_X)$ and $Y\colon (\Omega,\mathcal{A},P) \to (\overline{\mathbb{R}},\overline{\mathcal{B}})$ be random variables, where Y is nonnegative or with finite expectation. If $E(Y|X) \in \mathscr{E}(Y|X)$, then there is a measurable function $g\colon (\Omega'_X,\mathcal{A}'_X) \to (\overline{\mathbb{R}},\overline{\mathcal{B}})$ such that

$$E(Y|X) = g\circ X. \tag{10.22}$$

Remark 10.24 (Notation) Instead of $g\circ X$ we also use the notation $g(X)$. Figure 9.1 displays the random variable X, a factorization g, and a version $E(Y|X) = g(X) \in \mathscr{E}(Y|X)$. ◁

Definition 10.25 (Regression)
Under the assumptions of Corollary 10.23, the function $g\colon \Omega'_X \to \overline{\mathbb{R}}$ is also

called a factorization of $E(Y|X)$ or, synonymously, a regression of Y on X. Furthermore, Y is called the regressand and X the regressor.

Note that the definition of a regression does not require that the regressor is real-valued nor does it refer to any parametric function. In contrast, these requirements were made in the definition of a linear quasi-regression (see 7.2). The relationship between the regression and the linear quasi-regression is considered in the following section.

10.4.2 Conditional Expectation and Mean-Squared Error

In Definition 7.2 we introduced the linear quasi-regression as that linear function $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \alpha_0 + \alpha_1 x$, $x \in \mathbb{R}$, that minimizes $E\big([Y - (a_0 + a_1 X)]^2\big)$, where X is a real-valued random variable. In a sense, the linear quasi-regression is a function f such that the composition $Q_{lin}(Y|X) = f(X)$ is the best approximation of Y by a *linear function* of X. Now consider the approximation (with respect to the mean squared error) of Y by a more general function that is $\mathscr{C}$-measurable or X-measurable, respectively. Intuitively speaking, we ask for the best approximation of Y based on the information contained in $\mathscr{C}$ or in X.

Reading the following theorem, note that the right-hand sides of (10.23) and (10.24) do not depend on the particular choice of a version $E(Y|X) \in \mathscr{E}(Y|X)$ [see Th. 10.9 (ii) and Rule (viii) of Box 6.1].

Theorem 10.26 (Conditional Expectation and Mean-Squared Error)
Let $Y, Z: (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be real-valued random variables with $E(Y^2)$, $E(Z^2) < \infty$, let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, and suppose that Z is $\mathscr{C}$-measurable. Then

$$E[(Y - Z)^2] \;\geq\; E\big([Y - E(Y|\mathscr{C})]^2\big). \tag{10.23}$$

and

$$Z \underset{P}{=} E(Y|\mathscr{C}) \;\Leftrightarrow\; E[(Y - Z)^2] \;=\; E\big([Y - E(Y|\mathscr{C})]^2\big). \tag{10.24}$$

For a proof see Klenke (2008, p. 176, Cor. 8.16).

Remark 10.27 (Regression vs. Linear Quasi-Regression) If $E(Y^2) < \infty$, then Theorem 10.26 implies that $Z \in \mathscr{E}(Y|X)$ is an X-measurable random variable with $E(Z^2) < \infty$ minimizing the function $E[(Y - Z)^2]$. Vice versa, if Z is an X-measurable random variable with finite second moment minimizing the mean squared error $E[(Y - Z)^2]$, then $Z \in \mathscr{E}(Y|X)$, provided that $E(Y^2) < \infty$.

In contrast, $Q_{lin}(Y|X)$ is a *linear function* of X minimizing $E\big([Y - (a_0 + a_1 X)]^2\big)$, $a_0, a_1 \in \mathbb{R}$, provided that $E(Y^2) < \infty$. Hence, $Q_{lin}(Y|X)$ is the best (with respect to the mean squared error) approximation of Y in the set of all linear functions of

X, whereas $E(Y|X)$ is the best approximation of Y in the set of all X-measurable functions, provided that the second moment of Y is finite.

According to Definition 7.2, $Q_{lin}(Y|X) = f(X)$, where $f:\mathbb{R}\to\mathbb{R}$ defined by $f(x) = \alpha_0 + \alpha_1 x$, $x\in\mathbb{R}$, is the linear quasi-regression of Y on X. In contrast, the regression of Y on X is a function $g\colon \Omega'_X \to \overline{\mathbb{R}}$ such that there is a version $E(Y|X)\in \mathscr{E}(Y|X)$ with $E(Y|X) = g(X)$. Even if we consider a real-valued random variable X such that $\Omega'_X = \mathbb{R}$, a regression g does not require that it is a linear function. If g is a linear function with domain $\Omega'_X = \mathbb{R}$ and there is a version $E(Y|X)\in\mathscr{E}(Y|X)$ with $E(Y|X) = g(X)$, then $f = g$ and $Q_{lin}(Y|X) = E(Y|X)$. ◁

10.4.3 Uniqueness of a Factorization

A factorization of $E(Y|X)$ or synonymously, a regression of Y on X, is not necessarily uniquely defined. This even applies if we consider a fixed version $E(Y|X)$ of the conditional expectation.

Remark 10.28 (Uniqueness of a Factorization) For two elements V and V^* of $\mathscr{E}(Y|X)$ there can be different factorizations g and g^* with $V = g(X)$ and $V^* = g^*(X)$. This is true even if $V = V^*$. Hence, there can be different factorizations of a single element $V\in\mathscr{E}(Y|X)$ (see Example 10.32). In other words, $V = g(X) = g^*(X)$, with $g\neq g^*$ is not necessarily contradictory. In this case $g(x) = g^*(x)$ for all $x\in X(\Omega)$, whereas $g(x) = g^*(x)$ does *not* hold for all $x\in\Omega'_X$. However, Corollary 5.21 (i) implies the following corollary: ◁

Corollary 10.29 (P_X-Equivalence of Factorizations)
Let the assumptions 10.1 hold and let $g, g^*\colon (\Omega'_X,\mathscr{A}'_X)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be $(\mathscr{A}'_X,\overline{\mathscr{B}})$-measurable functions. If $g(X), g^*(X)\in\mathscr{E}(Y|X)$, then

$$g \underset{P_X}{=} g^*. \tag{10.25}$$

Remark 10.30 (P_X-Equivalence) Note that Equation (10.25) is equivalent to

$$g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x\in\Omega'_X. \tag{10.26}$$

◁

According to Remark 10.16, $P(X{=}x) > 0$ for all $x\in X(\Omega)$ implies that $E(Y|X)$ is uniquely defined. According to the following corollary, this also applies to the factorization of $E(Y|X)$ if we additionally assume $\Omega'_X = X(\Omega)$.

Corollary 10.31 (Uniqueness of the Factorization)
Let the assumptions 10.1 hold and assume $\Omega'_X = X(\Omega)$ with $P(X{=}x) > 0$ for all $x\in\Omega'_X$. Then the function $g\colon\Omega'_X\to\overline{\mathbb{R}}$ satisfying $E(Y|X) = g(X)$ is uniquely defined.

Example 10.32 (No Treatment for Joe – continued) In Example 9.22 we specified $E(Y|X)$ with its two values $E(Y|X=0) = .60$ and $E(Y|X=1) = .40$. If we consider the treatment variable $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, then $g\colon \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} .60, & \text{if } x = 0, \\ .40, & \text{if } x = 1, \\ \alpha, \quad \alpha \in \mathbb{R}, & \text{otherwise,} \end{cases}$$

is a factorization of the conditional expectation of Y given X for any choice of α. Note that the value $g(x)$ can be any real number for $x \notin \{0,1\}$. This implies that there are different factorizations g, g^* with $g(X), g^*(X) \in \mathscr{E}(Y|X)$. However, g and g^* are P_X-equivalent. In contrast, if we consider $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$, with $\Omega'_X = X(\Omega) = \{0,1\}$ and $\mathscr{A}'_X = \mathscr{P}(\Omega'_X)$, then there is only one single factorization $g\colon \Omega'_X \to \mathbb{R}$ with the two values $g(0) = .60$ and $g(1) = .40$ (see Cor. 10.31). ◁

10.4.4 Conditional Expectation Value

The concepts of a conditional expectation value $E(Y|X=x)$ and a conditional probability $P(A|X=x)$ have been introduced in Definition 9.2 and Remark 9.6 only for $P(X=x) > 0$. Now we drop this assumption and define these concepts more generally, again using the factorization of a conditional expectation.

Definition 10.33 ($(X=x)$-Conditional Expectation Value)
Let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable, let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be nonnegative or with finite expectation $E(Y)$, and let $g\colon (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a function satisfying (10.22)*. Then the value $g(x)$ of g is called an $(X=x)$-conditional expectation value of Y and is denoted by $E(Y|X=x)$, i.e.,*

$$E(Y|X=x) := g(x). \tag{10.27}$$

Remark 10.34 ($(X=x)$-Conditional Probability) If 1_A is the indicator of $A \in \mathscr{A}$, then $E(1_A|X=x)$ is also called an $(X=x)$-*conditional probability* of A and it is denoted by $P(A|X=x)$, i. e.,

$$P(A|X=x) := E(1_A|X=x). \tag{10.28}$$

Furthermore, considering the event $\{Y=y\}$, we also use the notation

$$P(Y=y|X=x) := P(\{Y=y\}|X=x) = E(1_{Y=y}|X=x). \tag{10.29}$$

◁

Remark 10.35 (Uniqueness and Consistency of Definitions) If $P(X=x) > 0$, then the $(X=x)$-conditional expectation value of Y is uniquely defined and it is identical to the term introduced in Definition 9.2, i. e.,

$$E(Y|X{=}x) \;=\; E^{X=x}(Y), \quad \text{if } P(X{=}x) > 0 \tag{10.30}$$

(see Exercise 10-7). In the general case, $E(Y|X{=}x)$ is not uniquely defined. However, g is uniquely defined up to P_X-equivalence (see Cor. 10.29). ◁

Remark 10.36 (Versions of a Conditional Expectation With a Discrete X) Let $Y\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a random variable that is nonnegative or with finite expectation and suppose that $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ is discrete implying that $\Omega'_0 \subset \Omega'_X$ is finite or countable with $P_X(\Omega'_0) = 1$ and $\{x\} \in \mathscr{A}'_X$ for all $x \in \Omega'_0$. Then, for all $\alpha \in \mathbb{R}$, the function $E(Y|X)\colon \Omega \to \mathbb{R}$ defined by

$$E(Y|X)(\omega) := \begin{cases} E(Y|X{=}x), & \text{if } X(\omega) = x \text{ and } P(X{=}x) > 0 \\ \alpha, & \text{otherwise.} \end{cases} \tag{10.31}$$

is a version $E(Y|X) \in \mathscr{E}(Y|X)$. This proposition follows from Remark 10.35 and proposition (10.9). ◁

Remark 10.37 (Values of the Conditional Expectation) Assume that $E(Y|X) = g(X) \in \mathscr{E}(Y|X)$. Then

$$E(Y|X)(\omega) \;=\; g(x) \;=\; E(Y|X{=}x), \quad \forall\, \omega \in \Omega\colon X(\omega) = x. \tag{10.32}$$

(see Exercise 10-8). This also implies that the value of $E(Y|X)$ is constant on all sets $\{X{=}x\} = \{\omega \in \Omega\colon X(\omega) = x\}$. Note that this also holds if Ω is finite or countable and some $\omega \in \{X{=}x\}$ have probability $P(\{\omega\}) = 0$. As an example, see $E(Y|X)(\omega)$ for $\omega \in \{(\textit{Joe}, \textit{yes}, -), (\textit{Joe}, \textit{yes}, +)\}$ in Table 9.2. These two values are equal to $E(Y|X{=}1) = .40$, although $P(\{(\textit{Joe}, \textit{yes}, -)\}) = P(\{(\textit{Joe}, \textit{yes}, +)\}) = 0$. ◁

Remark 10.38 (Equivalent Propositions) Let the assumptions 10.1 hold, let $g(X) \in \mathscr{E}(Y|X)$ and $g^*\colon (\Omega'_X,\mathscr{A}'_X) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ be an $(\mathscr{A}'_X,\overline{\mathscr{B}})$-measurable function. Then proposition (10.9) and Theorem 2.49 yield

$$g(X) \underset{P}{=} g^*(X) \quad \Leftrightarrow \quad g^*(X) \in \mathscr{E}(Y|X)\,. \tag{10.33}$$

◁

Definition 10.33, Corollary 5.21, Remark 5.22, and Equation (10.22) imply the following corollary, according to which we may either formulate propositions in terms of $(X{=}x)$-conditional expectation values or, equivalently, in terms of the corresponding conditional expectations.

Corollary 10.39 (Equivalent Propositions)
Let $Y_1, Y_2\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ be two numerical random variables that are nonnegative or with finite expectations and let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable. Then:

(i) $E(Y_1|X) \underset{P}{=} E(Y_2|X)$ *is equivalent to*

$$E(Y_1|X=x) = E(Y_2|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \tag{10.34}$$

(ii) $E(Y_1|X) \underset{P}{>} E(Y_2|X)$ *is equivalent to*

$$E(Y_1|X=x) > E(Y_2|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \tag{10.35}$$

(iii) $E(Y_1|X) \underset{P}{\geq} E(Y_2|X)$ *is equivalent to*

$$E(Y_1|X=x) \geq E(Y_2|X=x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \tag{10.36}$$

Remark 10.40 (($X=x$)-Conditional Expectation Value of $E(Y|X)$) Suppose that $f\colon (\Omega'_X, \mathscr{A}'_X) \to (\Omega'', \mathscr{A}'')$ is a measurable mapping, $w \in \Omega''$, $\{w\} \in \mathscr{A}''$, and $C = \{f(X)=w\}$, then Equation (10.6) can be written

$$\begin{gathered} E\big[E(Y|X) \,\big|\, f(X)=w\big] = E[Y|f(X)=w], \\ \text{if } w \in \Omega'' \text{ and } P[f(X)=w] > 0. \end{gathered} \tag{10.37}$$

As a special case, this equation implies

$$E[E(Y|X,Z)|X=x] = E(Y|X=x), \quad \text{if } x \in \Omega'_X \text{ with } P(X=x) > 0, \tag{10.38}$$

provided that $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ is a random variable, too. If Z takes on only a finite number of values $z_1, \ldots, z_m$ and $P(X=x, Z=z_i) > 0$, for all $i = 1, \ldots, m$, then Rule (ii) of Box 9.2 follows from Equations (10.38) and (9.20), which is consistent with our results already obtained in Exercise 9-2. Similarly, applying (9.19), Equation (10.38) yields Rule (iii) of Box 9.2 (see Exercise 10-9). ◁

Generalizing this result, proposition (i) of Corollary 10.39 and Rule (v) of Box 10.2 yield:

$$E[E(Y|X)|f(X)=w] = E[Y|f(X)=w], \quad \text{for } P_{f(X)}\text{-a.a } w \in \Omega''. \tag{10.39}$$

A special case is

$$E[E(Y|X,Z)|X=x] = E(Y|X=x), \quad \text{for } P_X\text{-a.a } x \in \Omega'_X. \tag{10.40}$$

10.5 Characterizing a Conditional Expectation by the Joint Distribution

Using the factorization and Equation (3.59) yields two conditions that are equivalent to those occurring in Definition 10.2. In these conditions we refer to the joint distribution $P_{X,Y}$ of X and Y.

Theorem 10.41 (Conditional Expectation and the Joint Distribution)
Let the assumptions 10.1 hold. Then $g(X) \in \mathscr{E}(Y|X)$ if and only if

(a) $g\colon (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is measurable.
(b) for all $C' \in \mathscr{A}'_X$,

$$\int 1_{C'}(x) \cdot g(x)\ P_X(dx) = \int 1_{C'}(x) \cdot y\ P_{X,Y}[d(x,y)]. \tag{10.41}$$

(Proof p. 312)

Remark 10.42 (Two Alternative Formulations) According to (3.28) and Equation (3.59), condition (b) of Theorem 10.41 is equivalent to

$$\int_{C'} g\, dP_X = \int_{\{X \in C'\}} Y\, dP, \quad \forall\, C' \in \mathscr{A}'_X. \tag{10.42}$$

If $E_X(\cdot)$ denotes the expectation with respect to the distribution P_X and $1_{X \in C'}$ denotes the indicator of $\{X \in C'\}$, then Equation (10.42) may also be written:

$$E_X(1_{C'}\, g) = E(1_{X \in C'}\, Y), \quad \forall\, C' \in \mathscr{A}'_X. \tag{10.43}$$

◁

Remark 10.43 (Conditional Expectation With Respect to a Joint Distribution) Note that in Equation (10.41) we do not explicitly refer to the measure P. Instead we refer to $P_{X,Y}$, the joint distribution of X and Y. Therefore $g(X) \in \mathscr{E}(Y|X)$ may also be called a version of the *conditional expectation* of Y on X with respect to $P_{X,Y}$. This can be used, e. g., to consider a conditional expectation with respect to the $(Z{=}z)$-conditional distribution $P_{X,Y|Z=z}$ (see Def. 17.8). ◁

10.6 Conditional Mean Independence

In section 5.4 we defined *independence* of two random variables X and Y. Furthermore, in chapter 7, introducing the covariance $Cov(X,Y)$ and the correlation $Corr(X,Y)$ of two numerical random variables, we also defined *correlational independence* by $Cov(X,Y) = 0$. Now we add two other related concepts: *mean independence* and *conditional mean independence*.

Definition 10.44 ($\mathscr{C}$-Conditional Mean Independence)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a numerical random variable that is nonnegative or has a finite expectation $E(Y)$ and let $\mathscr{D} \subset \mathscr{A}$ be a σ-algebra.

(i) Then Y is called mean independent from $\mathscr{D}$, if

$$E(Y|\mathscr{D}) \underset{P}{=} E(Y). \tag{10.44}$$

(ii) Let also $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra and let $E(Y|\mathscr{C},\mathscr{D})$ denote the conditional expectation of Y given $\sigma(\mathscr{C}\cup\mathscr{D})$. Then Y is called $\mathscr{C}$-conditionally mean independent from $\mathscr{D}$, if

$$E(Y|\mathscr{C},\mathscr{D}) \underset{P}{=} E(Y|\mathscr{C}). \tag{10.45}$$

Remark 10.45 (X- Conditional Mean Independence) Let $Y\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a numerical random variable that is nonnegative or has a finite expectation $E(Y)$ and let $Z\colon (\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ be a random variable.

(i) Then Y is called *mean independent from Z* if

$$E(Y|Z) \underset{P}{=} E(Y). \tag{10.46}$$

(ii) Let also $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable. Then Y is called *X-conditionally mean independent from Z* if

$$E(Y|X,Z) \underset{P}{=} E(Y|X). \tag{10.47}$$

◁

Remark 10.46 (A Special Case) Of course, if $\mathscr{D}\subset\mathscr{C}$, then Y is $\mathscr{C}$-conditionally mean independent from $\mathscr{D}$. In this case, $\sigma(\mathscr{C}\cup\mathscr{D}) = \sigma(\mathscr{C}) = \mathscr{C}$ and $E(Y|\mathscr{C},\mathscr{D})$ is just a different notation of $E(Y|\mathscr{C})$. Correspondingly, assume that Z is measurable with respect to X, i. e., $\sigma(Z)\subset\sigma(X)$. Then $\sigma(X,Z)=\sigma(X)$ and therefore,

$$E(Y|X,Z) \underset{P}{=} E(Y|X). \tag{10.48}$$

Hence, Y is X-conditionally mean independent from all random variables Z that are measurable with respect to X. In these cases Z does not carry any information additional to X. In more formal terms, Z does not represent any event that is not already represented by X, i. e., $\{Z\in A'\}\in\sigma(X)$, for all $A'\in\mathscr{A}'_Z$. ◁

Example 10.47 (Joe and Ann With no Treatment Effect) Table 10.1 displays an example for U-conditional mean independence of Y from X, i. e.,

$$E(Y|X,U) \underset{P}{=} E(Y|U).$$

The values of the conditional expectations $E(Y|X,U)$ and $E(Y|U)$ can be computed in the same way as in Example 9.20. This new example shows that $E(Y|X,U) \underset{P}{=} E(Y|U)$ does not imply $E(Y|X) \underset{P}{=} E(Y)$. Hence, although $E(Y|X) \underset{P}{=} E(Y)$ *does not hold* and the conditional expectation values $E(Y|X{=}x)$ *do depend* on the values x of X, in a sense, the treatment variable X is irrelevant once we condition on U. In other words, for *Joe* success does not depend on whether or not he receives treatment, and the same is true for *Ann* [see the column headed $E(Y|X,U)$]. This example shows that the conditional expectation $E(Y|X)$ is completely misleading if used for the evaluation of the effect of the treatment variable X on the outcome variable Y. ◁

Table 10.1. Joe and Ann With no Treatment Effects

	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\mid X,U)$	$E(Y\mid U)$	$E(Y\mid X)$	$P(X=1\mid U)$
(*Joe, no,* −)	0.16	*Joe*	0	0	0.2	0.2	0.56	0.6
(*Joe, no,* +)	0.04	*Joe*	0	1	0.2	0.2	0.56	0.6
(*Joe, yes,* −)	0.24	*Joe*	1	0	0.2	0.2	0.44	0.6
(*Joe, yes,* +)	0.06	*Joe*	1	1	0.2	0.2	0.44	0.6
(*Ann, no,* −)	0.06	*Ann*	0	0	0.8	0.8	0.56	0.4
(*Ann, no,* +)	0.24	*Ann*	0	1	0.8	0.8	0.56	0.4
(*Ann, yes,* -)	0.04	*Ann*	1	0	0.8	0.8	0.44	0.4
(*Ann, yes,* +)	0.16	*Ann*	1	1	0.8	0.8	0.44	0.4

Remark 10.48 (Implication Structure Among Different Kinds of Independence) According to Rule (vi) of Box 10.1, independence of Y and $\mathscr{C}$ implies that Y is mean independent from $\mathscr{C}$. Analogously, according to Rule (vi) of Box 10.2 independence of Y and X implies that Y is mean independent from X. Furthermore, mean independence of Y from X implies that X and Y are uncorrelated, provided that X and Y are numerical and $E(X^2), E(Y^2) < \infty$ (see Exercise 10-10). Hence, if $E(X^2), E(Y^2) < \infty$, then

$$Y \perp\!\!\!\perp_P X \quad \Rightarrow \quad E(Y\mid X) \underset{P}{=} E(Y) \quad \Rightarrow \quad Corr(X,Y) = Cov(X,Y) = 0. \tag{10.49}$$

◁

Now we turn to conditions that are equivalent to conditional mean independence. We start with a theorem that only applies to a *nonnegative* numerical random variable Y that *also has a finite expectation.* A second theorem also applies to a numerical random variable Y *with finite second moment.*

Theorem 10.49 ($\mathscr{C}$-Conditional Mean Independence I)
Let $Y\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a nonnegative random variable that has a finite expectation $E(Y)$ and let $\mathscr{C},\mathscr{D} \subset \mathscr{A}$ be σ-algebras. Then the following two propositions are equivalent to each other:

(a) $E(Y\mid\mathscr{C},\mathscr{D}) \underset{P}{=} E(Y\mid\mathscr{C})$.

(b) For all random variables $W\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ that are nonnegative and $\mathscr{D}$-measurable,

$$E(W\cdot Y\,|\,\mathscr{C}) \underset{P}{=} E(W\,|\,\mathscr{C})\cdot E(Y\,|\,\mathscr{C}). \tag{10.50}$$

(Proof p. 312)

Remark 10.50 (Mean Independence From a σ-Algebra) For $\mathscr{C} = \{\Omega,\emptyset\}$, Theorem 10.49 immediately yields the following proposition: If $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ is a nonnegative random variable that has a finite expectation $E(Y)$ and $\mathscr{D}\subset\mathscr{A}$ is a σ-algebra, then the following two propositions are equivalent to each other:

(a) $E(Y\,|\,\mathscr{D}) \underset{P}{=} E(Y)$.

(b) For all random variables $W\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ that are nonnegative and $\mathscr{D}$-measurable,

$$E(W\cdot Y) = E(W)\cdot E(Y). \tag{10.51}$$

◁

In Theorem 10.49 we required that Y is nonnegative and has a finite expectation. The implication of conditional mean independence formulated in proposition (i) of the following theorem is *not* restricted to nonnegative random variables Y. Instead we have to assume that Y has a finite second moment.

Theorem 10.51 ($\mathscr{C}$-Conditional Mean Independence II)
Let $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a random variable that has a finite second moment $E(Y^2)$, let $\mathscr{C},\mathscr{D}\subset\mathscr{A}$ be σ-algebras, and consider:

(a) $E(Y|\mathscr{C},\mathscr{D}) \underset{P}{=} E(Y|\mathscr{C})$.

(b) For all random variables $W\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ that are $\mathscr{D}$-measurable and have a finite second moment $E(W^2)$,

$$E(W\cdot Y\,|\,\mathscr{C}) \underset{P}{=} E(W\,|\,\mathscr{C})\cdot E(Y\,|\,\mathscr{C}). \tag{10.52}$$

Then:

(i) (a) implies (b).

(ii) If Y is also nonnegative, then (a) and (b) are equivalent to each other.

(Proof p. 313)

Remark 10.52 (Mean Independence II) For $\mathscr{C} = \{\Omega,\emptyset\}$, Theorem 10.51 immediately yields the following proposition. Let $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a random variable that has a finite second moment $E(Y^2)$, let $\mathscr{D}\subset\mathscr{A}$ be a σ-algebra, and consider:

(a) $E(Y|\mathscr{D}) \underset{P}{=} E(Y)$.

(b) For all random variables $W\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ that are $\mathscr{D}$-measurable and have a finite second moment $E(W^2)$,

$$E(W\cdot Y) = E(W)\cdot E(Y). \tag{10.53}$$

Then:

(i) (a) implies (b).
(ii) If Y is also nonnegative, then (a) and (b) are equivalent to each other.

◁

For $\mathscr{C} = \sigma(X)$ and $\mathscr{D} = \sigma(Z)$, Theorem 10.49 immediately implies the following corollary.

Corollary 10.53 (X-Conditional Mean Independence I)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a nonnegative random variable that has a finite expectation $E(Y)$ and let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$, $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ be random variables. Then the following two propositions are equivalent to each other:

(a) $E(Y|X,Z) \underset{P}{=} E(Y|X)$;
(b) For all random variables $W\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ that are nonnegative and Z-measurable,

$$E(W \cdot Y|X) \underset{P}{=} E(W|X) \cdot E(Y|X). \tag{10.54}$$

Similarly, for $\mathscr{C} = \sigma(X)$ and $\mathscr{D} = \sigma(Z)$, Theorem 10.51 immediately implies the following corollary.

Corollary 10.54 (X-Conditional Mean Independence II)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ be a random variable that has a finite second moment $E(Y^2)$, let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$, $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ be random variables, and consider:

(a) $E(Y|X,Z) \underset{P}{=} E(Y|X)$.
(b) For all random variables $W\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ that are Z-measurable and have a finite second moment $E(W^2)$,

$$E(W \cdot Y|X) \underset{P}{=} E(W|X) \cdot E(Y|X). \tag{10.55}$$

Then:

(i) (a) implies (b).
(ii) If Y is also nonnegative, then (a) and (b) are equivalent to each other.

Remark 10.55 (Mean Independence From a Random Variable) For $X = \alpha, \alpha \in \Omega'_X$, this corollary immediately yields the following proposition. Let $Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ and $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ be random variables with $E(Y^2) < \infty$, and consider:

(a) $E(Y|Z) \underset{P}{=} E(Y)$.

(b) For all random variables $W\colon (\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ that are Z-measurable and have a finite second moment $E(W^2)$,

$$E(W\cdot Y) \;=\; E(W)\cdot E(Y). \tag{10.56}$$

Then:

(i) (a) implies (b).
(ii) If Y is also nonnegative, then (a) and (b) are equivalent to each other.

$\lhd$

Remark 10.56 (Mean Independence and Correlational Independence) For $Z = W$, Proposition (i) of Remark 10.55 yields

$$E(Y|Z) \underset{P}{=} E(Y) \quad\Rightarrow\quad E(Z\cdot Y) \;=\; E(Z)\cdot E(Y), \tag{10.57}$$

provided that $E(Z^2), E(Y^2)<\infty$. Proposition (10.57) is equivalent to the second implication in (10.49), because $E(Z\cdot Y)=E(Z)\cdot E(Y)$ is equivalent to $Cov(Z,Y)=0$, provided that $E(Z^2), E(Y^2)<\infty$ [see Box 7.1 (i)]. However, $Cov(Z,Y)=0$ does *not* imply $E(Y|Z)\underset{P}{=}E(Y)$. In other words, correlational independence does not imply mean independence. $\lhd$

Further properties of conditional mean independence are treated in chapter 16, in particular the relationship between conditional independence and conditional mean independence.

10.7 Proofs

Proof of Theorem 10.17

Under the assumptions about $\mathscr{E}=\{A_1,A_2,\ldots\}$ and $\mathscr{C}$, a function $V\colon(\Omega,\mathscr{A})\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ is $\mathscr{C}$-measurable if and only if there are $\alpha_i\in\overline{\mathbb{R}}$, $i=1,2,\ldots$, such that $V=\sum_{i=1}^{\infty}\alpha_i 1_{A_i}$ (see Lemma 2.19). Hence, if $V,V^*\in\mathscr{E}(Y|\mathscr{C})$, then

$$V=\sum_{i=1}^{\infty}\alpha_i 1_{A_i} \qquad\text{and}\qquad V^*=\sum_{i=1}^{\infty}\alpha_i^* 1_{A_i}, \qquad \alpha_i,\alpha_i^*\in\overline{\mathbb{R}}.$$

This implies

$$P(\{V\neq V^*\}) \;=\; \sum_{i\colon \alpha_i\neq\alpha_i^*} P(A_i).$$

Because $P(\{V\neq V^*\})=0$ (see Th. 10.9 and Def. 2.68), we can conclude

$$\sum_{i\colon \alpha_i\neq\alpha_i^*} P(A_i) \;=\; 0. \tag{10.58}$$

Hence if (10.13) holds, then Equation (10.58) implies that there is no i with $\alpha_i\neq\alpha_i^*$, which implies $\{V\neq V^*\}=\emptyset$. Now assume that there is an $A_i\in\mathscr{E}$ with $P(A_i)=0$. Then there are $V,V^*\in\mathscr{E}(Y|\mathscr{C})$ and an i with $\alpha_i\neq\alpha_i^*$, which implies $\emptyset\neq A_i\subset\{V\neq V^*\}$. By contraposition, this proves that $\{V\neq V^*\}=\emptyset$ implies Equation (10.13).

Proof of Theorem 10.21

(i) This proof is found in Bauer (1996, (15.13), p. 115). Because the sequence $Y_i, i \in \mathbb{N}$, is increasing and the conditional expectation is monotone [see Rule (i) of Box 10.3], we can conclude: $\lim_{i\to\infty} Y_i = \sup_{i\in\mathbb{N}} Y_i$ and $\lim_{i\to\infty} E(Y_i\,|\mathscr{C}) = \sup_{i\in\mathbb{N}} E(Y_i\,|\mathscr{C})$.

(ii) If $Y_i \geq 0$, for all $i \in \mathbb{N}$, then $\tilde{Y}_n := \sum_{i=1}^{n} Y_i$, $n \in \mathbb{N}$, is increasing and $\lim_{n\to\infty} \tilde{Y}_n = \sum_{i=1}^{\infty} Y_i$. Hence,

$$\begin{aligned}
E\Big(\sum_{i=1}^{\infty} Y_i \,\Big|\mathscr{C}\Big) &\underset{P}{=} E\Big(\lim_{n\to\infty} \tilde{Y}_n \,\Big|\mathscr{C}\Big) \\
&\underset{P}{=} \lim_{n\to\infty} E\Big(\tilde{Y}_n \,\Big|\mathscr{C}\Big) && \text{[(10.19)]} \\
&\underset{P}{=} \lim_{n\to\infty} E\Big(\sum_{i=1}^{n} Y_i \,\Big|\mathscr{C}\Big) \\
&\underset{P}{=} \lim_{n\to\infty} \sum_{i=1}^{n} E(Y_i\,|\mathscr{C}) && \text{[Box 10.1 (xvi)]} \\
&\underset{P}{=} \sum_{i=1}^{\infty} E(Y_i\,|\mathscr{C}).
\end{aligned}$$

Proof of Theorem 10.41

According to Lemma 2.52, $g(X)$ is measurable with respect to $\sigma(X)$. Therefore, according to Definition 10.2 (b), we only have to show that

$$E[1_C \cdot g(X)] \;=\; E(1_C \cdot Y), \quad \forall\, C \in \sigma(X), \tag{10.59}$$

and Equation (10.41) are equivalent to each other.

(10.41) $\Leftrightarrow$ (10.59). By definition, $\sigma(X) = \{X^{-1}(C')\colon C' \in \mathscr{A}'_X\}$. Therefore,

$$\begin{aligned}
&\int 1_{C'}(x)\cdot g(x)\; P_X(dx) \;=\; \int 1_{C'}(x)\cdot y\; P_{X,Y}[d(x,y)], \quad \forall\, C' \in \mathscr{A}'_X, \\
\Leftrightarrow\; &\int 1_C \cdot g(X)\; dP \;=\; \int 1_C \cdot Y\; dP, \quad \forall\, C = X^{-1}(C') \in \sigma(X), && \text{[(3.59), (3.28)]} \\
\Leftrightarrow\; &E[1_C \cdot g(X)] \;=\; E(1_C \cdot Y), \quad \forall\, C \in \sigma(X). && \text{[(6.1)]}
\end{aligned}$$

Proof of Theorem 10.49

(a) $\Rightarrow$ (b) If Y and W a nonnegative, then $Y \cdot W$ is nonnegative as well and, for $\mathscr{D}$-measurable W,

$$\begin{aligned}
E(W\cdot Y\,|\mathscr{C}) &\underset{P}{=} E[E(W\cdot Y\,|\mathscr{C},\mathscr{D})\,|\mathscr{C}] && \text{[Box 10.1, (v)]} \\
&\underset{P}{=} E[W\cdot E(Y\,|\mathscr{C},\mathscr{D})\,|\mathscr{C}] && \text{[Box 10.1, (xiv)]} \\
&\underset{P}{=} E[W\cdot E(Y\,|\mathscr{C})\,|\mathscr{C}] && \text{[(a)]} \\
&\underset{P}{=} E(Y\,|\mathscr{C})\cdot E(W\,|\mathscr{C}). && \text{[Box 10.1, (xiv)]}
\end{aligned} \tag{10.60}$$

(b) $\Rightarrow$ (a)

$$\begin{aligned} E[W\cdot E(Y\,|\,\mathscr{C},\mathscr{D})\,|\,\mathscr{C}] &\underset{P}{=} E[E(W\cdot Y\,|\,\mathscr{C},\mathscr{D})\,|\,\mathscr{C}] && \text{[Box 10.1, (xiv)]}\\ &\underset{P}{=} E(W\cdot Y\,|\,\mathscr{C}) && \text{[Box 10.1, (v)]}\\ &\underset{P}{=} E(Y\,|\,\mathscr{C})\cdot E(W\,|\,\mathscr{C}) && \text{[(b)]}\\ &\underset{P}{=} E[W\cdot E(Y\,|\,\mathscr{C})\,|\,\mathscr{C}]. && \text{[Box 10.1, (xiv)]} \end{aligned} \tag{10.61}$$

Choosing in this equation $W = 1_D$, $D \in \mathscr{D}$, Definition 10.2 (b) yields,

$$E[1_C 1_D E(Y\,|\,\mathscr{C},\mathscr{D})] = E[1_C 1_D E(Y\,|\,\mathscr{C})], \quad \forall C \in \mathscr{C},$$

which equivalent to

$$\int 1_{C\cap D}\, E(Y\,|\,\mathscr{C},\mathscr{D})\, dP = \int 1_{C\cap D}\, E(Y\,|\,\mathscr{C})\, dP, \quad \forall C \in \mathscr{C}. \tag{10.62}$$

The set $\{C\cap D\colon C\in\mathscr{C}, D\in\mathscr{D}\}$ is $\cap$-stable and generates $\sigma(\mathscr{C}\cup\mathscr{D})$. Furthermore, finiteness of $E(Y) = E[E(Y\,|\,\mathscr{C},\mathscr{D})] = E[E(Y\,|\,\mathscr{C},)]$ [see Box 10.1, (iv)] implies that $E(Y\,|\,\mathscr{C},\mathscr{D})$ and $E(Y\,|\,\mathscr{C},)$ are integrable with respect to P. Hence, according to Theorem 3.68 (iv), we can conclude $E(Y\,|\,\mathscr{C},\mathscr{D}) \underset{P}{=} E(Y\,|\,\mathscr{C})$.

Proof of Theorem 10.51

(a) $\Rightarrow$ (b) If $E(Y^2), E(W^2) < \infty$, then $E(Y)$ and $E(Y\cdot W)$ are finite as well. Then, for $\mathscr{D}$-measurable W, (10.60) also applies to this case.

(b) $\Rightarrow$ (a)

$$\begin{aligned} E[W\cdot E(Y\,|\,\mathscr{C},\mathscr{D})\,|\,\mathscr{C}] &\underset{P}{=} E[E(W\cdot Y\,|\,\mathscr{C},\mathscr{D})\,|\,\mathscr{C}] && \text{[Box 10.1, (xiv)]}\\ &\underset{P}{=} E(W\cdot Y\,|\,\mathscr{C}) && \text{[Box 10.1, (v)]}\\ &\underset{P}{=} E(Y\,|\,\mathscr{C})\cdot E(W\,|\,\mathscr{C}) && \text{[(b)]}\\ &\underset{P}{=} E[W\cdot E(Y\,|\,\mathscr{C})\,|\,\mathscr{C}]. && \text{[Box 10.1, (xiv)]} \end{aligned}$$

Choosing in this equation $W = 1_D$, $D \in \mathscr{D}$, Definition 10.2 (b) yields,

$$E[1_C 1_D E(Y\,|\,\mathscr{C},\mathscr{D})] = E[1_C 1_D E(Y\,|\,\mathscr{C})], \quad \forall C \in \mathscr{C},$$

which equivalent to

$$\int 1_{C\cap D}\, E(Y\,|\,\mathscr{C},\mathscr{D})\, dP = \int 1_{C\cap D}\, E(Y\,|\,\mathscr{C})\, dP, \quad \forall C \in \mathscr{C}. \tag{10.63}$$

The set $\{C\cap D\colon C\in\mathscr{C}, D\in\mathscr{D}\}$ is $\cap$-stable and generates $\sigma(\mathscr{C}\cup\mathscr{D})$. Furthermore, finiteness of $E(Y) = E[E(Y\,|\,\mathscr{C},\mathscr{D})] = E[E(Y\,|\,\mathscr{C},)]$ [see Box 10.1, (iv)] implies that $E(Y\,|\,\mathscr{C},\mathscr{D})$ and $E(Y\,|\,\mathscr{C},)$ are integrable with respect to P. Hence, if additionally $Y \geq 0$, then there are versions $E(Y\,|\,\mathscr{C}), E(Y\,|\,\mathscr{C},\mathscr{D}) \geq 0$ [see Box 10.3 vi] and, according to Theorem 3.68 (iv), we can conclude $E(Y\,|\,\mathscr{C},\mathscr{D}) \underset{P}{=} E(Y\,|\,\mathscr{C})$.

10.8 Exercises

▷ **Exercise 10-1** Table 9.2 (p. 282) presents an element, say V, of $\mathscr{E}(Y|X,U)$. Define an alternative element $V^* \in \mathscr{E}(Y|X,U)$ and show that the two elements are P-equivalent.

▷ **Exercise 10-2** Show that, according to Definition 10.2, Equations (9.23) and (9.24) define an element of $\mathscr{E}(Y|X)$ provided that the assumptions of Definition 9.13 hold.

▷ **Exercise 10-3** Prove Equation (10.14) and that $\sum_{A_i\in\mathscr{E}} E(Y|A_i)\cdot 1_{A_i}$ is a version of $E(Y|\mathscr{C})$, provided that the assumptions of Theorem 10.17 and Equation (10.13) hold.

▷ **Exercise 10-4** Consider Table 9.2 (p. 282) and compute the expectation of $1_{X=0}\cdot E(Y|X)$.

▷ **Exercise 10-5** Prove the propositions of Box 10.1.

▷ **Exercise 10-6** Prove the propositions of Box 10.3.

▷ **Exercise 10-7** Show that $P(X{=}x)>0$ implies that the $(X{=}x)$-conditional expectation of Y defined by Equation (10.27) is uniquely defined and identical to the term introduced in Definition 9.2.

▷ **Exercise 10-8** Prove Equation (10.32).

▷ **Exercise 10-9** Show that Equation (10.38) implies Rule (iii) of Box 9.2.

▷ **Exercise 10-10** Show that mean independence of Y from X implies that X and Y are uncorrelated, provided that the second moments of X and Y are finite.

Solutions

▷ **Solution 10-1** Another element $V^*\in\mathscr{E}(Y|X,U)$ is obtained defining

$$V^*(\omega) \;=\; \begin{cases} 9, & \text{if } \omega=\omega_3 \text{ or } \omega=\omega_4 \\ V(\omega), & \text{if } \omega\in\Omega, \omega\neq\omega_3, \omega\neq\omega_4, \end{cases}$$

where $\omega_3=(\textit{Joe},\textit{yes},-)$ and $\omega_4=(\textit{Joe},\textit{yes},+)$. For V and V^*, $P(A_1)=1$, where $A_1=\{\omega\in\Omega: V(\omega)=V^*(\omega)\}$. The probability $P(A_1)=1$ is obtained from adding the probabilities of all six outcomes ω for which $P(\{\omega\})>0$ (see the second column of Table 9.2, p. 282).

▷ **Solution 10-2** Let $\Omega_0'\subset\Omega_X'$ denote the finite or countable set introduced in Definitions 9.13 (i) or (ii). Then $\{X{=}x\}\in\sigma(X)$ for all $x\in\Omega_0'$. This implies:
(a) For all $x\in\Omega_0'$, the indicator $1_{X=x}$ is X-measurable (see Example 2.12), which implies that $\sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{X=x}$ is X-measurable as well (see Lemma 2.19).
(b) For $C\in\sigma(X)$, define $C_0:=C\cap X^{-1}(\Omega_0')$. Because, by definition, $P[X^{-1}(\Omega_0')]=P_X(\Omega_0')=1$, this implies $P(C\setminus C_0)=0$ and

$$\begin{aligned} & E\Big(1_C\cdot\sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{X=x}\Big) \\ &= E\Big(1_{C_0}\cdot\sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{X=x}\Big) + E\Big(1_{C\setminus C_0}\cdot\sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{X=x}\Big) \qquad \text{[Box 6.1 (vi)]} \\ &= E\Big(\sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{C_0}\cdot 1_{X=x}\Big). \qquad \text{[Box 6.1 (iii)]} \end{aligned}$$

Furthermore, denote $C_0':=X(C_0)=\{x\in\Omega_0': X^{-1}(\{x\})\subset C\}$. Then

$$
\begin{aligned}
E\Big(1_C \cdot \sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{X=x}\Big) &= E\Big(\sum_{x\in\Omega_0'} E(Y|X{=}x)\cdot 1_{C_0}\cdot 1_{X=x}\Big)\\
&= \sum_{x\in\Omega_0'} E\left[E(Y|X{=}x)\cdot 1_{C_0}\cdot 1_{X=x}\right] && \text{[Box 6.1 (vii)]}\\
&= \sum_{x\in C_0'} E\left[E(Y|X{=}x)\cdot 1_{X=x}\right] && \text{[def. of } C_0\text{]}\\
&= \sum_{x\in C_0'} E(Y|X{=}x)\cdot E(1_{X=x}) && \text{[Box 6.1 (iii)]}\\
&= \sum_{x\in C_0'} E(Y|X{=}x)\cdot P(X{=}x) && \text{[(6.4)]}\\
&= \sum_{x\in C_0'} \Big(\frac{1}{P(X{=}x)}\cdot E(1_{X=x}\cdot Y)\Big)\cdot P(X{=}x) && \text{[(9.11)]}\\
&= \sum_{x\in C_0'} E(1_{X=x}\cdot Y)\\
&= E\Big(\sum_{x\in C_0'} 1_{X=x}\cdot Y\Big) && \text{[Box 6.1 (vii)]}\\
&= E(1_{C_0}Y) && \Big[1_{C_0} = \sum_{x\in C_0'} 1_{X=x}\Big]\\
&= E(1_{C_0}Y) \;+\; E(1_{C\setminus C_0}Y) && \text{[(6.21)]}\\
&= E(1_C Y). && \text{[(6.20)]}
\end{aligned}
$$

▷ **Solution 10-3** (a) If $\mathscr{E}$ is a finite or countable partition of Ω and $\sigma(\mathscr{E}) = \mathscr{C}$, then, for all $A_i \in \mathscr{E}$, the indicator 1_{A_i} is $\mathscr{C}$-measurable. This implies that $\sum_{A_i\in\mathscr{E}} E(Y|A_i)\cdot 1_{A_i}$ is $\mathscr{C}$-measurable as well (see Lemma 2.19). Hence, condition (a) of Definition 10.2 is satisfied.

(b) According to Lemma 1.20 and Equations (1.35), (1.36), for all $C \in \mathscr{C}$,

$$1_C \;=\; \sum_{A_i\in\mathscr{E},\, A_i\subset C} 1_{A_i}. \tag{10.64}$$

Hence,

$$
\begin{aligned}
E\left(1_C \cdot \sum_{A_i\in\mathscr{E}} E(Y|A_i)\cdot 1_{A_i}\right) &= E\left(\sum_{A_i\in\mathscr{E}} E(Y|A_i)\cdot 1_{A_i}1_C\right)\\
&= \sum_{A_i\in\mathscr{E}} E\left[E(Y|A_i)\cdot 1_{A_i}1_C\right] && \text{[Box 6.1 (vii)]}\\
&= \sum_{A_i\in\mathscr{E},\, A_i\subset C} E\left[E(Y|A_i)\cdot 1_{A_i}\right] && \text{[(10.64)]}\\
&= \sum_{A_i\in\mathscr{E},\, A_i\subset C} E(Y|A_i)\cdot E(1_{A_i}) && \text{[Box 6.1 (iii)]}\\
&= \sum_{A_i\in\mathscr{E},\, A_i\subset C} \frac{1}{P(A_i)}\cdot E(1_{A_i}\cdot Y)\cdot P(A_i) && \text{[(9.7), (6.4)]}\\
&= \sum_{A_i\in\mathscr{E},\, A_i\subset C} E(1_{A_i}\cdot Y)\\
&= E\left(\sum_{A_i\in\mathscr{E},\, A_i\subset C} 1_{A_i}\cdot Y\right) && \text{[Box 6.1 (vii)]}\\
&= E(1_C Y). && \text{[(10.64)]}
\end{aligned}
$$

This shows that condition (b) of Definition 10.2 is satisfied and that $\sum_{A_i\in\mathscr{E}} E(Y|A_i)\cdot 1_{A_i}$ is a version of $E(Y|\mathscr{C})$. Equation (10.14) then follows from Proposition (10.9) and the assumption that $\mathscr{E}$ is a countable partition of Ω and $P(A_i)>0$ for all $A_i\in\mathscr{E}$ (see Theorem 10.17).

▷ **Solution 10-4** Inspecting Table 9.2 (p. 282) shows that the random variable $1_{X=0}\cdot E(Y|X)$ has two values: .60 and 0. Hence,

$$\begin{aligned} E[1_{X=0}\cdot E(Y|X)] &= .60\cdot P(1_{X=0}=1, X=0) \;+\; 0\cdot P(1_{X=0}=1, X=1) \\ &= .60\cdot P(X=0) \;=\; .60\cdot(.152+.348+.096+.024) \;=\; .372. \end{aligned}$$

▷ **Solution 10-5** (i) The σ-algebra generated by α is $\{\Omega,\emptyset\}$, which is a subset of every σ-algebra on Ω. Hence, $\sigma(\alpha)\subset\mathscr{C}$, which shows that Rule (i) is a special case of Rule (vii).

(ii) Both sides are $\mathscr{C}$-measurable. Furthermore, for all $C\in\mathscr{C}$,

$$\begin{aligned} E[1_C(\alpha+Y)] &= E(1_C\alpha+1_C Y) \;=\; E(1_C\alpha)+E(1_C Y) && \text{[Box 6.1 (vi)]} \\ &= E(1_C\alpha)+E[1_C E(Y|\mathscr{C})] && \text{[Def. 10.2 (b)]} \\ &= E\big(1_C\alpha+1_C E(Y|\mathscr{C})\big] && \text{[Box 6.1 (vi)]} \\ &= E\big(1_C[\alpha+E(Y|\mathscr{C})]\big). \end{aligned}$$

Hence, according to conditions (a) and (b) of Definition 10.2 , $\alpha+E(Y|\mathscr{C})\in\mathscr{E}(\alpha+Y|\mathscr{C})$.

(iii) Both sides are $\mathscr{C}$-measurable. For all $C\in\mathscr{C}$,

$$\begin{aligned} E(1_C\cdot\alpha Y) &= \alpha E(1_C\cdot Y) && \text{[Box 6.1 (iii)]} \\ &= \alpha E[1_C\cdot E(Y|\mathscr{C})]. && \text{[Def. 10.2 (b)]} \end{aligned}$$

Hence, according to Definition 10.2, $\alpha E(Y|\mathscr{C})\in\mathscr{E}(\alpha Y|\mathscr{C})$.

(iv) This rule immediately follows from condition (b) of Definition 10.2 for $C=\Omega$, because

$$\begin{aligned} E[E(Y|\mathscr{C})] &= E[1_\Omega E(Y|\mathscr{C})] && \text{[(3.31)]} \\ &= E(1_\Omega Y) && \text{[Def. 10.2 (b)]} \\ &= E(Y). && \text{[(3.31)]} \end{aligned}$$

(v) The terms on both sides of this equation are $\mathscr{C}_0$-measurable, because of Definition 10.2 (a). Furthermore, for all $C\in\mathscr{C}_0\subset\mathscr{C}$,

$$\begin{aligned} E\big(1_C E[E(Y|\mathscr{C})|\mathscr{C}_0]\big) &= E\big(1_C E(Y|\mathscr{C})\big) && \text{[Def. 10.2 (b)]} \\ &= E(1_C Y). && \text{[Def. 10.2 (b)]} \end{aligned}$$

In the first equation, we apply Definition 10.2 (b) to $E[E(Y|\mathscr{C})|\mathscr{C}_0]$ and $\mathscr{C}_0$, whereas in the second, we apply it to $E(Y|\mathscr{C})$ and $\mathscr{C}$. The last equation shows that $E[E(Y|\mathscr{C})|\mathscr{C}_0]\in\mathscr{E}(Y|\mathscr{C}_0)$.

(vi) The constant $E(Y)$ is measurable with respect to any σ-algebra $\mathscr{C}$ on Ω. Furthermore, if Y and $\mathscr{C}$ are independent, then Y and 1_C are independent for all $C\in\mathscr{C}$ (see Rem. 5.42). Hence, for $C\in\mathscr{C}$, $E(1_C Y)=E(1_C)E(Y)$ for all $C\in\mathscr{C}$ (see Th. 6.24). Therefore, Rule (ii) of Box 6.1 yields

$$E(1_C Y) \;=\; E(1_C)E(Y) \;=\; E[1_C E(Y)], \quad \forall C\in\mathscr{C}.$$

(vii) We assume that Y is $\mathscr{C}$-measurable. Furthermore,

$$E(1_C Y) \;=\; E(1_C Y), \quad \forall C\in\mathscr{C}.$$

obviously holds. Hence, according to Definition 10.2 (b) and Remark 10.10, this implies that $Y\in\mathscr{E}(Y|\mathscr{C})$ and $E(Y|\mathscr{C}) \underset{P}{=} Y$.

(viii) By definition, $E(Y|\mathscr{C})$ is $\mathscr{C}$-measurable. Hence,

$$\begin{aligned} E\big[Y\big|E(Y|\mathscr{C})\big] &= E\big[E(Y|\mathscr{C})\big|E(Y|\mathscr{C})\big] && \big[\text{(v) with } \mathscr{C}_0 = \sigma[E(Y|\mathscr{C})]\big] \\ &= E(Y|\mathscr{C})\,. && [\text{(vii)}] \end{aligned}$$

(ix)

$$\begin{aligned} Y_1 \underset{P}{=} Y_2 &\Rightarrow \forall\, C \in \mathscr{C}\colon\ E(1_C \cdot Y_1) = E(1_C \cdot Y_2) && [\text{Th. 3.48, } \mathscr{C} \subset \mathscr{A}] \\ &\Rightarrow \forall\, C \in \mathscr{C}\colon\ E[1_C \cdot E(Y_1|\mathscr{C})] = E[1_C \cdot E(Y_2|\mathscr{C})] && [\text{Def. 10.2 (b)}] \\ &\Rightarrow E(Y_1|\mathscr{C}) = E(Y_2|\mathscr{C})\,. && [\text{Th. 3.48, Def. 10.2 (a)}] \end{aligned}$$

(x)

$$\begin{aligned} E(Y) < \infty &\Rightarrow \forall\, V \in \mathscr{E}(Y|\mathscr{C})\colon\ E(V) < \infty && [E(V) = E(Y), \text{(iv)}] \\ &\Rightarrow \forall\, V \in \mathscr{E}(Y|\mathscr{C})\colon\ V \text{ is real-valued } P\text{-a.s.} && [\text{Lemma 3.41}] \end{aligned}$$

Now let $V^* \in \mathscr{E}(Y|\mathscr{C})$ and $A := \{\omega \in \Omega\colon V(\omega) \notin \mathbb{R}\}$. Then $A \in \mathscr{C}$ and $P(A) = 0$. Define $V := 1_{\Omega\setminus A} \cdot V^*$. Then V is real-valued, $\mathscr{C}$-measurable, and $V \underset{P}{=} V^*$, which implies $V \in \mathscr{E}(Y|\mathscr{C})$.

(xv) If Y_1 (or Y_2) is real-valued and with finite expectation, then there is a real-valued version $E(Y_1|\mathscr{C}) \in \mathscr{E}(Y_1|\mathscr{C})$ [or $E(Y_2|\mathscr{C}) \in \mathscr{E}(Y_2|\mathscr{C})$] [see Th. 10.9 (i) and Rem. 6.4]. If Y_1 (or Y_2) is nonnegative, then there is a nonnegative version $E(Y_1|\mathscr{C}) \in \mathscr{E}(Y_1|\mathscr{C})$ [or $E(Y_2|\mathscr{C}) \in \mathscr{E}(Y_2|\mathscr{C})$] [see Box 10.3 (vi) for $\alpha = 0$]. [Note that the proof of Box 10.3 (vi) uses Box 10.2 (ii).]

For versions $E(Y_1|\mathscr{C}) \in \mathscr{E}(Y_1|\mathscr{C})$, $E(Y_2|\mathscr{C}) \in \mathscr{E}(Y_2|\mathscr{C})$ and all $C \in \mathscr{C}$,

$$\begin{aligned} & E\big(1_C\,[E(Y_1|\mathscr{C}) + E(Y_2|\mathscr{C})]\big) \\ &= E\big(1_C\ E(Y_1|\mathscr{C})\big) + E\big(1_C\ E(Y_2|\mathscr{C})\big) && [\text{Box 6.1 (vi)}] \\ &= E(1_C\ Y_1)\ +\ E(1_C\ Y_2) && [\text{Def. 10.2 (i) (b)}] \\ &= E[1_C\ (Y_1 + Y_2)]. && [\text{Box 6.1 (vi)}] \end{aligned}$$

(xvi) The conditional expectations $E(Y_1|\mathscr{C})$ and $E(Y_2|\mathscr{C})$ are $\mathscr{C}$-measurable by definition. According to Example 2.61 this implies that $\alpha_1\,E(Y_1|\mathscr{C}) + \alpha_2\,E(Y_2|\mathscr{C})$ is $\mathscr{C}$-measurable as well. Furthermore, for $C \in \mathscr{C}$,

$$\begin{aligned} & E\big(1_C\,[\alpha_1\,E(Y_1|\mathscr{C}) + \alpha_2\,E(Y_2|\mathscr{C})]\big) \\ &= \alpha_1\,E\big(1_C\ E(Y_1|\mathscr{C})\big) + \alpha_2 E\big(1_C\ E(Y_2|\mathscr{C})\big) && [\text{Box 6.1 (vii)}] \\ &= \alpha_1\,E(1_C\ Y_1)\ +\ \alpha_2 E(1_C\ Y_2) && [\text{Def. 10.2 (i) (b)}] \\ &= E[1_C\ (\alpha_1\ Y_1 + \alpha_2\ Y_2)]. && [\text{Box 6.1 (vii)}] \end{aligned}$$

The equation for n summands follows by induction.

(xi) Cor. 8.16 of Klenke, 2008, p. 176).

(xiv) If Y_1 is $\mathscr{C}$-measurable, then $Y_1 \cdot E(Y_2|\mathscr{C})$ is $\mathscr{C}$-measurable as well [see Def. 10.2 (a), Th. 2.57]. First, consider the case $Y_1 = 1_{C^*}$, for $C^* \in \mathscr{C}$. Then, for all $C \in \mathscr{C}$,

$$\begin{aligned} \int 1_C \cdot E(1_{C^*} \cdot Y_2|\mathscr{C})\,dP &= \int 1_C \cdot 1_{C^*} \cdot Y_2\ dP && [\text{Def. 10.2 (a)}] \\ &= \int 1_{C\cap C^*} \cdot Y_2\ dP && [(1.32)] \end{aligned}$$

$$= \int 1_{C\cap C^*}\cdot E(Y_2\,|\mathscr{C})\,dP \qquad [C\cap C^*\in\mathscr{C}, \text{Def. 10.2 (b)}]$$
$$= \int 1_C\cdot 1_{C^*}\cdot E(Y_2\,|\mathscr{C})\,dP. \qquad [(1.32)]$$

If $E(Y_1^2), E(Y_2^2)<\infty$ or Y_1, Y_2 nonnegative, then Remark 3.30, Box 10.1 (xi), (xv), (xvi) imply, for all $\mathscr{C}$-measurable Y_1,

$$\int 1_C\cdot E(Y_1\cdot Y_2\,|\mathscr{C})\,dP = \int 1_C\cdot Y_1\cdot E(Y_2\,|\mathscr{C})\,dP, \quad \forall\, C\in\mathscr{C}.$$

Now Theorem 3.48 yields

$$E(Y_1\cdot Y_2\,|\mathscr{C}) \underset{P}{=} Y_1\cdot E(Y_2\,|\mathscr{C}).$$

(xii)

$$\begin{aligned}
&Cov[Y, E(Y|\mathscr{C})]\\
&= E[Y\cdot E(Y|\mathscr{C})] - E(Y)\cdot E[E(Y|\mathscr{C})] && [\text{Box 7.1 (i)}]\\
&= E\big(E[Y\cdot E(Y|\mathscr{C})\,|\mathscr{C}]\big) - E(Y)\cdot E[E(Y|\mathscr{C})] && [\text{Box 10.1 (iv)}]\\
&= E[E(Y|\mathscr{C})\cdot E(Y|\mathscr{C})] - E[E(Y|\mathscr{C})]\cdot E[E(Y|\mathscr{C})] && [\text{Box 10.1 (xiv), (iv)}]\\
&= Var[E(Y|\mathscr{C})]. && [\text{Box 6.2 (i)}]
\end{aligned}$$

(xiii) Note that $E({Y_2}^2)<\infty$ implies $E[E(Y_2\,|\mathscr{C})^2]<\infty$ [see Box 10.1 (xi)]. Hence,

$$\begin{aligned}
Cov(Y_1, Y_2) &= E(Y_1\cdot Y_2) - E(Y_1)\cdot E(Y_2) && [\text{Box 7.1 (i)}]\\
&= E[Y_1\cdot Y_2\,|\mathscr{C}] - E(Y_1)\cdot E[E(Y_2\,|\mathscr{C})] && [\text{(iv)}]\\
&= E[Y_1\cdot E(Y_2\,|\mathscr{C})] - E(Y_1)\cdot E[E(Y_2\,|\mathscr{C})] && [\sigma(Y_1)\subset\mathscr{C}, \text{(xiv)}]\\
&= Cov[Y_1, E(Y_2\,|\mathscr{C})]. && [\text{Box 7.1 (i)}]
\end{aligned}$$

▷ **Solution 10-6** (vi) First, we prove

$$Y \underset{P}{\geq} 0 \;\Rightarrow\; \exists V\in\mathscr{E}(Y|\mathscr{C})\colon V\geq 0. \tag{10.65}$$

Let $V^*\in\mathscr{E}(Y|\mathscr{C})$. Then

$$\int V^*\,dP = \int Y\,dP \geq 0 \qquad [\text{Box 10.2 (iv), (3.50)}]$$

By contraposition, (3.51) implies $V^* \underset{P}{\geq} 0$, i. e.,

$$\exists A\in\mathscr{C}\colon\; P(A)=0 \;\wedge\; \forall\,\omega\in\Omega\setminus A\colon V^*(\omega)\geq 0.$$

If we define $V := 1_{\Omega\setminus A}\cdot V^*\geq 0$, then V is $\mathscr{C}$-measurable (see Th. 2.57), $V \underset{P}{=} V^*$, and $V\in\mathscr{E}(Y|\mathscr{C})$ [see (10.9)]. For $\alpha\in\mathbb{R}$, applying (10.65),

$$Y \underset{P}{\geq} \alpha \;\Rightarrow\; Y-\alpha \underset{P}{\geq} 0 \;\Rightarrow\; \exists V_\alpha\in\mathscr{E}(Y-\alpha|\mathscr{C})\colon V_\alpha \underset{P}{\geq} 0.$$

Now Rule (ii) of Box 10.1 implies that there is a $V := V_\alpha+\alpha$ such that $V\in\mathscr{E}(Y|\mathscr{C})$ and $V\geq\alpha$.

(viii)

$$\begin{aligned} Y \underset{P}{\leq} \alpha \;&\Rightarrow\; -Y \underset{P}{\geq} -\alpha \\ &\Rightarrow\; \exists\, V^* \in \mathscr{E}(-Y\,|\mathscr{C})\colon V^* \geq -\alpha \\ &\Rightarrow\; \exists\, V \in \mathscr{E}(Y\,|\mathscr{C})\colon V \leq \alpha. \qquad \text{[Box 10.1 (iii), } V := -V\text{]} \end{aligned}$$

(vii)

$$\begin{aligned} & Y \underset{P}{=} \alpha \\ \Rightarrow\; & Y \underset{P}{\geq} \alpha \wedge Y \underset{P}{\leq} \alpha \\ \Rightarrow\; & \exists V_1 \in \mathscr{E}(Y|\mathscr{C})\colon\ V_1 \geq \alpha \wedge \exists V_2 \in \mathscr{E}(Y|\mathscr{C})\colon\ V_2 \leq \alpha && \text{[(vi), (viii)]} \\ \Rightarrow\; & \exists V_1, V_2 \in \mathscr{E}(Y|\mathscr{C})\colon\ \exists A \in \mathscr{C}\colon P(A) = 0 \wedge \forall \omega \in \Omega \setminus A\colon V_1(\omega) = V_2(\omega) = \alpha && \text{[(10.9)]} \\ \Rightarrow\; & \exists V \in \mathscr{E}(Y|\mathscr{C})\colon V = \alpha. && \text{[}V := \alpha \cdot 1_A + V_1 \cdot 1_{\Omega\setminus A}\text{]} \end{aligned}$$

(iii), (v), and (iv) are direct implications of (vi), (vii), and (viii).
(x), (xi) These rules follow from (v) and (5.10).
(ii) is a straightforward implication of Theorem 3.43 and (vii).
(i) If $Y \underset{P}{\leq} Z$, define $W\colon \Omega \to \mathbb{R}$ by

$$W(\omega) = \begin{cases} Z(\omega) - Y(\omega), & \text{if } Y(\omega), Z(\omega) \in \mathbb{R} \\ 0, & \text{if } Y(\omega) = \infty \text{ or } Z(\omega) = -\infty. \end{cases}$$

Then $W \underset{P}{\geq} 0$ and $Y + W = Z$. Hence, choosing a nonnegative version $V^* \in \mathscr{E}(W\,|\mathscr{C})$ and a nonnegative or real-valued version $V \in \mathscr{E}(Y\,|\mathscr{C})$ [see Box 10.1 (x)],

$$\begin{aligned} E(Z|\mathscr{C}) &\underset{P}{=} E(Y + W\,|\mathscr{C}) \\ &\underset{P}{=} V + V^* && \text{[Box 10.1 (xv)]} \\ &\underset{P}{\geq} E(Y|\mathscr{C})\,. && \text{[(i)]} \end{aligned}$$

(iii) The proof is by contraposition. Assume that there is a $E(Y|\mathscr{C})$ such that $P(A) > 0$ for $A := \{\omega \in \Omega\colon E(Y|\mathscr{C})(\omega) < \alpha\}$. Now,

$$A = \bigcup_{i=1}^{\infty} A_i \;\text{ for }\; A_i := \{\omega \in \Omega\colon E(Y|\mathscr{C})(\omega) < \alpha - 1/i\}, \quad i \in \mathbb{N},$$

and $A_1 \subset A_2 \subset \ldots$ implies $P(A) = \lim_{i\to\infty} P(A_i)$. Therefore, there is an $i \in \mathbb{N}$ such that $P(A_i) > 0$. Because $E(Y|\mathscr{C})$ is $\mathscr{C}$-measurable, we can conclude $A, A_i \in \mathscr{C}$ [see Def. 2.5 and Eq. (1.19)]. Now condition (b) of Definition 10.2 and (3.50) imply

$$\int Y \cdot 1_{A_i}\, dP = \int E(Y|\mathscr{C}) \cdot 1_{A_i}\, dP \leq \left(\alpha - \frac{1}{i}\right) P(A_i) < \alpha \cdot P(A_i),$$

which is a contradiction to $Y \underset{P}{\geq} \alpha$ [see (3.50)].

▷ **Solution 10-7** Let $g(X) \in \mathscr{E}(Y\,|X)$. Then, according to Equation (10.26), for all $g^*(X) \in \mathscr{E}(Y\,|X)$,

$$g(x) = g^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X.$$

Hence, if $P(X{=}x) > 0$ for an $x \in \Omega'_X$, then according to Remark 2.71, $g(x) = g^*(x)$, i. e., $g(x)$ is uniquely defined. Furthermore, Equation (9.6) yields

$$\begin{aligned}
E(Y|X{=}x) &= \int Y\,dP^{X=x} && [(9.5)]\\
&= \frac{1}{P(X{=}x)}\cdot\int 1_{X=x}\cdot Y\,dP && [(9.7)]\\
&= \frac{1}{P(X{=}x)}\cdot\int 1_{X=x}\cdot g(X)\,dP && [\text{Def. 10.2 (b)}]\\
&= \frac{1}{P(X{=}x)}\cdot\int 1_{X=x}\cdot g(x)\,dP && [1_{X=x}\cdot g(X) = 1_{X=x}\cdot g(x)]\\
&= \frac{1}{P(X{=}x)}\cdot g(x)\cdot\int 1_{X=x}\,dP && [(3.32)]\\
&= \frac{1}{P(X{=}x)}\cdot g(x)\cdot P(X{=}x) \;=\; g(x). && [(3.8)]
\end{aligned}$$

▷ **Solution 10-8** If g is a factorization of $E(Y|X)\in\mathscr{E}(Y|X)$, then, for all $\omega\in\{X{=}x\}$,

$$\begin{aligned}
E(Y|X)(\omega) &= (g\circ X)(\omega) && [(10.22)]\\
&= g[X(\omega)] \;=\; g(x) && [\omega\in\{X{=}x\}]\\
&= E(Y|X{=}x)\,. && [(10.27)]
\end{aligned}$$

▷ **Solution 10-9** Assume that Z is a discrete random variable with values $z_1,z_2,\ldots\in\Omega'_Z$ such that $P_Z(\{z_1,z_2,\ldots\})=1$ and, for all $i=1,2\ldots$, $\{z_i\}\in\mathscr{A}'_Z$. Then, for all $x\in\Omega'_X$ with $P(X{=}x,Z{=}z_i)>0$ for all $i\in\mathbb{N}$,

$$\begin{aligned}
E(Y|X{=}x) &= E[E(Y|X,Z)\,|X{=}x] && [(10.38)]\\
&= E[g(X,Z)\,|\,X{=}x] && [(10.22)]\\
&= \sum_{i=1}^{\infty} g(x,z_i)\cdot P(X{=}x,Z{=}z_i\,|X{=}x) && [\text{Rem. 10.35, (9.15), (9.21)}]\\
&= \sum_{i=1}^{\infty} E(Y|X{=}x,Z{=}z_i)\cdot\frac{P(X{=}x,Z{=}z_i,X{=}x)}{P(X{=}x)} && [(10.27),\,(4.2)]\\
&= \sum_{i=1}^{\infty} E(Y|X{=}x,Z{=}z_i)\cdot\frac{P(X{=}x,Z{=}z_i)}{P(X{=}x)}\\
&= \sum_{i=1}^{\infty} E(Y|X{=}x,Z{=}z_i)\cdot P(Z{=}z_i\,|X{=}x). && [(4.2)]
\end{aligned}$$

▷ **Solution 10-10** If $E(Y|X)\underset{P}{=}E(Y)$, then

$$\begin{aligned}
Cov(X,Y) &= Cov[X,E(Y|X)] && [\text{Box 10.2 (xiii)}]\\
&= Cov[X,E(Y)] && [E(Y|X)\underset{P}{=}E(Y),\ \text{Box 7.1 (x)}]\\
&= 0. && [\text{Box 7.1 (vii)}]
\end{aligned}$$

Chapter 11
Residual, Conditional Variance, and Conditional Covariance

In chapters 9 and 10 we introduced the concepts conditional expectation and regression. In this chapter we turn to the *residual* of a conditional expectation. In a sense, a residual is 'the other side' of a conditional expectation and its properties supplements the properties of conditional expectations. Oftentimes a residual is what econometricians call a 'disturbance', applied statisticians call an 'error term', and psychometricians call a 'measurement error'. Furthermore, we define the *coefficient of determination*, which represents the proportion of variance of a regressand explained by the regressor. It appears under different names in special areas of applied statistics, ranging from 'intra-class correlation' to 'reliability' in psychometrics. The square root of the coefficient of determination is known as the 'multiple correlation'. Next, we will define the concepts of a *conditional variance* and a *conditional covariance* given a σ-algebra, as well as the *partial correlation*. Just like the expectation has been used to define variance, covariance, and correlation, the conditional expectation can be used to define conditional variance, conditional covariance, and the partial correlation.

11.1 Residual With Respect to a Conditional Expectation

In section 10.4.2 we have shown that a conditional expectation $E(Y|\mathscr{C})$ is the best approximation of Y in the sense of minimizing the mean-squared error function. Now we will study the properties of the deviation of Y from $E(Y|\mathscr{C})$. Defining this deviation, we refer to the following assumptions.

Assumptions 11.1
$Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ is a real-valued random variable with finite expectation and $\mathscr{C}\subset\mathscr{A}$ is a σ-algebra.

According to Rule (x) of Box 10.1, $E(Y)<\infty$ implies that there is a real-valued version $E(Y|\mathscr{C})$. In this chapter it would also suffice to assume that there is a real-valued version $E(Y|\mathscr{C})$. Referring to a real-valued version $E(Y|\mathscr{C})$ avoids the subtraction of ∞ and ∞ for values of Y and $E(Y|\mathscr{C})$, respectively.

Definition 11.2 (Residual With Respect to a Conditional Expectation)
Let the assumptions 11.1 hold and let $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$. *Then*

$$\varepsilon := Y - E(Y|\mathscr{C}) \tag{11.1}$$

is called a version of the $residual\ of\ Y\ with\ respect\ to\ E(Y|\mathscr{C})$.

Remark 11.3 (Versions of the Residual) If $E(Y|\mathscr{C})$, $E(Y|\mathscr{C})^* \in \mathscr{E}(Y|\mathscr{C})$ and ε, ε^* are the respective residuals, then $\varepsilon \underset{P}{=} \varepsilon^*$. According to Rule (x) of Box 10.1, finiteness of $E(Y)$ implies that there is a real-valued version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$. ◁

Box 11.1 summarizes some properties of the residual, which are proved in Exercise 11-1. All these properties follow from the definition of a residual, provided that we consider only *real-valued versions of* $E(Y|\mathscr{C})$ (see Rem. 11.3).

Remark 11.4 (Some Special Cases) Because $E(Y|\mathscr{C})$ is $\mathscr{C}$-measurable, the following equations are special cases of Rules (vii) and (viii) of Box 11.1, respectively.

$$E\big[\varepsilon \,\big|\, E(Y|\mathscr{C})\big] \underset{P}{=} 0, \tag{11.2}$$

$$Cov[\varepsilon, E(Y|\mathscr{C})] = 0, \quad \text{if } E(Y^2) < \infty. \tag{11.3}$$

According to Equation (11.2), the conditional expectation of the residual ε given $E(Y|\mathscr{C})$ is 0 with probability 1. According to the second equation, the residual $\varepsilon = Y - E(Y|\mathscr{C})$ is uncorrelated with $E(Y|\mathscr{C})$ if $E(Y^2) < \infty$. [Note that finiteness of $E\big(E(Y|\mathscr{C})^2\big)$ follows from $E(Y^2) < \infty$; see Box 10.1 (xi)].

Now consider a random variable $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ with $\sigma(X) = \mathscr{C}$. Then two other special cases of Rule (vii) are

$$E\big[\varepsilon \,\big|\, E(Y|X)\big] \underset{P}{=} E(\varepsilon|X) \underset{P}{=} 0. \tag{11.4}$$

If $f(X)$ denotes the composition of X and a function $f\colon (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ that is $(\mathscr{A}'_X, \overline{\mathscr{B}})$-measurable, and if $E(Y^2)$, $E[f(X)]^2 < \infty$, then

$$Cov[\varepsilon, f(X)] = 0 \tag{11.5}$$

is a special case of Rule (viii). Hence, if $(\Omega'_X, \mathscr{A}'_X) = (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ and $E(X^2)$, $E(Y^2) < \infty$, then

$$Cov(\varepsilon, X) = 0, \tag{11.6}$$

is another special case of Rule (viii).

Now consider the residual $\varepsilon := Y - E(Y|X,Z)$, where $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ and $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ are not necessarily real-valued. In this case,

$$E(\varepsilon|X,Z) \underset{P}{=} E(\varepsilon|X) \underset{P}{=} E(\varepsilon|Z) \underset{P}{=} 0 \tag{11.7}$$

Box 11.1 Rules of Computation for a Residual

Let the assumptions 11.1 hold. Then the following properties hold for all real-valued versions of $E(Y|\mathscr{C})$ and all versions of the residual ε defined in (11.1):

$$\varepsilon \underset{P}{=} Y - E(Y|\mathscr{C}). \tag{i}$$

$$Y \underset{P}{=} E(Y|\mathscr{C}) + \varepsilon. \tag{ii}$$

$$E(\varepsilon) = 0. \tag{iii}$$

$$Var(Y) = Var[E(Y|\mathscr{C})] + Var(\varepsilon), \quad \text{if } E(Y^2) < \infty. \tag{iv}$$

$$\varepsilon \underset{P}{=} 0, \quad \text{if } Y \underset{P}{=} E(Y|\mathscr{C}). \tag{v}$$

Additionally, let $\mathscr{C}_0$ be a σ-algebra and $W\colon (\Omega, \mathscr{A}, P) \to (\Omega'_W, \mathscr{A}'_W)$ be a random variable. Then

$$E(\varepsilon|\mathscr{C}_0) \underset{P}{=} 0, \quad \text{if } \mathscr{C}_0 \subset \mathscr{C}. \tag{vi}$$

$$E(\varepsilon|W) \underset{P}{=} 0, \quad \text{if } \sigma(W) \subset \mathscr{C}. \tag{vii}$$

If W is real-valued, $\sigma(W) \subset \mathscr{C}$, and $E(W^2), E(Y^2) < \infty$, then

$$Cov(\varepsilon, W) = E(\varepsilon \cdot W) = 0. \tag{viii}$$

$$Cov[W, E(Y|\mathscr{C})] = Cov[W, E(Y|\mathscr{C}) + \varepsilon] = Cov(W, Y). \tag{ix}$$

are special cases of Rule (vii), where $\mathscr{C} = \sigma(X, Z)$. If we additionally assume X and Z to be numerical and $E(Y^2), E(X^2), E(Z^2) < \infty$, then

$$Cov(X, \varepsilon) = E(X \cdot \varepsilon) = Cov(Z, \varepsilon) = E(Z \cdot \varepsilon) = 0 \tag{11.8}$$

are special cases of Rule (viii) (see Exercise 11-2). ◁

Example 11.5 (No Treatment for Joe – continued) Table 11.1 displays the conditional expectations $E(Y|X)$, $E(Y|X,U)$, and $P(X=1|U)$ and their residuals. First, we illustrate the property $E(\varepsilon) = 0$ for $\varepsilon = Y - E(Y|X)$. Looking at the table reveals that $\varepsilon = Y - E(Y|X)$ has four different values: $-.60$, $.40$, $.60$, and $-.40$. Hence, according to Equation (6.3),

$$\begin{aligned} E(\varepsilon) &= -.60 \cdot (.152 + .096) + .40 \cdot (.348 + .024) + .60 \cdot (0 + .152) - .40 \cdot (0 + .228) \\ &= 0. \end{aligned}$$

Second, we illustrate the property

$$E(\varepsilon|X) = 0,$$

Table 11.1. No Treatment for Joe With Conditional Expectations and Residuals

Elements of Ω		Observables			Conditional expectations			Residuals		
Unit, Treatment, Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\mid X,U)$	$E(Y\mid X)$	$P(X=1\mid U)$	$Y-E(Y\mid X,U)$	$Y-E(Y\mid X)$	$X-P(X=1\mid U)$
(*Joe*, *no*, −)	.152	*Joe*	0	0	.696	.60	0	−.696	−.60	0
(*Joe*, *no*, +)	.348	*Joe*	0	1	.696	.60	0	.304	.40	0
(*Joe*, *yes*, −)	0	*Joe*	1	0	0	.40	0	0	−.40	1
(*Joe*, *yes*, +)	0	*Joe*	1	1	0	.40	0	1	.60	1
(*Ann*, *no*, −)	.096	*Ann*	0	0	.20	.60	.76	−.20	−.60	−.76
(*Ann*, *no*, +)	.024	*Ann*	0	1	.20	.60	.76	.80	.40	−.76
(*Ann*, *yes*, −)	.228	*Ann*	1	0	.40	.40	.76	−.40	−.40	.24
(*Ann*, *yes*, +)	.152	*Ann*	1	1	.40	.40	.76	.60	.60	.24

which follows from Equation (11.7) and Remark 10.16. Because X is an indicator variable, according to Equation (9.23) and Remark 10.35, it suffices to show that $E(\varepsilon\mid X=0)=0$ and $E(\varepsilon\mid X=1)=0$. The four values of $\varepsilon = Y-E(Y\mid X)$ occur with $(X=0)$-conditional probabilities

$$\frac{.152+.096}{.152+.348+.096+.024} = .40, \qquad \frac{.348+.024}{.152+.348+.096+.024} = .60, \qquad 0, \qquad 0,$$

and with $(X=1)$-conditional probabilities

$$0, \qquad 0, \qquad \frac{.152}{0+0+.152+.228} = .40, \qquad \frac{.228}{0+0+.152+.228} = .60,$$

respectively. Hence, according to Equation (9.20),

$$E(\varepsilon\mid X=0) = -.60\cdot.40 + .40\cdot.60 + .60\cdot 0 - .40\cdot 0 = 0$$

and

$$E(\varepsilon\mid X=1) = -.60\cdot 0 + .40\cdot 0 + .60\cdot.40 - .40\cdot.60 = 0.$$

Because X is dichotomous and $P(X=0), P(X=1)>0$, we can conclude:

$$E(\varepsilon\mid X) = E(\varepsilon\mid X=0)\cdot 1_{X=0} + E(\varepsilon\mid X=1)\cdot 1_{X=1} = 0\cdot 1_{X=0} + 0\cdot 1_{X=1} = 0$$

[see Eq. (9.23) and Rem. 10.35]. ◁

11.2 Coefficient of Determination and Multiple Correlation

The *coefficient of determination* quantifies the strength of the dependence of a numerical random variable Y on a σ-algebra $\mathscr{C}$, where we refer to the dependence described by the conditional expectation $E(Y|\mathscr{C})$ (see Rem. 10.18). As we will see, the *multiple correlation* is a closely related concept. Reading the following definition, remember that $E(Y^2) < \infty$ implies $Var(Y) < \infty$ and $E(Y) < \infty$ (see Rem. 6.26). It also implies that the conditional expectation $E(Y|\mathscr{C})$ is defined.

Definition 11.6 (Coefficient of Determination)
Let $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a real-valued random variable with $E(Y^2) < \infty$ and $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra. Then

$$R^2_{Y|\mathscr{C}} := \begin{cases} \dfrac{Var[E(Y|\mathscr{C})]}{Var(Y)}, & \text{if } Var(Y) > 0, \\ 0, & \text{if } Var(Y) = 0, \end{cases} \tag{11.9}$$

is called the coefficient of determination of $E(Y|\mathscr{C})$.

Remark 11.7 (The Case $Var(Y) = 0$) If $Var(Y) = 0$, then $Var[E(Y|\mathscr{C})] = 0$ [see Box 6.2 (iv) and Box 10.3 (v)]. Defining $R^2_{Y|\mathscr{C}} = 0$ if $Var(Y) = 0$ is arbitrary. However, this choice implies

$$\mathscr{C} \perp\!\!\!\perp_P Y \quad \Rightarrow \quad E(Y|\mathscr{C}) \underset{P}{=} E(Y) \quad \Rightarrow \quad R^2_{Y|\mathscr{C}} = 0, \tag{11.10}$$

even if $Var(Y) = 0$. In other words, with this definition, *independence of $\mathscr{C}$ and Y* implies *mean independence* of Y from $\mathscr{C}$, which itself implies $R^2_{Y|\mathscr{C}} = 0$. ◁

Remark 11.8 (Range of the Coefficient of Determination) Using Rule (iv) of Box 11.1 yields

$$R^2_{Y|\mathscr{C}} = \frac{Var[E(Y|\mathscr{C})]}{Var[E(Y|\mathscr{C})] + Var(\varepsilon)}, \tag{11.11}$$

provided that $Var(Y) > 0$. Because $Var(\varepsilon)$ is nonnegative, $0 \le R^2_{Y|\mathscr{C}} \le 1$. The number $R^2_{Y|\mathscr{C}}$ is close to 1 if the variance of the residual $\varepsilon = Y - E(Y|\mathscr{C})$ is small compared to the variance of the conditional expectation $E(Y|\mathscr{C})$. In contrast, $R^2_{Y|\mathscr{C}}$ is close to 0 if the variance of the residual is large compared to the variance of $E(Y|\mathscr{C})$. ◁

Remark 11.9 (Conditions Implying $R^2_{Y|\mathscr{C}} = 1$) If $Var(Y) > 0$ and we assume that $Y \underset{P}{=} E(Y|\mathscr{C})$, then $R^2_{Y|\mathscr{C}} = 1$ [see Eq. (11.9) and Box 6.2 (v)]. Note that this does not necessarily mean that Y is $\mathscr{C}$-measurable. However, if Y is $\mathscr{C}$-measurable, then $Y \underset{P}{=} E(Y|\mathscr{C})$ already follows from Box 10.1 (vii). ◁

Remark 11.10 (Alternative Notation) Suppose that $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ is a random variable and $\mathscr{C} = \sigma(X)$, then we also use the notation $R^2_{Y|X}$ instead of $R^2_{Y|\mathscr{C}}$, i. e.,

$$R^2_{Y|X} := R^2_{Y|\sigma(X)}. \tag{11.12}$$

Equations (10.1) and (11.9) yield

$$R^2_{Y|X} := \begin{cases} \dfrac{Var[E(Y|X)]}{Var(Y)}, & \text{if } Var(Y) > 0, \\ 0, & \text{if } Var(Y) = 0. \end{cases} \tag{11.13}$$

If we consider the multivariate regressor $X = (X_1, \ldots, X_n)$, then we also use the notation

$$R^2_{Y|X_1,\ldots,X_n} := R^2_{Y|X}. \tag{11.14}$$

◁

Remark 11.11 (Correlation and the Coefficient of Determination) Assume that $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ is a real-valued random variable, $E(X^2), E(Y^2) < \infty$, and that there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ with

$$E(Y|X) = Q_{lin}(Y|X) = \beta_0 + \beta_1 X. \tag{11.15}$$

Then

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}, \tag{11.16}$$

and

$$R^2_{Y|X} = Corr(X,Y)^2, \tag{11.17}$$

which implies

$$R^2_{Y|X} = 0 \quad \Leftrightarrow \quad Corr(X,Y) = 0 \tag{11.18}$$

(see Exercise 11-3). Hence, under these assumptions, the correlation $Corr(X,Y)$ also quantifies the strength of the dependence of Y on X described by $E(Y|X)$. Both, $R^2_{Y|X}$ and $Corr(X,Y)$ are normed quantities. The first takes on its values in the interval $[0,1]$, the latter in the interval $[-1,1]$. In contrast, the slope β_1 as well as $Cov(X,Y)$ quantify the strength of the dependence described by (11.15) by real numbers without bounds. ◁

Remark 11.12 (Quantifying the Strength of Dependence) The term $R^2_{Y|X}$ quantifies the strength of the dependence of Y on X described by $E(Y|X)$, irrespective on whether or not Equation (11.15) holds. While $E(Y|X) = g(X)$ describes how the conditional expectation values $E(Y|X{=}x)$ of Y depend on the values x of X, the coefficient of determination $R^2_{Y|X}$ quantifies the strength of this dependence by a single real number between 0 and 1. Similarly, $R^2_{Y|\mathscr{C}}$ quantifies the strength of dependence of Y on $\mathscr{C}$ described by $E(Y|\mathscr{C})$ (see Rem. 10.18). ◁

Remark 11.13 (Uniqueness) If $V, V^* \in \mathscr{E}(Y|\mathscr{C})$, then V and V^* are P-equivalent and, according to Rule (v) of Box 6.2, this implies $Var(V) = Var(V^*)$. Hence, Equation (11.9) implies that $R^2_{Y|\mathscr{C}}$ is identical for all versions $V \in \mathscr{E}(Y|\mathscr{C})$. ◁

Remark 11.14 (Correlation of Y and the Conditional Expectation) If $E(Y^2)$ is finite, then the coefficient of determination $R^2_{Y|\mathscr{C}}$ is identical to the squared correlation of Y and $E(Y|\mathscr{C})$, i. e.,

$$R^2_{Y|\mathscr{C}} \;=\; Corr\,[Y,\, E(Y|\mathscr{C})]^2 \tag{11.19}$$

(see Exercise 11-4). Correspondingly,

$$R^2_{Y|X} \;=\; Corr\,[Y,\, E(Y|X)]^2. \tag{11.20}$$

Note that this equation does not rely on any parametrization of $E(Y|X)$. ◁

Definition 11.15 (Multiple Correlation)
Let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ and $Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be random variables and assume $E(Y^2)<\infty$. Then

$$R_{Y|X} \;:=\; \sqrt{R^2_{Y|X}} \tag{11.21}$$

is called the multiple correlation of Y and X. If $X=(X_1,\ldots,X_n)$ is a multivariate random variable, then we also use the notation

$$R_{Y|X_1,\ldots,X_n} \;:=\; R_{Y|X}\,. \tag{11.22}$$

Equations (11.20) and (11.21) immediately imply

$$R_{Y|X} \;=\; Corr\,[Y,\, E(Y|X)]. \tag{11.23}$$

Remark 11.16 (The Multiple Correlation is Not Symmetric) Note that, in contrast to a correlation of two numerical random variables, the multiple correlation of Y and X is not symmetric. In general, the roles of Y and X cannot be exchanged. ◁

Example 11.17 (No Treatment for Joe – continued) In Table 9.2 (p. 282) we displayed the conditional expectations $E(Y|X)$, $E(Y|X,U)$, and $P(X{=}1\,|\,U)$. Let us compute $R^2_{Y|X}$ for the conditional expectation $E(Y|X)$. Looking at the table reveals that $E(Y|X)$ has two different values: .60, which occurs with probability $.152+.348+.096+.024=.62$, and .40, which occurs with probability $0+0+.228+.152=.38$. Furthermore, the expectation of Y is

$$E(Y) \;=\; P(Y{=}1) \;=\; .348+0+.024+.152 \;=\; .524$$

Hence, according to Equation (i) of Box 6.2,

$$\begin{aligned} Var[E(Y|X)] \;&=\; E[E(Y|X)^2] \;-\; E[E(Y|X)]^2 \\ &=\; E[E(Y|X)^2] \;-\; E(Y)^2 \qquad\qquad \text{[Box 10.2 (iv)]} \\ &=\; (.60^2\cdot.62 \;+\; .40^2\cdot.38) \;-\; .524^2 \\ &\approx\; .284 \;-\; .2746 \;=\; .0094. \end{aligned}$$

The variance of Y is $Var(Y) = .524 \cdot (1 - .524) \approx 0.2494$. This yields

$$R^2_{Y|X} = \frac{Var[E(Y|X)]}{Var(Y)} \approx \frac{.0094}{.2494} \approx .0377.$$

Similarly, the conditional expectation $E(Y|U)$ takes on each of the two values .696 and .352 with probability .50. Hence,

$$\begin{aligned} Var[E(Y|U)] &= E[E(Y|U)^2] - E[E(Y|U)]^2 \\ &= E[E(Y|U)^2] - E(Y)^2 \qquad \text{[Box 10.2 (iv)]} \\ &= (.696^2 \cdot .50 + .352^2 \cdot .50) - .524^2 \\ &\approx .3042 - .2746 = .0296. \end{aligned}$$

Hence,

$$R^2_{Y|U} = \frac{Var[E(Y|U)]}{Var(Y)} \approx \frac{.0296}{.2494} \approx .1187.$$

Finally,

$$\begin{aligned} Var[E(Y|X,U)] &= E[E(Y|X,U)^2] - E[E(Y|X,U)]^2 \\ &= E[E(Y|X,U)^2] - E(Y)^2 \qquad \text{[Box 10.2 (iv)]} \\ &= (.696^2 \cdot .50 + .20^2 \cdot .12 + .40^2 \cdot .38) - .524^2 \\ &\approx .3078 - .2746 = .0332. \end{aligned}$$

Hence,

$$R^2_{Y|X,U} = \frac{Var[E(Y|X,U)]}{Var(Y)} \approx \frac{.0332}{.2494} \approx .1332.$$

Note that, in this example, $R^2_{Y|X,U}$ is smaller than the sum of $R^2_{Y|X}$ and $R^2_{Y|U}$. ◁

In the following theorem we present a condition under which the coefficients of determination are additive [see Eq. (11.30)]. This theorem also contains a condition under which the coefficient of X in the equation $E(Y|X) = \alpha_0 + \alpha_1 X$ is identical to the coefficient of X in the equation $E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z$ (see section 12.8 for a generalization).

Theorem 11.18 (Additivity of the Coefficients of Determination)
Let $X, Y, Z : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ be real-valued random variables with finite second moments and positive variances and assume that there are $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1 \in \mathbb{R}$, $E(Y|X,Z) \in \mathcal{E}(Y|X,Z)$, and $E(Z|X) \in \mathcal{E}(Z|X)$ such that

$$E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z, \tag{11.24}$$

$$E(Z|X) = \gamma_0 + \gamma_1 X. \tag{11.25}$$

(i) Then there are $\alpha_0, \alpha_1 \in \mathbb{R}$ such that

$$E(Y|X) = \alpha_0 + \alpha_1 X. \tag{11.26}$$

(ii) If $E(Z|X) \underset{P}{=} E(Z)$ *or* $\beta_2 = 0$*, then for* α_0, α_1 *of (i),*

$$\alpha_0 \;=\; \beta_0 \;+\; \beta_2\, E(Z) \tag{11.27}$$

and

$$\alpha_1 \;=\; \beta_1 \tag{11.28}$$

(iii) Finally, if

$$\Big(E(Z|X) \underset{P}{=} E(Z) \;\; or \;\; \beta_2 = 0\Big) \quad and \quad \Big(E(X|Z) \underset{P}{=} E(X) \;\; or \;\; \beta_1 = 0\Big),$$

then

$$Var[E(Y|X,Z)] \;=\; Var[E(Y|X)] + Var[E(Y|Z)] \tag{11.29}$$

and

$$R^2_{Y|X,Z} \;=\; R^2_{Y|X} \;+\; R^2_{Y|Z}. \tag{11.30}$$

(Proof p. 338)

Remark 11.19 (Independence of X and Z) Note that the condition specified in proposition (iii) of Theorem 11.18 is satisfied, e. g., if X and Z are independent [see Rule (vi) of Box 10.2]. It is also satisfied, e. g., if

$$E(Y|X,Z) \underset{P}{=} E(Y|X) \quad \text{and} \quad E(Y|X,Z) \underset{P}{=} E(Y|Z)\,. \tag{11.31}$$

◁

11.3 Conditional Variance and Covariance Given a σ-Algebra

The covariance $Cov(Y_1, Y_2)$ has been defined as the expectation of the product of the mean centered random variables $Y_1 - E(Y_1)$ and $Y_2 - E(Y_2)$, i. e.,

$$Cov(Y_1, Y_2) \;=\; E\big([Y_1 - E(Y_1)] \cdot [Y_2 - E(Y_2)]\big) \tag{11.32}$$

(cf. Def. 7.7). Similarly, we define the $\mathscr{C}$-conditional covariance $Cov(Y_1, Y_2|\mathscr{C})$ as the *$\mathscr{C}$-conditional expectation* of the product $[Y_1 - E(Y_1|\mathscr{C})] \cdot [Y_2 - E(Y_2|\mathscr{C})] = \varepsilon_1 \cdot \varepsilon_2$ of the residuals of Y_1 and Y_2 with respect to their $\mathscr{C}$-conditional expectations.

Definition 11.20 (Conditional Covariance Given a σ-Algebra)
For $i = 1,2$*, let* $Y_i \colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ *be real-valued random variables with* $E(Y_i^2) < \infty$*, let* $\mathscr{C} \subset \mathscr{A}$ *be a* σ*-algebra, and define* $\varepsilon_i := Y_i - E(Y_i|\mathscr{C})$*. Then*

$$Cov(Y_1, Y_2|\mathscr{C}) \;:=\; E(\varepsilon_1 \cdot \varepsilon_2|\mathscr{C}) \tag{11.33}$$

is called a version of the $\mathscr{C}$*-conditional covariance of* Y_1 *and* Y_2*.*

Remark 11.21 (X-Conditional Covariance) Let the assumptions of Definition 11.20 hold and let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable. Then

$$Cov(Y_1,Y_2\,|\,X) \;:=\; Cov\big(Y_1,Y_2\,|\,\sigma(X)\big) \tag{11.34}$$

is called a version of the *X-conditional covariance* of Y_1 and Y_2. ◁

The $\mathscr{C}$-conditional variance is defined analogously.

Definition 11.22 (Conditional Variance Given a σ-Algebra)
Let $Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be a real-valued random variable with $E(Y^2)<\infty$, let $\mathscr{C}\subset\mathscr{A}$ be a σ-algebra, and define $\varepsilon := Y-E(Y|\mathscr{C})$.

(i) Then

$$Var(Y|\mathscr{C}) \;:=\; E(\varepsilon^2|\mathscr{C}) \tag{11.35}$$

is called a version of the $\mathscr{C}$-conditional variance of Y.

(ii) Assume that $Var(Y|\mathscr{C})$ is a nonnegative version of the $\mathscr{C}$-conditional variance of Y. Then

$$SD(Y|\mathscr{C}) \;:=\; \sqrt{Var(Y|\mathscr{C})} \tag{11.36}$$

is called a version of the $\mathscr{C}$-conditional standard deviation of Y.

Remark 11.23 (Conditional Variance Given X) Let the assumptions of Definition 11.22 hold and let $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable. Then

$$Var(Y\,|\,X) \;:=\; Var\big(Y|\sigma(X)\big) \tag{11.37}$$

is called a version of the *X-conditional variance* of Y. Correspondingly, assuming that $Var(Y|X)$ is a nonnegative version of the X-conditional variance of Y, we call

$$SD(Y\,|\,X) \;:=\; \sqrt{Var(Y|X)} \tag{11.38}$$

a version of the *X-conditional standard deviation* of Y. ◁

11.4 Conditional Variance and Covariance Given a Value of a Random Variable

While the concepts defined above are random variables, the $(X{=}x)$-conditional covariance is a number. It is defined using the $(X{=}x)$-conditional expectation value $E(\varepsilon_1\cdot\varepsilon_2\,|\,X{=}x)$ that has been introduced as a value $g(x)$ of a factorization g of an X-conditional expectation $E(\varepsilon_1\cdot\varepsilon_2\,|\,X)=g(X)$ (see section 10.4.4).

Definition 11.24 ($(X=x)$-Conditional Variance and Covariance)
Let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ be a random variable.

(i) For $i=1,2$, let $Y_i\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be real-valued random variables with $E(Y_i^2)<\infty$, and let $\varepsilon_i := Y_i - E(Y_i|X)$. Then we call

$$Cov(Y_1,Y_2|X=x) := E(\varepsilon_1\cdot\varepsilon_2|X=x) \tag{11.39}$$

an $(X=x)$-conditional covariance of Y_1 and Y_2.

(ii) Let $Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be a real-valued random variable with $E(Y^2)<\infty$ and let $\varepsilon := Y-E(Y|X)$. Then we call

$$Var(Y|X=x) := E(\varepsilon^2|X=x) \tag{11.40}$$

an $(X=x)$-conditional variance of Y.

(iii) If, under the assumptions of (ii), $Var(Y|X)$ is a nonnegative version of the X-conditional variance of Y, then we call

$$SD(Y|X=x) := \sqrt{Var(Y|X=x)} \tag{11.41}$$

an $(X=x)$-conditional standard deviation of Y.

Remark 11.25 (Equivalent Propositions) Note that $Cov(Y_1,Y_2|X=x)$ is uniquely defined only if $P(X=x)>0$. However, even if $P(X=x)=0$ for all $x\in\Omega'_X$, then we can still make propositions such as

$$Cov(Y_1,Y_2|X=x) = Cov(Z_1,Z_2|X=x), \quad \text{for } P_X\text{-almost } x\in\Omega'_X, \tag{11.42}$$

provided that Z_1,Z_2 are real-valued random variables on $(\Omega,\mathscr{A},P)$ with finite second moments. According to Corollary 10.39, this proposition is equivalent to

$$Cov(Y_1,Y_2|X) \underset{P}{=} Cov(Z_1,Z_2|X). \tag{11.43}$$

Of course, the same applies to the X-conditional variance. ◁

Remark 11.26 (Values of the Conditional Covariance) As mentioned before, the term defined in Remark 11.21 is a *random variable*. Its values are

$$Cov(Y_1,Y_2|X)(\omega) = Cov(Y_1,Y_2|X=x), \quad \text{if } X(\omega)=x \tag{11.44}$$

(see Rem. 10.37). This also implies that the value of $Cov(Y_1,Y_2|X)$ is constant on all sets $\{X=x\}$. Similarly,

$$Var(Y|X)(\omega) = Var(Y|X=x), \quad \text{if } X(\omega)=x. \tag{11.45}$$

◁

Table 11.2. Joe and Ann With Self-Selection and Residuals

Outcomes ω				Observables			Conditional expectations				Residuals	
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y \mid X,U)$	$E(Y \mid X)$	$E(Y \mid U)$	$P(X=1 \mid U)$	$\varepsilon_Y = Y - E(Y \mid U)$	$\varepsilon_X = X - P(X=1 \mid U)$
Joe	*no*	−	.144	*Joe*	0	0	.70	.60	.704	.04	−.704	−.04
Joe	*no*	+	.336	*Joe*	0	1	.70	.60	.704	.04	.296	−.04
Joe	*yes*	−	.004	*Joe*	1	0	.80	.42	.704	.04	−.704	.96
Joe	*yes*	+	.016	*Joe*	1	1	.80	.42	.704	.04	.296	.96
Ann	*no*	−	.096	*Ann*	0	0	.20	.60	.352	.76	−.352	−.76
Ann	*no*	+	.024	*Ann*	0	1	.20	.60	.352	.76	.648	−.76
Ann	*yes*	−	.228	*Ann*	1	0	.40	.42	.352	.76	−.352	.24
Ann	*yes*	+	.152	*Ann*	1	1	.40	.42	.352	.76	.648	.24

Note. The probabilities of the elementary events are fictive

Example 11.27 (Covariance With Respect to the Measure $P^{X=x}$) Suppose that X represents *sex* with values *male* and *female*. Then $Cov(Y_1, Y_2 \mid X=x)$ is the covariance of Y_1 and Y_2 given $x = male$ or given $x = female$, or loosely speaking, the covariance of Y_1 and Y_2 in one of the two subpopulations of males and females. In more precise terms, if $P(X=x) > 0$, then $Cov(Y_1, Y_2 \mid X=x)$ is identical to the covariance $Cov^{X=x}(Y_1, Y_2)$ of Y_1 and Y_2 with respect to the conditional-probability measure $P^{X=x}$. This immediately follows from $Cov(Y_1, Y_2 \mid X=x) = E(\varepsilon_1 \cdot \varepsilon_2 \mid X=x) = E^{X=x}(\varepsilon_1 \cdot \varepsilon_2)$ (see Def. 9.2). ◁

Example 11.28 (Joe and Ann With Self-Selection – continued) For the example presented in Table 11.2 we consider the (unconditional) covariance of the treatment variable X and the outcome variable Y. Note that X and Y are indicator variables with values 0 and 1. Therefore, $E(X) = P(X=1)$, $E(Y) = P(Y=1)$, $E(X \cdot Y) = P(X=1, Y=1)$, and

$$\begin{aligned} Cov(X,Y) &= E(X \cdot Y) - E(X) \cdot E(Y) = P(X=1, Y=1) - P(X=1) \cdot P(Y=1) \\ &= (.016 + .152) - (.004 + .016 + .228 + .152) \cdot (.336 + .016 + .024 + .152) \\ &= .168 - .40 \cdot .528 = -0.0432. \end{aligned}$$

Hence, the treatment variable and the outcome variable have a *negative* covariance.

Now let us compute the $(U=u)$-conditional covariances of X and Y for $u=Joe$ and for $u=Ann$. First of all, note that $P(U=Joe) = .144+.336+.004+.016 = .50$ and $P(U=Ann) = .096+.024+.228+.152 = .50$. According to Equation (9.20) we have to sum the values of the product variable $\varepsilon_Y \cdot \varepsilon_X$ weighted by their $(U=u)$-conditional probabilities. Hence,

$$\begin{aligned} &E(\varepsilon_Y \cdot \varepsilon_X \,|\, U=Joe) \\ &\quad = .704 \cdot .04 \cdot \frac{.144}{.50} - .296 \cdot .04 \cdot \frac{.336}{.50} - .704 \cdot .96 \cdot \frac{.004}{.50} + .296 \cdot .96 \cdot \frac{.016}{.50} \\ &\quad = .00384. \end{aligned}$$

$$\begin{aligned} &E(\varepsilon_Y \cdot \varepsilon_X \,|\, U=Ann) \\ &\quad = .352 \cdot .76 \cdot \frac{.096}{.50} - .648 \cdot .76 \cdot \frac{.024}{.50} - .352 \cdot .24 \cdot \frac{.228}{.50} + .648 \cdot .24 \cdot \frac{.152}{.50} \\ &\quad = .03648. \end{aligned}$$

Both conditional covariances are *positive*. Hence, in this example, the (unconditional) covariance of X and Y (which is negative) is highly misleading if used to evaluate the effects of the treatment on success, because for both persons the $(U=u)$-conditional (or person-specific) covariances of X and Y are positive. ◁

11.5 Properties of Conditional Variances and Covariances

Boxes 11.2 and 11.3 summarize some important properties of conditional covariances and conditional variances. The rules for conditional variances are special cases of the corresponding rules for conditional covariances with $Y_1 = Y_2 = Y$ and $A = B$, respectively [see (xv)]. Hence we only have to prove the rules for the conditional covariances (see Exercise 11-5).

For $n=2$ variables Y_i and $m=2$ variables Z_j, Equation (xiv) of Box 11.2 can be written

$$\begin{aligned} &Cov(\alpha_1 Y_1 + \alpha_2 Y_2 \,,\, \beta_1 Z_1 + \beta_2 Z_2 \,|\, \mathscr{C}) \\ &\underset{P}{=} \alpha_1\beta_1 \, Cov(Y_1, Z_1 \,|\, \mathscr{C}) \;+\; \alpha_1\beta_2 \, Cov(Y_1, Z_2 \,|\, \mathscr{C}) \;+ \\ &\quad \alpha_2\beta_1 \, Cov(Y_2, Z_1 \,|\, \mathscr{C}) \;+\; \alpha_2\beta_2 \, Cov(Y_2, Z_2 \,|\, \mathscr{C}). \end{aligned} \tag{11.46}$$

Similarly, for two random variables Y_1 and Y_2, Rule (xi) of Box 11.3 can also be written

$$\begin{aligned} &Var(\alpha_1 Y_1 + \alpha_2 Y_2 \,|\, \mathscr{C}) \\ &\underset{P}{=} \alpha_1^2 \, Var(Y_1 \,|\, \mathscr{C}) \;+\; \alpha_2^2 \, Var(Y_2 \,|\, \mathscr{C}) \;+\; 2\alpha_1\alpha_2 \, Cov(Y_1, Y_2 \,|\, \mathscr{C}). \end{aligned} \tag{11.47}$$

Example 11.29 (Conditional Variance of an Indicator) Let $(\Omega, \mathscr{A}, P)$ be a probability space, let 1_A denote the indicator variable of $A \in \mathscr{A}$, and consider the random variable $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$. Then, according to Rule (xii) of Box 11.3,

$$Var(1_A \,|\, X) \;\underset{P}{=}\; P(A \,|\, X) \cdot [1 - P(A \,|\, X)], \tag{11.48}$$

Box 11.2 Rules of Computation for $\mathscr{C}$-Conditional Covariances

For $i = 1,2$, let $Y_i\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ be real-valued random variables with $E(Y_i^2) < \infty$. Furthermore, let $\alpha,\beta \in \mathbb{R}$ and let $\mathscr{C}_0,\mathscr{C} \subset \mathscr{A}$ be σ-algebras. Then the following properties hold for all real-valued versions of $E(Y_i\,|\,\mathscr{C})$ and all versions of the residuals $\varepsilon_i := Y_i - E(Y_i\,|\,\mathscr{C})$, $i = 1,2$:

$$E\big[\,Y_1 \cdot E(Y_2|\mathscr{C}_0)\,\big|\mathscr{C}\big] \underset{P}{=} E(Y_1|\mathscr{C}) \cdot E(Y_2|\mathscr{C}_0), \quad \text{if } \mathscr{C}_0 \subset \mathscr{C}. \tag{i}$$

$$Cov\,(Y_1, Y_2\,|\mathscr{C}) \underset{P}{=} E(Y_1 \cdot Y_2\,|\mathscr{C}) - E(Y_1|\mathscr{C}) \cdot E(Y_2|\mathscr{C})\,. \tag{ii}$$

$$Cov\,(Y_1, Y_2\,|\mathscr{C}) \underset{P}{=} Cov\,(\varepsilon_1, \varepsilon_2\,|\mathscr{C}). \tag{iii}$$

$$Cov\,(Y_1, Y_2\,|\mathscr{C}) \underset{P}{=} 0, \quad \text{if } Y_1 \underset{P}{=} \alpha. \tag{iv}$$

$$Cov\,(\alpha + Y_1\,, \beta + Y_2\,|\mathscr{C}) \underset{P}{=} Cov\,(Y_1, Y_2\,|\mathscr{C}). \tag{v}$$

$$Cov\,(\alpha\,Y_1\,, \beta\,Y_2\,|\mathscr{C}) \underset{P}{=} \alpha\beta\,Cov\,(Y_1, Y_2\,|\mathscr{C}). \tag{vi}$$

$$E\big[Cov\,(Y_1, Y_2\,|\mathscr{C})\,\big|\mathscr{C}_0\big] \underset{P}{=} E(\varepsilon_1 \cdot \varepsilon_2\,|\mathscr{C}_0) \underset{P}{=} Cov\,(Y_1, Y_2\,|\mathscr{C}_0), \quad \text{if } \mathscr{C}_0 \subset \mathscr{C}. \tag{vii}$$

$$E\,[Cov\,(Y_1, Y_2\,|\mathscr{C})\,] = E(\varepsilon_1 \cdot \varepsilon_2) = Cov\,(\varepsilon_1, \varepsilon_2). \tag{viii}$$

$$Cov\,(Y_1, Y_2) = Cov\,[E(Y_1|\mathscr{C}), E(Y_2\,|\mathscr{C})] + Cov\,(\varepsilon_1, \varepsilon_2). \tag{ix}$$

If $W\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ is a random variable and $E(W^2) < \infty$, then

$$Cov\,(\varepsilon_1, W\,|\mathscr{C}_0) \underset{P}{=} E(\varepsilon_1 \cdot W\,|\mathscr{C}_0), \quad \text{if } \mathscr{C}_0 \subset \mathscr{C}. \tag{x}$$

$$Cov\,(\varepsilon_1, W\,|\mathscr{C}) \underset{P}{=} 0, \quad \text{if } \sigma(W) \subset \mathscr{C}. \tag{xi}$$

$$Cov\,(Y_1, W\,|\mathscr{C}) \underset{P}{=} 0, \quad \text{if } \sigma(W) \subset \mathscr{C}. \tag{xii}$$

Let $W_1, W_2, Y_1, Y_2\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ be random variables such that $E(W_1^4)$, $E(W_2^4)$, $E(Y_1^4)$, $E(Y_2^4) < \infty$. If W_1 and W_2 are $\mathscr{C}$-measurable, then

$$Cov\,(W_1 \cdot Y_1\,, W_2 \cdot Y_2\,|\mathscr{C}) \underset{P}{=} W_1 \cdot W_2 \cdot Cov\,(Y_1, Y_2\,|\mathscr{C}). \tag{xiii}$$

For $i = 1,\ldots,n$, $j = 1,\ldots,m$, let $Y_i, Z_j\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ be random variables with $E(Y_i^2)$, $E(Z_j^2) < \infty$, and let $\alpha_i,\beta_j \in \mathbb{R}$. Then:

$$Cov\Big(\sum_{i=1}^{n} \alpha_i\,Y_i\,,\, \sum_{j=1}^{m} \beta_j\,Z_j\,\Big|\,\mathscr{C}\Big) \underset{P}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i\beta_j\,Cov\,(Y_i, Z_j\,|\mathscr{C}). \tag{xiv}$$

If $A, B \in \mathscr{A}$, then

$$Cov\,(1_A, 1_B\,|\mathscr{C}) \underset{P}{=} P(A \cap B\,|\mathscr{C}) - P(A\,|\mathscr{C}) \cdot P(B\,|\mathscr{C}). \tag{xv}$$

Box 11.3 Rules of Computation for Conditional Variances

Let $Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ denote a real-valued random variable with $E(Y^2) < \infty$. Furthermore, let $\alpha \in \mathbb{R}$ and let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra. Then the following properties hold for all real-valued versions of $E(Y|\mathscr{C})$ and all versions of the residual $\varepsilon := Y - E(Y|\mathscr{C})$:

$$Var(Y|\mathscr{C}) \underset{P}{=} E(Y^2|\mathscr{C}) - E(Y|\mathscr{C})^2 \tag{i}$$

$$Y \underset{P}{=} \alpha \quad \Rightarrow \quad Var(Y|\mathscr{C}) \underset{P}{=} 0 \tag{ii}$$

$$Var(\alpha + Y|\mathscr{C}) \underset{P}{=} Var(Y|\mathscr{C}) \tag{iii}$$

$$Var(\alpha Y|\mathscr{C}) \underset{P}{=} \alpha^2 Var(Y|\mathscr{C}) \tag{iv}$$

$$E[Var(Y|\mathscr{C})] = E(\varepsilon^2) = Var(\varepsilon) \tag{v}$$

$$Var(Y) = Var[E(Y|\mathscr{C})] + E[Var(Y|\mathscr{C})] \tag{vi}$$

$$= Var[E(Y|\mathscr{C})] + Var(\varepsilon). \tag{vii}$$

If we additionally assume that the random variable $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ is $\mathscr{C}$-measurable, then

$$Var(Y|X) \underset{P}{=} Var\big[E(Y|\mathscr{C})\,\big|\,X\big] + E\big[Var(Y|\mathscr{C})\,\big|\,X\big] \tag{viii}$$

$$\underset{P}{=} Var\big[E(Y|\mathscr{C})\,\big|\,X\big] + Var(\varepsilon|X). \tag{ix}$$

Let $X, Y\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be random variables with $E(X^4)$, $E(Y^4) < \infty$. If X is $\mathscr{C}$-measurable, then

$$Var(X \cdot Y|\mathscr{C}) \underset{P}{=} X^2 \cdot Var(Y|\mathscr{C}). \tag{x}$$

Let $Y_i\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be random variables with $E(Y_i^2) < \infty$ and $\alpha_i \in \mathbb{R}$, where $i = 1, \ldots, n$. Then:

$$Var\Big(\sum_{i=1}^{n} \alpha_i Y_i \,\Big|\, \mathscr{C}\Big) \underset{P}{=} \sum_{i=1}^{n} \alpha_i^2 Var(Y_i|\mathscr{C}) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \alpha_i \alpha_j Cov(Y_i, Y_j|\mathscr{C}). \tag{xi}$$

If $A \in \mathscr{A}$, then

$$Var(1_A|\mathscr{C}) \underset{P}{=} P(A|\mathscr{C}) \cdot [1 - P(A|\mathscr{C})]. \tag{xii}$$

which implies that the $(X{=}x)$-conditional variances $Var(1_A|X{=}x)$ depend on the values x of X unless $P(A|X) \underset{P}{=} P(A)$ (see Exercise 11-6). Hence, we cannot assume homogenous residual variances in statistical models for the analysis of the conditional expectation $E(1_A|X) \underset{P}{=} P(A|X)$ (see, e. g., Agresti, 2007). Furthermore, note that the X-conditional variance of an indicator variable does not contain any information additional to the X-conditional expectation $E(1_A|X) \underset{P}{=} P(A|X)$. ◁

11.6 Partial Correlation

Another concept used to describe a certain kind of dependence between two random variables Y_1 and Y_2 is the *partial correlation*, which is the correlation of the residuals of Y_1 and Y_2 with respect to the conditional expectations $E(Y_1|\mathscr{C})$ and $E(Y_2|\mathscr{C})$, respectively.

Definition 11.30 (Partial Correlation)
For $i = 1,2$, let $Y_i\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ be random variables with $E(Y_i^2) < \infty$, let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, and define $\varepsilon_i := Y_i - E(Y_i|\mathscr{C})$. Then we call

$$Corr(Y_1, Y_2; \mathscr{C}) := Corr(\varepsilon_1, \varepsilon_2) \tag{11.49}$$

the partial correlation of Y_1 and Y_2 given $\mathscr{C}$. If $X\colon (\Omega,\mathscr{A},P) \to (\Omega'_X,\mathscr{A}'_X)$ is a random variable, then we call

$$Corr(Y_1, Y_2; X) := Corr[Y_1, Y_2; \sigma(X)] \tag{11.50}$$

the partial correlation of Y_1 and Y_2 given X.

Remark 11.31 (Formulas for the Partial Correlation) If $Var(Y_1), Var(Y_2) > 0$ and $R_{Y_1|\mathscr{C}}, R_{Y_2|\mathscr{C}} < 1$, then

$$Corr(Y_1, Y_2; \mathscr{C}) = \frac{Corr(Y_1, Y_2) - R_{Y_1|\mathscr{C}} \cdot R_{Y_2|\mathscr{C}} \cdot Corr[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})]}{\sqrt{1 - R^2_{Y_1|\mathscr{C}}} \cdot \sqrt{1 - R^2_{Y_2|\mathscr{C}}}}, \tag{11.51}$$

where $R^2_{Y_i|\mathscr{C}} = Var[E(Y_i|\mathscr{C})]/Var(Y_i)$, $i = 1,2$, denotes the coefficient of determination (see Exercise 11-7). Similarly, if, for $i = 1,2$, there are versions $E(Y_i|X)$ with $E(Y_i|X) = \beta_{i0} + \beta_{i1}X$ with slopes $\beta_{i1} \neq 0$ and $Corr(Y_i, X)^2 < 1$, then Equation (11.51) simplifies to

$$Corr(Y_1, Y_2; X) = \frac{Corr(Y_1, Y_2) - Corr(Y_1, X) \cdot Corr(Y_2, X)}{\sqrt{1 - Corr(Y_1, X)^2} \cdot \sqrt{1 - Corr(Y_2, X)^2}} \tag{11.52}$$

(see Exercise 11-8). ◁

Remark 11.32 ($(X{=}x)$-Conditional Correlation) If $P(X{=}x) > 0$ and, for $i = 1,2$, $Var(Y_i \,|\, X{=}x) > 0$, then we define

$$Corr(Y_1, Y_2 \,|\, X{=}x) \;=\; \frac{Cov(Y_1, Y_2 \,|\, X{=}x)}{\sqrt{Var(Y_1 \,|\, X{=}x)} \cdot \sqrt{Var(Y_2 \,|\, X{=}x)}}, \tag{11.53}$$

and call it the *$(X{=}x)$-conditional correlation of Y_1 and Y_2*. If $Var(Y_1 \,|\, X{=}x) = 0$ or $Var(Y_2 \,|\, X{=}x) = 0$ we define $Corr(Y_1, Y_2 \,|\, X{=}x) = 0$. ◁

Remark 11.33 (Interpretation of the Partial Correlation) In a sense, the partial correlation is an average correlation across the *$(X{=}x)$-conditional correlations*. More precisely,

$$Corr(Y_1, Y_2 \,;\, X) \;=\; \frac{E[Cov(Y_1, Y_2 \,|\, X)]}{\sqrt{E[Var(Y_1 \,|\, X)]} \cdot \sqrt{E[Var(Y_2 \,|\, X)]}}. \tag{11.54}$$

This equation immediately follows from the definition of $Corr(Y_1, Y_2 \,;\, X)$, Rule (v) of Box 11.2, and Rule (viii) of Box 11.3. ◁

Example 11.34 (Joe and Ann With Self-Selection – continued) We compute the partial correlation $Corr(Y, X; U)$ in the example presented in Table 11.2. For this purpose we use Equations (11.49), (11.50), and (7.17). The covariance $Cov(\varepsilon_Y, \varepsilon_X)$ of the two residuals is

$$\begin{aligned} Cov(\varepsilon_Y, \varepsilon_X) &= E(\varepsilon_Y \cdot \varepsilon_X) \\ &= (-.704 \cdot -.04) \cdot .144 + (.296 \cdot -.04) \cdot .336 + (-.704 \cdot .96) \cdot .004 \\ &\quad + (.296 \cdot .96) \cdot .016 + (-.352 \cdot -.76) \cdot .096 + (.648 \cdot -.76) \cdot .024 \\ &\quad + (-.352 \cdot .24) \cdot .228 + (.648 \cdot .24) \cdot .152 \;=\; 0.02016, \end{aligned}$$

the variance $Var(\varepsilon_Y)$ of ε_Y is

$$\begin{aligned} Var(\varepsilon_Y) \;=\; E(\varepsilon_Y^2) &= (-.704)^2 \cdot .144 + .296^2 \cdot .336 + (-.704)^2 \cdot .004 \\ &\quad + .296^2 \cdot .016 + (-.352)^2 \cdot .096 + .648^2 \cdot .024 \\ &\quad + (-.352)^2 \cdot .228 + .648^2 \cdot .152 \;=\; 0.21824, \end{aligned}$$

and the variance $Var(\varepsilon_X)$ of ε_X is

$$\begin{aligned} Var(\varepsilon_X) \;=\; E(\varepsilon_X^2) &= .04^2 \cdot .144 + .04^2 \cdot .336 + .96^2 \cdot .004 + .96^2 \cdot .016 \\ &\quad + .76^2 \cdot .096 + .76^2 \cdot .024 + .24^2 \cdot .228 + .24^2 \cdot .152 \;=\; 0.1104. \end{aligned}$$

Hence,

$$Corr(Y, X; U) \;=\; \frac{Cov(\varepsilon_Y, \varepsilon_X)}{SD(\varepsilon_Y) \cdot SD(\varepsilon_X)} \;=\; \frac{0.02016}{\sqrt{0.21824} \cdot \sqrt{0.1104}} \;=\; 0.129879,$$

which is a *positive* number. Again, this indicates that using the (unconditional) covariance of X and Y or the (unconditional) correlation—which both are negative—for the evaluation of the effects of the treatment on success would be highly misleading. ◁

11.7 Proofs

Proof of Theorem 11.18

(i) Equation (11.26) can be derived as follows:

$$\begin{aligned}
E(Y|X) &\underset{P}{=} E[E(Y|X,Z)|X] && \text{[Box 10.2 (v)]}\\
&\underset{P}{=} E(\beta_0+\beta_1 X+\beta_2 Z|X) && \text{[(11.24)]}\\
&\underset{P}{=} \beta_0+\beta_1 X+\beta_2 E(Z|X) && \text{[Box 10.2 (xvi), (vii)]}\\
&\underset{P}{=} \beta_0+\beta_1 X+\beta_2(\gamma_0+\gamma_1 X) && \text{[(11.25)]}\\
&\underset{P}{=} (\beta_0+\beta_2\gamma_0)+(\beta_1+\gamma_1)X\\
&\underset{P}{=} \alpha_0+\alpha_1 X,
\end{aligned}$$

with $\alpha_0 := \beta_0+\beta_2\gamma_0$ and $\alpha_1 := \beta_1+\gamma_1$.

(ii) If $E(Z|X) \underset{P}{=} E(Z)$, then the third line of the equations above yields Equations (11.27) and (11.28). If $\beta_2 = 0$, then the proof of (i) shows that $E(Y|X) \underset{P}{=} \beta_0+\beta_1 X \underset{P}{=} \alpha_0+\alpha_1 X$ which proves the proposition.

(iii) If $E(X|Z) \underset{P}{=} E(X)$ or $\beta_1 = 0$, then

$$\begin{aligned}
E(Y|Z) &\underset{P}{=} E[E(Y|X,Z)|Z] && \text{[Box 10.2 (v)]}\\
&\underset{P}{=} E(\beta_0+\beta_1 X+\beta_2 Z|Z) && \text{[(11.24)]}\\
&\underset{P}{=} \beta_0+\beta_1 E(X|Z)+\beta_2 Z && \text{[Box 10.2 (xvi), (vii)]}\\
&\underset{P}{=} \beta_0+\beta_1 E(X)+\beta_2 Z. && [E(X|Z) \underset{P}{=} E(X) \text{ or } \beta_1 = 0]
\end{aligned}$$

Hence, our assumption implies $Var[E(Y|Z)] = \beta_2^2 Var(Z)$ [Box 6.2 (ii), (iii)]. Now consider

$$\begin{aligned}
Var[E(Y|X,Z)] &= Var(\beta_0+\beta_1 X+\beta_2 Z) && \text{[(11.24)]}\\
&= \beta_1^2 Var(X)+\beta_2^2 Var(Z)+2\beta_1\beta_2 Cov(X,Z). && \text{[Box 7.1 (viii)]}
\end{aligned}$$

Assuming $[E(Z|X) \underset{P}{=} E(Z)$ or $\beta_2 = 0]$ and $[E(X|Z) \underset{P}{=} E(X)$ or $\beta_1 = 0]$ yields $2\beta_1\beta_2 Cov(X,Z) = 0$, because $E(Z|X) \underset{P}{=} E(Z)$ and $E(X|Z) \underset{P}{=} E(X)$ both imply $Cov(X,Z) = 0$ [see Eq. (10.49)]. Hence,

$$\begin{aligned}
Var[E(Y|X,Z)] &= \beta_1^2 Var(X)+\beta_2^2 Var(Z)\\
&= \alpha_1^2 Var(X)+\beta_2^2 Var(Z) && \text{[(11.28)]}\\
&= Var[E(Y|X)]+Var[E(Y|Z)],
\end{aligned}$$

because $Var[E(Y|X)] = \alpha_1^2 Var(X)$ and $Var[E(Y|Z)] = \beta_2^2 Var(Z)$. Now Equation (11.30) follows dividing both sides by $Var(Y)$ and using the definition of the coefficient of determination.

11.8 Exercises

▷ **Exercise 11-1** Prove the rules of computation for the residual $\varepsilon := Y - E(Y|\mathscr{C})$ summarized in Box 11.1.

▷ **Exercise 11-2** Show that for $\varepsilon := Y - E(Y|X,Z)$, the equations $Cov(X,\varepsilon) = Cov(Z,\varepsilon) = 0$ are special cases of Rule (viii) of Box 11.1 if we consider the conditional expectation $E(Y|X,Z)$, assume X and Z to be numerical, and $E(Y^2)$, $E(X^2)$, $E(Z^2) < \infty$.

▷ **Exercise 11-3** Prove Equations (11.16) and (11.17).

▷ **Exercise 11-4** Show that $R^2_{Y|\mathscr{C}} = Corr[Y, E(Y|\mathscr{C})]^2$.

▷ **Exercise 11-5** Prove the rules of Box 11.2.

▷ **Exercise 11-6** Show: If Y is a dichotomous random variable on $(\Omega, \mathscr{A}, P)$ with values 0 and 1, and we assume $P(X{=}x) > 0$, then $Var(Y \mid X{=}x) = P(Y{=}1 \mid X{=}x) \cdot [1 - P(Y{=}1 \mid X{=}x)]$. Furthermore, if $P(X{=}x_1), P(X{=}x_2) > 0$, $P(Y{=}1 \mid X{=}x_1) \neq P(Y{=}1 \mid X{=}x_2)$ and $P(Y{=}1 \mid X{=}x_1) \neq 1 - P(Y{=}1 \mid X{=}x_2)$, then $Var(Y|X{=}x_1) \neq Var(Y|X{=}x_2)$.

▷ **Exercise 11-7** Show that Equation (11.51) holds for $Corr(Y_1, Y_2; \mathscr{C})$.

▷ **Exercise 11-8** Prove Equation (11.52).

Solutions

▷ **Solution 11-1** (i) This rule directly follows from Theorem 10.9 (ii) and Proposition (2.36).

(ii) This rule directly follows from the definition $\varepsilon := Y - E(Y|\mathscr{C})$ and the assumption that Y and $E(Y|\mathscr{C})$ are real-valued.

(v)

$$\begin{aligned} \varepsilon &\underset{P}{=} Y - E(Y|\mathscr{C}) && \text{[Box 11.1 (i)]} \\ &\underset{P}{=} Y - Y \;=\; 0. && [Y \underset{P}{=} E(Y|\mathscr{C})] \end{aligned}$$

(vi) This rule can be derived as follows:

$$\begin{aligned} E(\varepsilon|\mathscr{C}_0) &\underset{P}{=} E[Y - E(Y|\mathscr{C})|\mathscr{C}_0] && \text{[def. of } \varepsilon\text{]} \\ &\underset{P}{=} E(Y|\mathscr{C}_0) - E[E(Y|\mathscr{C})|\mathscr{C}_0] && \text{[Box 10.1 (xvi)]} \\ &\underset{P}{=} E(Y|\mathscr{C}_0) - E(Y|\mathscr{C}_0) && \text{[Box 10.1 (v)]} \\ &\underset{P}{=} 0. \end{aligned}$$

(iii) This rule is a special case of Rule (vi) for $\mathscr{C}_0 = \{\Omega, \emptyset\}$.

(vii) This rule is a special case of Rule (vi) for $\mathscr{C}_0 := \sigma(W) \subset \mathscr{C}$.

(viii) Note that $E(Y^2) < \infty$ implies $E[E(Y|\mathscr{C})^2] < \infty$ [see Box 10.1 (xi)], which in turn implies $E(\varepsilon^2) < \infty$ (see Rem. 7.1).

$$\begin{aligned} Cov(\varepsilon, W) &= E(\varepsilon \cdot W) - E(\varepsilon) \cdot E(W) && \text{[Box 7.1 (i)]} \\ &= E(\varepsilon \cdot W) - 0 && \text{[Box 11.1 (iii)]} \\ &= E[E(\varepsilon \cdot W|\mathscr{C})] && \text{[Box 10.1 (v)]} \\ &= E[W \cdot E(\varepsilon|\mathscr{C})] && \text{[Box 10.1 (xiv)]} \\ &= E(W \cdot 0) && \text{[Box 11.1 (vi)]} \\ &= E(0) \;=\; 0. && \text{[Box 6.1 (i)]} \end{aligned}$$

The second equation of (viii) following from Box 6.2 (i).

(iv)

$$\begin{aligned} Var(Y) &= Var[E(Y|\mathscr{C})+\varepsilon] && \text{[Box 11.1 (ii)]} \\ &= Var[E(Y|\mathscr{C})] + Var(\varepsilon) + 2\cdot Cov[E(Y|\mathscr{C}),\varepsilon] && \text{[Box 7.1 (viii)]} \\ &= Var[E(Y|\mathscr{C})] + Var(\varepsilon). && \text{[Box 11.1 (viii)]} \end{aligned}$$

In the last equation we used the fact that $E(Y|\mathscr{C})$ is $\mathscr{C}$-measurable, thus taking the role of W in Rule (viii) of Box 11.1.

(ix) Note that $E(Y^2) < \infty$ implies $E[E(Y|\mathscr{C})^2] < \infty$ [see Box 10.1 (xi)]. Hence,

$$\begin{aligned} Cov(W,Y) &= Cov[W, E(Y|\mathscr{C})+\varepsilon] && \text{[Box 11.1 (ii)]} \\ &= Cov[W, E(Y|\mathscr{C})] + Cov(W,\varepsilon) && \text{[Box 7.1 (ix)]} \\ &= Cov[W, E(Y|\mathscr{C})] + 0. && \text{[Box 11.1 (viii)]} \end{aligned}$$

▷ **Solution 11-2** $E(Y|X,Z) \underset{P}{=} E(Y|\mathscr{C})$ with $\mathscr{C} := \sigma(X,Z)$. Hence, X and Z are both $\mathscr{C}$-measurable [see Eq. (2.17)] and Rule (viii) of Box 11.1 applies if we assume $E(Y^2)$, $E(X^2)$, $E(Z^2) < \infty$.

▷ **Solution 11-3** Assume that $Var(X), Var(Y) > 0$. Then

$$\begin{aligned} Cov(X,Y) &= Cov(X, \beta_0 + \beta_1 X + \varepsilon) && \text{[(11.15), (11.1), (10.1)]} \\ &= \beta_1 Var(X). && \text{[Box 7.1 (ii), (iii), (iv)]} \end{aligned}$$

Dividing both sides by $Var(X)$ yields Equation (11.16). Now

$$\begin{aligned} R^2_{Y|X} &= \frac{Var[E(Y|X)]}{Var(Y)} && \text{[(11.13)]} \\ &= \frac{Var(\beta_0 + \beta_1 X)}{Var(Y)} && \text{[(11.15)]} \\ &= \frac{\beta_1^2 Var(X)}{Var(Y)} && \text{[Box 6.2 (ii), (iii)]} \\ &= \frac{\left(\frac{Cov(X,Y)}{Var(X)}\right)^2 Var(X)}{Var(Y)} && \text{[(11.16)]} \\ &= \frac{Cov(X,Y)^2}{Var(X)\cdot Var(Y)} \\ &= Corr(X,Y)^2. && \text{[(7.17)]} \end{aligned}$$

If $Var(Y) = 0$, then $R^2_{Y|X} = Corr(X,Y) = 0$ by the definitions of the two terms. If $Var(X) = 0, Var(Y) > 0$, then $Corr(X,Y) = 0$ by definition and $Var[E(Y|X)] = Var(\beta_0 + \beta_1 X) = \beta_1^2 Var(X) = 0$, and Equation (11.13) implies $R^2_{Y|X} = 0$ as well. Hence, in both cases Equation (11.17) holds.

▷ **Solution 11-4** Assume that $Var(Y), Var[E(Y|\mathscr{C})] > 0$. Then

$$\begin{aligned}
Corr[Y, E(Y|\mathscr{C})] &= \frac{Cov[Y, E(Y|\mathscr{C})]}{SD(Y)\cdot SD[E(Y|\mathscr{C})]} && [(7.17)]\\
&= \frac{Var[E(Y|\mathscr{C})]}{SD(Y)\cdot SD[E(Y|\mathscr{C})]} && [\text{Box 10.1 (xii)}]\\
&= \frac{SD[E(Y|\mathscr{C})]\cdot SD[E(Y|\mathscr{C})]}{SD(Y)\cdot SD[E(Y|\mathscr{C})]}\\
&= \frac{SD[E(Y|\mathscr{C})]}{SD(Y)}.
\end{aligned}$$

Squaring both sides and inserting the definition of $R^2_{Y|\mathscr{C}}$ yields $R^2_{Y|\mathscr{C}} = Corr[Y, E(Y|\mathscr{C})]^2$. If $Var(Y) = 0$, then $Corr[Y, E(Y|\mathscr{C})]^2 = 0 = R^2_{Y|\mathscr{C}}$ by the definitions of $Corr[Y, E(Y|\mathscr{C})]$ and $R^2_{Y|\mathscr{C}}$. If $Var[E(Y|\mathscr{C})] = 0$, then $Corr[Y, E(Y|\mathscr{C})]^2 = 0$ by definition of $Corr[Y, E(Y|\mathscr{C})]$, and $R^2_{Y|\mathscr{C}} = 0$, either because $Var[E(Y|\mathscr{C})] = 0$, or, if $Var(Y) = 0$, by definition of $R^2_{Y|\mathscr{C}}$.

▷ **Solution 11-5** (i) If $\mathscr{C}_0 \subset \mathscr{C}$, then $E(Y_2|\mathscr{C}_0)$ is $\mathscr{C}$-measurable. Therefore, Box 10.1 (xiv) yields

$$E\left[\, Y_1 \cdot E(Y_2|\mathscr{C}_0) \,\middle|\, \mathscr{C}\right] \underset{P}{=} E(Y_1|\mathscr{C}) \cdot E(Y_2|\mathscr{C}_0).$$

(ii)

$$\begin{aligned}
&Cov(Y_1, Y_2|\mathscr{C})\\
&\underset{P}{=} E(\varepsilon_1 \cdot \varepsilon_2|\mathscr{C}) && [(11.33)]\\
&\underset{P}{=} E\left([Y_1 - E(Y_1|\mathscr{C})]\cdot[Y_2 - E(Y_2|\mathscr{C})]\,\middle|\,\mathscr{C}\right) && [\text{def. of } \varepsilon_i]\\
&\underset{P}{=} E\left[\, Y_1\cdot Y_2 - Y_1\cdot E(Y_2|\mathscr{C}) - E(Y_1|\mathscr{C})\cdot Y_2 + E(Y_1|\mathscr{C})\cdot E(Y_2|\mathscr{C}) \,\middle|\, \mathscr{C}\right]\\
&\underset{P}{=} E(Y_1\cdot Y_2|\mathscr{C}) - E[Y_1\cdot E(Y_2|\mathscr{C})|\mathscr{C}] - E[E(Y_1|\mathscr{C})\cdot Y_2|\mathscr{C}]\\
&\quad + E[E(Y_1|\mathscr{C})\cdot E(Y_2|\mathscr{C})|\mathscr{C}] && [\text{Box 10.1 (xvi)}]\\
&\underset{P}{=} E(Y_1\cdot Y_2|\mathscr{C}) - E(Y_1|\mathscr{C})\cdot E(Y_2|\mathscr{C}). && [\text{Box 11.2 (i)}]
\end{aligned}$$

(iii)

$$\begin{aligned}
Cov(\varepsilon_1, \varepsilon_2|\mathscr{C}) &\underset{P}{=} E(\varepsilon_1\cdot\varepsilon_2|\mathscr{C}) - E(\varepsilon_1|\mathscr{C})\cdot E(\varepsilon_2|\mathscr{C}) && [\text{Box 11.2 (ii)}]\\
&\underset{P}{=} E(\varepsilon_1\cdot\varepsilon_2|\mathscr{C}) && [\text{Box 11.1 (vi)}]\\
&\underset{P}{=} Cov(Y_1, Y_2|\mathscr{C}). && [(11.33)]
\end{aligned}$$

(iv)

$$\begin{aligned}
Cov(Y_1, Y_2|\mathscr{C}) &\underset{P}{=} E(Y_1\cdot Y_2|\mathscr{C}) - E(Y_1|\mathscr{C})\cdot E(Y_2|\mathscr{C}) && [\text{Box 11.2 (ii)}]\\
&\underset{P}{=} E(\alpha\, Y_2|\mathscr{C}) - \alpha\, E(Y_2|\mathscr{C}) && [Y_1 \underset{P}{=} \alpha, \text{ Box 10.1 (i)}]\\
&\underset{P}{=} \alpha\, E(Y_2|\mathscr{C}) - \alpha\, E(Y_2|\mathscr{C}) \underset{P}{=} 0. && [\text{Box 10.1 (ii)}]
\end{aligned}$$

(v)

$$\begin{aligned}
&Cov(\alpha + Y_1, \beta + Y_2|\mathscr{C})\\
&\underset{P}{=} E\Big(\big[\alpha + Y_1 - E(\alpha + Y_1|\mathscr{C})\big]\cdot\big[\beta + Y_2 - E(\beta + Y_2|\mathscr{C})\big]\,\Big|\,\mathscr{C}\Big) && [(11.33)]\\
&\underset{P}{=} E\Big(\big[\, Y_1 - E(Y_1|\mathscr{C})\big]\cdot\big[\, Y_2 - E(Y_2|\mathscr{C})\big]\,\Big|\,\mathscr{C}\Big) && [\text{Box 10.1 (ii)}]\\
&\underset{P}{=} E(\varepsilon_1\cdot\varepsilon_1|\mathscr{C}) && [\text{def. of } \varepsilon_i]\\
&\underset{P}{=} Cov(Y_1, Y_2|\mathscr{C}). && [(11.33)]
\end{aligned}$$

(vi)

$$
\begin{aligned}
& Cov(\alpha\, Y_1, \beta\, Y_2 \,|\mathscr{C}) \\
\underset{P}{=}\ & E\Big(\big[\alpha\, Y_1 - E(\alpha\, Y_1|\mathscr{C})\big]\cdot\big[\beta\, Y_2 - E(\beta\, Y_2|\mathscr{C})\big]\,\Big|\,\mathscr{C}\Big) && \text{[(11.33)]} \\
\underset{P}{=}\ & E\Big(\big[\alpha\cdot[Y_1 - E(Y_1|\mathscr{C})]\big]\cdot\big[\beta\cdot[Y_2 - E(Y_2|\mathscr{C})]\big]\,\Big|\,\mathscr{C}\Big) && \text{[Box 10.1 (iii)]} \\
\underset{P}{=}\ & E(\alpha\,\varepsilon_1\cdot\beta\,\varepsilon_2|\mathscr{C}) && \text{[def. of } \varepsilon_i\text{]} \\
\underset{P}{=}\ & \alpha\,\beta\, E(\varepsilon_1\cdot\varepsilon_2|\mathscr{C}) && \text{[Box 10.1 (iii)]} \\
\underset{P}{=}\ & \alpha\,\beta\, Cov(Y_1, Y_2|\mathscr{C}). && \text{[(11.33)]}
\end{aligned}
$$

(vii)

$$
\begin{aligned}
Cov(Y_1, Y_2|\mathscr{C}_0) &\underset{P}{=} E(\varepsilon_1\cdot\varepsilon_2|\mathscr{C}_0) && \text{[(11.33)]} \\
&\underset{P}{=} E\big[E(\varepsilon_1\cdot\varepsilon_2|\mathscr{C})\,\big|\,\mathscr{C}_0\big] && \text{[Box 10.1 (v)]} \\
&\underset{P}{=} E\big[Cov(Y_1, Y_2|\mathscr{C})\,\big|\,\mathscr{C}_0\big]. && \text{[(11.33)]}
\end{aligned}
$$

(viii)

$$
\begin{aligned}
E\big[Cov(Y_1, Y_2|\mathscr{C})\big] &= E\big[E(\varepsilon_1\cdot\varepsilon_2|\mathscr{C})\big] && \text{[(11.33)]} \\
&= E(\varepsilon_1\cdot\varepsilon_2) && \text{[Box 10.1 (iv)]} \\
&= Cov(\varepsilon_1, \varepsilon_2). && \text{[Box 7.1 (i), Box 11.1 (iii)]}
\end{aligned}
$$

(ix)

$$
\begin{aligned}
Cov(Y_1, Y_2) &= Cov\big(E(Y_1|\mathscr{C}) + \varepsilon_1, E(Y_2|\mathscr{C}) + \varepsilon_2\big) && \text{[Box 11.1 (ii)]} \\
&= Cov\big(E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})\big) + Cov\big(E(Y_1|\mathscr{C}), \varepsilon_2\big) \\
&\quad + Cov\big(\varepsilon_1, E(Y_2|\mathscr{C})\big) + Cov(\varepsilon_1, \varepsilon_2) && \text{[Box 7.1 (ix)]} \\
&= Cov\big(E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})\big) + Cov(\varepsilon_1, \varepsilon_2). && \text{[Box 11.1 (viii)]}
\end{aligned}
$$

(x)

$$
\begin{aligned}
Cov(\varepsilon_1, W|\mathscr{C}_0) &\underset{P}{=} E\big([\varepsilon_1 - E(\varepsilon_1|\mathscr{C}_0)]\cdot[W - E(W|\mathscr{C}_0)]\,\big|\,\mathscr{C}_0\big) && \text{[(11.33)]} \\
&\underset{P}{=} E\big(\varepsilon_1\cdot[W - E(W|\mathscr{C}_0)]\,\big|\,\mathscr{C}_0\big) && \text{[}\mathscr{C}_0\subset\mathscr{C}\text{, Box 11.1 (vi)]} \\
&\underset{P}{=} E\big(\varepsilon_1\cdot W - \varepsilon_1\cdot E(W|\mathscr{C}_0)\,\big|\,\mathscr{C}_0\big) \\
&\underset{P}{=} E(\varepsilon_1\cdot W|\mathscr{C}_0) - E\big(\varepsilon_1\cdot E(W|\mathscr{C}_0)\,\big|\,\mathscr{C}_0\big) && \text{[Box 10.1 (xvi)]} \\
&\underset{P}{=} E(\varepsilon_1\cdot W|\mathscr{C}_0) - E\big(\varepsilon_1|\mathscr{C}_0\big)\cdot E(W|\mathscr{C}_0) && \text{[Box 10.1 (xiv)]} \\
&\underset{P}{=} E(\varepsilon_1\cdot W|\mathscr{C}_0). && \text{[Box 11.1 (vi)]}
\end{aligned}
$$

(xi)

$$
\begin{aligned}
Cov(\varepsilon_1, W|\mathscr{C}) &\underset{P}{=} E(\varepsilon_1\cdot W|\mathscr{C}). && \text{[Box 11.2 (x)]} \\
&\underset{P}{=} E(\varepsilon_1|\mathscr{C})\cdot W && \text{[}\sigma(W)\subset\mathscr{C}\text{, Box 10.1 (xiv)]} \\
&\underset{P}{=} 0. && \text{[Box 11.1 (vi)]}
\end{aligned}
$$

(xii)

$$
\begin{aligned}
& Cov(Y_1, W \,|\mathscr{C}) \\
\underset{P}{=}\ & E\Big(\big[\,Y_1 - E(Y_1|\mathscr{C})\,\big]\cdot\big[W - E(W\,|\mathscr{C})\big]\,\Big|\mathscr{C}\Big) && \text{[(11.33)]} \\
\underset{P}{=}\ & E\big(\,Y_1 - E(Y_1|\mathscr{C})\,\big|\mathscr{C}\big)\cdot\big(W - E(W\,|\mathscr{C})\big) && \text{[}\sigma(W)\subset\mathscr{C}\text{, Box 10.1 (xiv)]} \\
\underset{P}{=}\ & 0. && \text{[Box 11.1 (vi)]}
\end{aligned}
$$

(xiii) For $i = 1,2$, consider the residual of $W_i \cdot Y_i$ with respect to its $\mathscr{C}$-conditional expectation,

$$W_i \cdot Y_i - E(W_i \cdot Y_i \,|\mathscr{C}) \underset{P}{=} W_i \cdot Y_i - W_i E(Y_i \,|\mathscr{C}) \underset{P}{=} W_i \cdot [Y_i - E(Y_i \,|\mathscr{C})] \underset{P}{=} W_i \cdot \varepsilon_i.$$

This equation implies

$$
\begin{aligned}
Cov(W_1 \cdot Y_1, W_2 \cdot Y_2 \,|\mathscr{C}) &\underset{P}{=} E(W_1 \cdot \varepsilon_1 \cdot W_2 \cdot \varepsilon_2 \,|\mathscr{C}) && \text{[(11.33)]} \\
&\underset{P}{=} W_1 \cdot W_2 \cdot E(\varepsilon_1 \cdot \varepsilon_2 \,|\mathscr{C}) && \text{[Box 10.1 (xiv)]} \\
&\underset{P}{=} W_1 \cdot W_2 \cdot Cov(Y_1, Y_2 \,|\mathscr{C}). && \text{[(11.33)]}
\end{aligned}
$$

(xiv) Define

$$\varepsilon_i := Y_i - E(Y_i|\mathscr{C}), \quad i = 1,\ldots,n, \quad \text{and} \quad \delta_j := Z_j - E(Y_j|\mathscr{C}), \quad j = 1,\ldots,m.$$

Then

$$
\begin{aligned}
\sum_{i=1}^{n} \alpha_i Y_i \;-\; E\left(\sum_{i=1}^{n} \alpha_i Y_i \,\middle|\mathscr{C}\right) &\underset{P}{=} \sum_{i=1}^{n} \alpha_i [Y_i - E(Y_i|\mathscr{C})] && \text{[Box 10.1 (xvi)]} \\
&\underset{P}{=} \sum_{i=1}^{n} \alpha_i \varepsilon_i,
\end{aligned}
$$

and, analogously,

$$\sum_{j=1}^{m} \beta_j Z_j \;-\; E\left(\sum_{j=1}^{m} \beta_j Z_j \,\middle|\mathscr{C}\right) = \sum_{j=1}^{m} \beta_j \delta_j.$$

Hence,

$$
\begin{aligned}
Cov\left(\sum_{i=1}^{n} \alpha_i Y_i,\, \sum_{j=1}^{m} \beta_j Z_j \,\middle|\mathscr{C}\right) &\underset{P}{=} E\left[\left(\sum_{i=1}^{n} \alpha_i \varepsilon_i\right)\cdot\left(\sum_{j=1}^{m} \beta_j \delta_j\right)\,\middle|\mathscr{C}\right] && \text{[(11.33)]} \\
&\underset{P}{=} E\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j \cdot \varepsilon_i \delta_j \,\middle|\mathscr{C}\right) \\
&\underset{P}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j \, E\left(\varepsilon_i \cdot \delta_j \,|\mathscr{C}\right) && \text{[Box 10.1 (xvi)]} \\
&\underset{P}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j \, Cov\left(Y_i, Z_j \,|\mathscr{C}\right). && \text{[(11.33)]}
\end{aligned}
$$

(xv)

$$
\begin{aligned}
Cov(1_A, 1_B \,|\mathscr{C}) &\underset{P}{=} E(1_A 1_B \,|\mathscr{C}) - E(1_A \,|\mathscr{C}) \cdot E(1_B \,|\mathscr{C}) && \text{[Box 11.2 (ii)]} \\
&\underset{P}{=} P(A \cap B \,|\mathscr{C}) - P(A \,|\mathscr{C}) \cdot P(B \,|\mathscr{C}). && \text{[(1.32), (10.2)]}
\end{aligned}
$$

▷ **Solution 11-6** If $P(X{=}x) > 0$, then (11.48) and Remark 10.35 imply

$$Var(Y|X{=}x) \;=\; P(Y{=}1\,|\,X{=}x)\cdot[1-P(Y{=}1\,|\,X{=}x)].$$

Using some algebra shows that this equation implies

$$Var(Y\,|\,X{=}x_1) \;\neq Var(Y\,|\,X{=}x_2),$$

provided that $P(X{=}x_1), P(X{=}x_2) > 0$, $P(Y{=}1\,|\,X{=}x_1) \neq P(Y{=}1\,|\,X{=}x_2)$, and $P(Y{=}1\,|\,X{=}x_1) \neq 1-P(Y{=}1\,|\,X{=}x_2)$.

▷ **Solution 11-7** Because $E(Y_1|\mathscr{C})$ and $E(Y_2|\mathscr{C})$ are $\mathscr{C}$-measurable, Rule (ix) of Box 11.1 implies

$$Cov\,[Y_1, E(Y_2|\mathscr{C})] \;=\; Cov\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})] \;=\; Cov\,[Y_2, E(Y_1|\mathscr{C})].$$

Therefore,

$$\begin{aligned}
&Cov\,(\varepsilon_1,\varepsilon_2)\\
&= Cov\,[Y_1 - E(Y_1|\mathscr{C}), Y_2 - E(Y_2|\mathscr{C})]\\
&= Cov\,(Y_1,Y_2) + Cov\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})] -\\
&\quad Cov\,[Y_1, E(Y_2|\mathscr{C})] - Cov\,[Y_2, E(Y_1|\mathscr{C})] && \text{[Box 7.1 (ix)]}\\
&= Cov\,(Y_1,Y_2) - Cov\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})]\\
&= SD(Y_1)\cdot SD(Y_2)\cdot Corr\,(Y_1,Y_2) - Cov\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})] && \text{[(7.17)]}\\
&= SD(Y_1)\cdot SD(Y_2)\cdot \left[Corr\,(Y_1,Y_2) - \frac{Cov\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})]}{SD(Y_1)\cdot SD(Y_2)} \right]\\
&= SD(Y_1)\cdot SD(Y_2)\cdot \left[Corr\,(Y_1,Y_2) - R_{Y_1|\mathscr{C}} R_{Y_2|\mathscr{C}} \cdot \frac{Cov\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})]}{SD[E(Y_1|\mathscr{C})]\cdot SD[E(Y_2|\mathscr{C})]} \right] && \text{[(11.21)]}\\
&= SD(Y_1)\cdot SD(Y_2)\cdot \big[Corr\,(Y_1,Y_2) - R_{Y_1|\mathscr{C}} R_{Y_2|\mathscr{C}} \cdot Corr\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})] \big].
\end{aligned}$$

Furthermore, for $i = 1,2$,

$$\begin{aligned}
SD(\varepsilon_i) \;&=\; \sqrt{Var(\varepsilon_i)} \;=\; \sqrt{Var(Y_i) - Var[E(Y_i|\mathscr{C})]} && \text{[Box 11.1 (iv)]}\\
&= \sqrt{Var(Y_i) - Var(Y_i)\cdot R^2_{Y_i|\mathscr{C}}} \;=\; \sqrt{Var(Y_i)\cdot\big(1-R^2_{Y_i|\mathscr{C}}\big)} && \text{[(11.9)]}\\
&= SD(Y_i)\cdot\sqrt{1-R^2_{Y_i|\mathscr{C}}},
\end{aligned}$$

which implies

$$SD(\varepsilon_1)\cdot SD(\varepsilon_2) \;=\; SD(Y_1)\cdot SD(Y_2)\cdot\sqrt{1-R^2_{Y_1|\mathscr{C}}}\cdot\sqrt{1-R^2_{Y_2|\mathscr{C}}}.$$

Using these results, Definition (11.49) yields

$$\begin{aligned}
Corr\,(Y_1,Y_2\,;\mathscr{C}) \;:=\; Corr\,(\varepsilon_1,\varepsilon_2) \;&=\; \frac{Cov\,(\varepsilon_1,\varepsilon_2)}{SD(\varepsilon_1)\cdot SD(\varepsilon_2)}\\
&= \frac{SD(Y_1)\cdot SD(Y_2)\cdot\big[Corr\,(Y_1,Y_2) - R_{Y_1|\mathscr{C}} R_{Y_2|\mathscr{C}}\cdot Corr\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})]\big]}{SD(Y_1)\cdot SD(Y_2)\cdot\sqrt{1-R^2_{Y_1|\mathscr{C}}}\cdot\sqrt{1-R^2_{Y_2|\mathscr{C}}}}\\
&= \frac{Corr\,(Y_1,Y_2) - R_{Y_1|\mathscr{C}}\cdot R_{Y_2|\mathscr{C}}\cdot Corr\,[E(Y_1|\mathscr{C}), E(Y_2|\mathscr{C})]}{\sqrt{1-R^2_{Y_1|\mathscr{C}}}\;\cdot\;\sqrt{1-R^2_{Y_2|\mathscr{C}}}},
\end{aligned}$$

which is Equation (11.51).

▷ **Solution 11-8** We have already proven Equation (11.51) and we assume that, for $i = 1,2$, the conditional expectations $E(Y_i \mid X)$ have linear parametrizations with nonzero slopes and that $Corr(Y_i, X)^2 < 1$. Hence, there are $\beta_{i0}, \beta_{i1} \in \mathbb{R}$, $i = 1,2$, such that

$$E(Y_1 \mid X) \underset{P}{=} \beta_{10} + \beta_{11} X \qquad \text{and} \qquad E(Y_2 \mid X) \underset{P}{=} \beta_{20} + \beta_{21} X, \qquad \beta_{11}, \beta_{21} \neq 0.$$

Therefore,

$$E(Y_1 \mid X) \underset{P}{=} a_0 + a_1 E(Y_2 \mid X), \quad \text{with} \quad a_0 = \beta_{10} - \beta_{20} \cdot \frac{\beta_{11}}{\beta_{21}} \quad \text{and} \quad a_1 = \frac{\beta_{11}}{\beta_{21}}.$$

Furthermore, $\beta_{i1} \neq 0$, $Corr(Y_i, X)^2 < 1$, $i = 1,2$, and Equations (7.17) and (11.16) imply $0 < Corr(Y_i, X)^2 < 1$. Now we consider two cases.

Case 1: The slopes β_{11} and β_{21} have *identical signs*. Then $a_1 > 0$ and the correlations $Corr(Y_1, X)$, $Corr(Y_2, X)$ have identical signs as well [see (11.16)], and

$$R_{Y_1|X} \cdot R_{Y_2|X} = Corr(Y_1, X) \cdot Corr(Y_2, X)$$

[see Eq. (11.23)] and $Corr[E(Y_1 \mid X), E(Y_2 \mid X)] = 1$ [see Cor. 7.22]. This implies that Equation (11.51) simplifies to

$$Corr(Y_1, Y_2 ; X) = \frac{Corr(Y_1, Y_2) - Corr(Y_1, X) \cdot Corr(Y_2, X)}{\sqrt{1 - Corr(Y_1, X)^2} \cdot \sqrt{1 - Corr(Y_2, X)^2}}.$$

Case 2: The slopes β_{11} and β_{21} have *different signs*. Then $a_1 < 0$ and the correlations $Corr(Y_1, X)$, $Corr(Y_2, X)$ have different signs as well [see (11.16)]. In this case, the same equation holds, because $R_{Y_1|X} \cdot R_{Y_2|X} = -Corr(Y_1, X) \cdot Corr(Y_2, X)$ [see Eq. (11.23)] and, according to Corollary 7.22, $Corr[E(Y_1 \mid X), E(Y_2 \mid X)] = -1$.

Chapter 12
Linear Regression

In chapter 10, we introduced the general concepts of a conditional expectation and a regression, and in chapter 11 we treated the *residual* with respect to a conditional expectation, the concepts of a *conditional variance*, a *conditional covariance*, and a *partial correlation*. Now we turn to *parametrizations* of a conditional expectation. A parametrization serves to describe a conditional expectation with a few parameters (real numbers). Oftentimes these parameters have important substantive meanings. We treat a *linear parametrization* of a conditional expectation, which is also called the *linear regression*, if it is uniquely defined. We start with the basic ideas, present the definitions, treat some examples, consider the relationship between a linear regression and a linear quasi-regression, and deal with uniqueness of a linear parametrization and the identification of the regression coefficients. Finally, we present a theorem on the invariance of regression coefficients and a theorem on the existence of a linear regression if the regressand and the regressors have a joint multivariate normal distribution.

12.1 Basic Ideas

Consider the random variables $X\colon (\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ and $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$, and let Y be nonnegative or with finite expectation. Furthermore, let $\mathscr{C}\subset\mathscr{A}$ be a σ-algebra and assume that X is $\mathscr{C}$-measurable. Now assume that there is a real-valued version $E(Y|\mathscr{C})$ of the $\mathscr{C}$-conditional expectation of Y and coefficients $\beta_0,\beta_1\in\mathbb{R}$ such that

$$E(Y|\mathscr{C}) = \beta_0+\beta_1 X. \tag{12.1}$$

Then we call the function $g\colon\mathbb{R}\to\mathbb{R}$ defined by

$$g(x) := \beta_0+\beta_1 x, \quad \forall\, x\in\mathbb{R}, \tag{12.2}$$

a *linear parametrization of* $E(Y|\mathscr{C})$. This definition implies $g(X)\in\mathscr{E}(Y|\mathscr{C})$, where $g(X)$ denotes the composition of X and g.

If $\mathscr{C}=\sigma(X)$, then a function $g\colon\mathbb{R}\to\overline{\mathbb{R}}$ satisfying $E(Y|\mathscr{C})=E(Y|X)=g(X)$ always exists (see Rem. 10.3 and Cor. 10.23). However, g is not necessarily a linear function. Hence, even if $\mathscr{C}=\sigma(X)$, a *linear parametrization* of $E(Y|\mathscr{C})=E(Y|X)$ does *not necessarily exist.* Yet, if we assume that

(a) Equations (12.1) and (12.2) hold, and
(b) the variance of X is positive and finite,

then $E(Y|X)$ and g, and therefore, the coefficients β_0 and β_1, are uniquely defined. Under the assumptions (a) and (b), the function g is also called the *linear regression* of Y on X and the numbers β_0 and β_1 are called *regression coefficients*.

Remark 12.1 (Composition of X and a Linear Function) Because the conditional expectation $E(Y|\mathscr{C})$ is a function with domain Ω, strictly speaking, it is not a linear function itself. Assuming that Equation (12.1) holds and saying that $E(Y|\mathscr{C})$ is a linear function of X we mean that $E(Y|\mathscr{C})$ *is the composition* of the random variable $X\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ and the linear function $g\colon \mathbb{R}\to\mathbb{R}$ satisfying (12.2). This is why g and not $E(Y|\mathscr{C})$ itself is called a linear parametrization and a linear regression if the assumptions (a) and (b) hold. ◁

Remark 12.2 (Estimation) Although estimation is beyond the scope of this book, it is worthwhile noting that estimation is one of the reasons why a parametrization is useful. The definition of a concrete version of a conditional expectation $E(Y|X)$ requires that we know for all $\omega\in\Omega$ which values $E(Y|X)(\omega)$ are assigned to ω. In empirical applications, these values are often *unknown*, i. e., we do not know which concrete number $E(Y|X)(\omega)$ is assigned to a concrete ω. In these cases estimating the values of the conditional expectation may be an issue. In particular, if Equations (12.1) and (12.2) hold and the variance of X is positive and finite, then the coefficients β_0 and β_1 — and with them $E(Y|X)=\beta_0+\beta_1 X$ — can be computed from estimable quantities such as the variance of X and the expectations and the covariance of X and Y. In this case, estimation of the values of $E(Y|X)$ is relatively simple because the variance of X and the expectations and the covariance of X and Y can be estimated in a data sample. ◁

Example 12.3 (Joe and Ann With Self-Selection – continued) Table 12.1 shows nine random variables, the first five of which may be called *observable* (or *manifest*), whereas the last four are *unobservable* (or *latent*). The difference between the two kinds of random variables is that, in empirical applications, the values of the conditional expectations, the unobservable random variables, are unknown parameters that we might wish to estimate in a sample. These parameters can be computed from the joint distributions of the random variables involved. In this fictitious example, the information about the joint distribution of the random variables U, X, and Y is contained in the second column of the table, whereas in empirical applications these parameters usually have to be estimated using a data sample. Examples in case are the conditional expectation values $E(Y|U{=}\mathit{Joe},X{=}0)=.70$ and $E(Y|X{=}0)=.60$. In contrast to the values of the conditional expectations, the values of the five observables are known for all eight possible outcomes $\omega\in\Omega$ of the random experiment. For example, if $\omega=(\mathit{Joe},\mathit{no},-)$, then $U(\omega)=\mathit{Joe}$, $X(\omega)=0$, and $Y(\omega)=0$, and these values are known, because the definitions of these observables do not involve unknown parameters that depend on the joint distribution of the random variables involved.

In this example, we may consider, for instance, the conditional expectations

Table 12.1. Joe and Ann With Self-Selection: Conditional Expectations

Outcomes ω				Observables					Conditional Expectations			
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Indicator for Joe $1_{U=Joe}$	Indicator for Ann $1_{U=Ann}$	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$E(Y\|U)$	$P(X=1\|U)$
(*Joe*,	*no*,	−)	.144	*Joe*	1	0	0	0	.70	.60	.704	.04
(*Joe*,	*no*,	+)	.336	*Joe*	1	0	0	1	.70	.60	.704	.04
(*Joe*,	*yes*,	−)	.004	*Joe*	1	0	1	0	.80	.42	.704	.04
(*Joe*,	*yes*,	+)	.016	*Joe*	1	0	1	1	.80	.42	.704	.04
(*Ann*,	*no*,	−)	.096	*Ann*	0	1	0	0	.20	.60	.352	.76
(*Ann*,	*no*,	+)	.024	*Ann*	0	1	0	1	.20	.60	.352	.76
(*Ann*,	*yes*,	−)	.228	*Ann*	0	1	1	0	.40	.42	.352	.76
(*Ann*,	*yes*,	+)	.152	*Ann*	0	1	1	1	.40	.42	.352	.76

Note. The probabilities of the elementary events are fictive

$$E(Y|X) = .60 - .18 \cdot X, \tag{12.3}$$

$$E(Y|U) = .352 + \cdot .352 \cdot 1_{U=Joe}, \tag{12.4}$$

and

$$E(Y|X,U) = .20 + .20 \cdot X + .50 \cdot 1_{U=Joe} - .10 \cdot X \cdot 1_{U=Joe}. \tag{12.5}$$

The computation of the parameters of these equations is illustrated in Examples 12.16 and 12.23. ◁

12.2 Assumptions and Definitions

In this section we often refer to the following assumptions and the following notation.

Notation and Assumptions 12.4
$Y\colon (\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ and $X:=(X_1,\ldots,X_n)\colon(\Omega,\mathscr{A},P)\to(\mathbb{R}^n,\mathscr{B}_n)$ are random variables, where Y is nonnegative or has a finite expectation $E(Y)$. Furthermore, $\mathscr{C}\subset\mathscr{A}$ is a σ-algebra and X is $\mathscr{C}$-measurable.

Referring to these assumptions we define a linear parametrization as follows:

Definition 12.5 (Linear Parametrization)
Let the assumptions 12.4 hold and let $\beta_0, \beta_1, \ldots, \beta_n \in \mathbb{R}$. *If there is a real-valued version* $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ *such that*

$$E(Y|\mathscr{C}) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i, \tag{12.6}$$

then the function $g\colon \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$g(x) := \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{12.7}$$

is called a $\mathit{linear\ parametrization\ of}$ $E(Y|\mathscr{C})$.

Note that, even if it exists, then a linear parametrization of $E(Y|\mathscr{C})$ is not uniquely defined unless additional assumptions hold (see Example 12.9). Uniqueness of a linear parametrization is treated in Corollary 12.30.

Remark 12.6 (Another Notation) If $\sigma(X) = \mathscr{C}$, then Equation (12.6) is equivalent to

$$E(Y|X) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i. \tag{12.8}$$

◁

Remark 12.7 (X-Conditional Mean Independence) Equation (12.6) implies $\sigma(X) \subset \mathscr{C}$ and that Y is X-conditionally mean independent from $\mathscr{C}$ (see Def. 10.44). If $\sigma(X) \neq \mathscr{C}$, then this conditional mean independence is nontrivial (see Rem. 10.46).
◁

Remark 12.8 (Other Versions and Other Factorizations) Note that a linear parametrization g of the conditional expectation $E(Y|\mathscr{C})$ is a factorization of $E(Y|X)$, i. e., $E(Y|X) = g(X)$ is the composition of X and g (see section 10.4). Also note that there may be other factorizations g^* of $E(Y|X)$ and versions in $V^* \in \mathscr{E}(Y|\mathscr{C})$ that do not satisfy Equations (12.6) and (12.7), respectively. However, according to Theorem 10.9 (ii), Equation (12.6) implies

$$V^* \underset{P}{=} \beta_0 + \sum_{i=1}^{n} \beta_i X_i, \quad \forall\, V^* \in \mathscr{E}(Y|\mathscr{C}). \tag{12.9}$$

Furthermore, if g, g^* are factorizations of versions $V, V^* \in \mathscr{E}(Y|X)$, respectively, then

$$g^* \underset{P_X}{=} g \tag{12.10}$$

(see Cor. 10.29). ◁

Example 12.9 (Constant Regressor) Suppose that $E(Y|\mathscr{C}) = g(X) = \beta_0 + \beta_1 X$ and X is a constant function such that, for all $\omega \in \Omega$, $X(\omega) = \alpha$, $\alpha \in \mathbb{R}$. Then $Var(X) = 0$ and $g, g^*: \mathbb{R} \to \mathbb{R}$ defined by

$$\forall x \in \mathbb{R}: \; g(x) = \beta_0 + \beta_1 x \qquad \text{and} \qquad g^*(x) = (\beta_0 - \gamma\alpha) + (\beta_1 + \gamma)x, \quad 0 \neq \gamma \in \mathbb{R},$$

are two linear parametrizations that differ from each other but satisfy $g(X) = g^*(X) = E(Y|X)$. This example shows that a linear parametrization g is not uniquely defined unless additional assumptions hold. ◁

Remark 12.10 (Other Linear Parametrizations) Remember that $E(Y|X)$ is just a simplified notation for $E(Y|\mathscr{C})$ with $\mathscr{C} = \sigma(X)$. Hence, if $Z = (Z_1, \ldots, Z_m): \Omega \to \mathbb{R}^m$ is an m-variate random variable on $(\Omega, \mathscr{A}, P)$, and $\sigma(Z) = \mathscr{C}$, then for one and the same version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ there may be another parametrization $f: \mathbb{R}^m \to \mathbb{R}$ with coefficients $\gamma_0, \gamma_1, \ldots, \gamma_m \in \mathbb{R}$ satisfying

$$f(z) = \gamma_0 + \sum_{i=1}^{m} \gamma_i z_i, \quad \forall z = (z_1, \ldots, z_m) \in \mathbb{R}^m. \tag{12.11}$$

In other words, one and the same version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ may have several linear parametrizations such as f and g with different regression coefficients γ_0, $\gamma_1, \ldots, \gamma_m \in \mathbb{R}$ and $\beta_0, \beta_1, \ldots, \beta_n \in \mathbb{R}$ pertaining to different random variables $Z = (Z_1, \ldots, Z_m)$ and $X = (X_1, \ldots, X_n)$, respectively. ◁

Remark 12.11 (Conditional Expectation Values) If g is a linear parametrization of $E(Y|X)$ satisfying Equation (12.7), then, according to Definition 10.33,

$$\begin{aligned} E(Y|X{=}x) &= E(Y|X_1{=}x_1, \ldots, X_n{=}x_n) \\ &= g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n. \end{aligned} \tag{12.12}$$

Note that another factorization g^* of $E(Y|X)$ might yield other conditional expectation values $E(Y|X{=}x)$. ◁

Remark 12.12 (P_X-Equivalence of Different Parametrizations) If g, g^* are factorizations of versions $V, V^* \in \mathscr{E}(Y|X)$, then, according to Equation (12.10),

$$g(x) = g^*(x), \quad \text{for } P_X\text{-almost all } x \in \mathbb{R}^n, \tag{12.13}$$

[see Eq. (10.26)]. ◁

12.3 Examples

Example 12.13 (Univariate Real-Valued X) If $X: (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ and

$$E(Y|X) = \beta_0 + \beta_1 X, \tag{12.14}$$

then the function $g: \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = \beta_0 + \beta_1 x, \quad x \in \mathbb{R}, \tag{12.15}$$

is a linear parametrization of $E(Y|X)$. If $0 < Var(X) < \infty$, then g is uniquely defined and also called the *simple linear regression* of Y on X (see Cor. 12.30 and Rem. 12.33). ◁

Example 12.14 (Intercept and Slope) If $E(Y|X) = \beta_0 + \beta_1 X$, then

$$\beta_0 = E(Y|X=0). \tag{12.16}$$

Furthermore, if $x_1, x_2 \in \mathbb{R}$ and $x_2 > x_1$, then

$$\beta_1 = \frac{1}{x_2 - x_1} \cdot \big[E(Y|X=x_2) - E(Y|X=x_1)\big] \tag{12.17}$$

[see Exercise 7-4 and Eq. (12.12)]. Equation (12.17) yields

$$\beta_1 = E(Y|X=x_2) - E(Y|X=x_1), \quad \text{if } x_2 - x_1 = 1. \tag{12.18}$$

This justifies calling β_0 the *intercept* and β_1 the *slope* of $E(Y|X)$, respectively (see Fig. 7.3). Note that these equations also apply if $P(X=0) = P(X=x_1) = P(X=x_2) = 0$. They even apply if $0, x_1, x_2 \notin X(\Omega)$. ◁

Example 12.15 (Dichotomous Regressor) If X is dichotomous with values 0 and 1 (see Def. 5.10), then there is always a version $E(Y|X) \in \mathscr{E}(Y|X)$ such that

$$E(Y|X) = \beta_0 + \beta_1 X \tag{12.19}$$

with

$$\beta_0 = E(Y|X=0), \tag{12.20}$$

and

$$\beta_1 = E(Y|X=1) - E(Y|X=0) \tag{12.21}$$

(for a proof see Th. 12.36). ◁

Example 12.16 (Joe and Ann With Self-Selection – continued) In the example of Table 12.1,

$$E(Y|X) = .60 - .18 \cdot X, \tag{12.22}$$

and the function $g\colon \mathbb{R} \to \mathbb{R}$ defined by $g(x) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$, is a linear parametrization of $E(Y|X)$.

The intercept is most easily obtained via Equation (12.20):

$$\begin{aligned} \beta_0 &= E(Y|X=0) = P(Y=1|X=0) \\ &= \frac{P(Y=1, X=0)}{P(X=0)} = \frac{.336 + .024}{.144 + .336 + .096 + .024} = .60. \end{aligned} \tag{12.23}$$

The slope is obtained via Equation (12.21):

$$\begin{aligned}\beta_1 &= E(Y|X=1) - E(Y|X=0) = P(Y=1|X=1) - P(Y=1|X=0)\\ &= \frac{P(Y=1,X=1)}{P(X=1)} - P(Y=1|X=0)\\ &= \frac{.016+.152}{.004+.016+.228+.152} - .60 = .42 - .60 = -.18.\end{aligned} \tag{12.24}$$

◁

Example 12.17 (Dichotomous Regressor – continued) Continue Example 12.15 and define the random variable $Z\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ by $Z := 2X-1$. Then Z is dichotomous with values -1 and 1, and

$$\forall\,\omega\in\Omega\colon\ X(\omega)=0 \Leftrightarrow Z(\omega)=-1 \quad \text{and} \quad X(\omega)=1 \Leftrightarrow Z(\omega)=1.$$

Note that $\sigma(Z)=\sigma(X)$ holds for the σ-algebras generated by X and Z. Because $X=\frac{1}{2}(Z+1)$,

$$E(Y|X) = \beta_0+\beta_1 X = \beta_0+\frac{\beta_1}{2}\cdot(Z+1) = \beta_0+\frac{\beta_1}{2}+\frac{\beta_1}{2}\cdot Z = E(Y|Z). \tag{12.25}$$

the function $g^*\colon \mathbb{R}\to\mathbb{R}$ defined by $g^*(z)=\alpha_0+\alpha_1 z,\ z\in\mathbb{R}$, is a linear parametrization of $E(Y|X)$, where

$$\begin{aligned}\alpha_0 = \beta_0 + \frac{\beta_1}{2} &= E(Y|X=0) + \frac{E(Y|X=1)-E(Y|X=0)}{2}\\ &= \frac{E(Y|X=1)+E(Y|X=0)}{2}\end{aligned} \tag{12.26}$$

and

$$\alpha_1 = \frac{\beta_1}{2} = \frac{E(Y|X=1)-E(Y|X=0)}{2}. \tag{12.27}$$

Note that

$$\begin{aligned}\{X=1\} &= \left\{\frac{1}{2}(Z+1)=1\right\} = \{Z=1\},\\ \{X=0\} &= \left\{\frac{1}{2}(Z+1)=0\right\} = \{Z=-1\}.\end{aligned} \tag{12.28}$$

Because X and Z are dichotomous with positive probabilities for both of their values, Equation (12.28) and $E(Y|X=x)=E(Y|\{X=x\})$ [see Eq. (9.6)] imply

$$\alpha_0 = \frac{E(Y|Z=1)+E(Y|Z=-1)}{2}, \tag{12.29}$$

$$\alpha_1 = \frac{E(Y|Z=1)-E(Y|Z=-1)}{2}. \tag{12.30}$$

Comparing Equations (12.26) and (12.27) to Equations (12.20) and (12.21) shows that the meaning of the regression coefficients depends on the choice of the particular parametrization. ◁

Example 12.18 (Joe and Ann With Self-Selection–continued) Using the results of Example 12.16 as well as Equations (12.25) to (12.27) yields

$$E(Y|X) \;=\; E(Y|Z) \;=\; \frac{.42+.60}{2} \;+\; \frac{.42-.60}{2} \;=\; .51 \;-\; .09\cdot Z. \tag{12.31}$$

Note again that $E(Y|X)$ and $E(Y|Z)$ are only different notations for a version of $\mathscr{E}(Y|\mathscr{C})$, where $\mathscr{C} = \sigma(X) = \sigma(Z)$ and that $g\colon \mathbb{R} \to \mathbb{R}$ defined by $g(x) = .60 - .18\cdot x$, $x \in \mathbb{R}$, and $g^*\colon \mathbb{R} \to \mathbb{R}$ defined by $g^*(z) = .51 - .09\cdot z,\ z \in \mathbb{R}$, are two different linear parametrizations of one and the same version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$. ◁

Example 12.19 (Quadratic Function) Let $X_1\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a real-valued random variable, let $X_2 := X_1^2$, $X := (X_1, X_2)$, and assume that there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ with

$$E(Y|X) \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \tag{12.32}$$

Then the function $g\colon \mathbb{R}^2 \to \mathbb{R}$ defined by

$$g(x) \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{12.33}$$

is a linear parametrization of $E(Y|X)$. ◁

Example 12.20 (Logarithmic Function) Consider $Z\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, a real-valued and positive random variable, define $X := \ln Z$, and assume that there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ with

$$E(Y|X) \;=\; \beta_0 + \beta_1 \ln Z \;=\; \beta_0 + \beta_1 X. \tag{12.34}$$

Then the function $g\colon \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = \beta_0 + \beta_1 x, \quad x \in \mathbb{R}, \tag{12.35}$$

is a linear parametrization of $E(Y|X)$. Note that the definition of X implies that there is also a version $E(Y|Z) \in \mathscr{E}(Y|Z)$ with $E(Y|Z) = E(Y|X)$, because $\sigma(X) = \sigma(Z)$ (see Exercise 12-1). ◁

Example 12.21 (Two Regressors) If $X_i\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, $i = 1,2$, are univariate real-valued random variables, $X := (X_1, X_2)$, and there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ with

$$E(Y|X) \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{12.36}$$

then the function $g\colon \mathbb{R}^2 \to \mathbb{R}$ defined by

$$g(x) \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{12.37}$$

is a linear parametrization of $E(Y|X)$. ◁

Example 12.22 (Two Regressors and Their Product) Let $X_i : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, $i = 1,2$, be univariate real-valued random variables, define $X_3 := X_1 \cdot X_2$ and $X := (X_1, X_2, X_3)$, and assume that there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ with

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2. \tag{12.38}$$

Then the function $g : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \quad \forall x = (x_1, x_2, x_3) \in \mathbb{R}^3, \tag{12.39}$$

is a linear parametrization of $E(Y|X)$. ◁

Example 12.23 (Joe and Ann With Self-Selection – continued) Consider the random variables X and $1_{U=Joe}$ specified in Table 12.1 and define $Z := (Z_1, Z_2, Z_3) := (X, 1_{U=Joe}, X \cdot 1_{U=Joe})$. Then

$$E(Y|Z) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot 1_{U=Joe} + \beta_3 \cdot X \cdot 1_{U=Joe}, \tag{12.40}$$

and the function $g : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$g(z) = .20 + .20 \cdot z_1 + .50 \cdot z_2 - .10 \cdot z_3, \quad \forall z = (z_1, z_2, z_3) \in \mathbb{R}^3,$$

is a linear parametrization of $E(Y|Z)$.

The coefficients β_0 to β_3 in Equation (12.40) are most easily obtained as follows. Table 12.1 and Equation (12.40) yield:

$$\begin{aligned}
\beta_0 &= E(Y|Z_1{=}0, Z_2{=}0, Z_3{=}0) = P(Y{=}1|X{=}0, U{=}Ann) = .20,\\
\beta_0 + \beta_1 &= E(Y|Z_1{=}1, Z_2{=}0, Z_3{=}0) = P(Y{=}1|X{=}1, U{=}Ann) = .40,\\
\beta_0 + \beta_2 &= E(Y|Z_1{=}0, Z_2{=}1, Z_3{=}0) = P(Y{=}1|X{=}0, U{=}Joe) = .70,\\
\beta_0 + \beta_1 + \beta_2 + \beta_3 &= E(Y|Z_1{=}1, Z_2{=}1, Z_3{=}1) = P(Y{=}1|X{=}1, U{=}Joe) = .80.
\end{aligned}$$

Solving these equations for the four coefficients and inserting them into Equation (12.40) yields

$$E(Y|Z) = .20 + .20 \cdot X + .50 \cdot 1_{U=Joe} - .10 \cdot X \cdot 1_{U=Joe}. \tag{12.41}$$

In this example, $\sigma(Z) = \sigma(X, U)$ and $Z(\Omega) = \{0,1\}^3$. According to Remark 10.16, this implies $E(Y|Z) = E(Y|X,U)$ and that the function g is also a linear parametrization of $E(Y|X,U)$. ◁

Remark 12.24 (Generalizing the Examples) Generalizing the Examples 12.19 to 12.23, let $Z : (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ be a (univariate or multivariate) random variable. For $i = 1, \ldots, n$, let $h_i : (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ be measurable functions, define $X_i := h_i(Z)$, and $X := (X_1, \ldots, X_n)$. If there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ such that

$$E(Y|X) = \beta_0 + \sum_{i=1}^{n} \beta_i h_i(Z), \tag{12.42}$$

then $g\colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$g(x) \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i\, x_i, \quad \forall\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{12.43}$$

is a linear parametrization of $E(Y|X)$. Remember that $\sigma(X) \subset \sigma(Z)$, but note that $\sigma(X) = \sigma(Z)$ does not necessarily hold. If $\sigma(X) \neq \sigma(Z)$, then there is not necessarily a version $E(Y|Z) \in \mathscr{E}(Y|Z)$ with $E(Y|Z) = E(Y|X)$. However, if we assume that there is a version $E(Y|Z) \in \mathscr{E}(Y|Z)$ with $E(Y|Z) = E(Y|X)$, then

$$E(Y|Z) \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i\, h_i(Z). \tag{12.44}$$

In this case, the function g is also a linear parametrization of $E(Y|Z)$. ◁

12.4 Linear Quasi-Regression

In the following corollary $Q_{lin}(Y|X_1, \ldots, X_n)$ denotes the function that has been introduced in Definition 7.26. This corollary immediately follows from Theorem 10.26 and Definition 12.5.

Corollary 12.25 (Linear Regression and Linear Quasi-Regression)
Let the assumptions 12.4 hold and suppose that $E(Y^2), E(X_i^2) < \infty$, $i = 1, \ldots, n$. If there is a version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ with $E(Y|\mathscr{C}) = \beta_0 + \sum_{i=1}^{n} \beta_i\, X_i$, where $\beta_0, \beta_1, \ldots, \beta_n \in \mathbb{R}$, then

$$E(Y|\mathscr{C}) \;=\; Q_{lin}(Y|X_1, \ldots, X_n) \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i\, X_i. \tag{12.45}$$

Hence, if $E(Y|\mathscr{C}) = \beta_0 + \sum_{i=1}^{n} \beta_i\, X_i$, then $E(Y|\mathscr{C})$ and $Q_{lin}(Y|X_1, \ldots, X_n) = f(X)$, then the composition of X and the linear quasi-regression $f\colon \mathbb{R}^n \to \mathbb{R}$ (see Def. 7.26), are identical.

Remark 12.26 (Dichotomous Regressor) If X is dichotomous with values 0 and 1 (see Def. 5.10), then $E(Y|X) = Q_{lin}(Y|X)$. If $P(X{=}x_i) > 0$ for at least three different $x_i \in \mathbb{R}$, then it is not necessarily true that $Q_{lin}(Y|X) \in \mathscr{E}(Y|X)$. If $Q_{lin}(Y|X) \notin \mathscr{E}(Y|X)$, then there are no $\beta_0, \beta_1 \in \mathbb{R}$ such that the function g defined by $g(x) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$, is a linear parametrization of a version $E(Y|X) \in \mathscr{E}(Y|X)$ (see Example 12.40). ◁

Remark 12.27 (Unbounded Regressor, Dichotomous Regressand) Suppose that X and Y are real-valued random variables on $(\Omega, \mathscr{A}, P)$, and that Y is dichotomous with values 0 and 1. Because $P(0 \leq Y \leq 1) = 1$, Rules (iii) and (iv) of Box 10.3 imply

$$0 \underset{P}{\leq} E(Y|X) = P(Y{=}1|X) \underset{P}{\leq} 1. \tag{12.46}$$

Suppose that the regressor X is *not P-almost surely bounded*, i. e., suppose

$$\forall c \in \mathbb{R}, c > 0: \; P(X < -c) + P(X > c) > 0. \tag{12.47}$$

Then there is no linear parametrization of $E(Y|X)$ with slope $\beta_1 \neq 0$ (see Exercise 12-2). Note that the premise (12.47) holds, e. g., if X has a normal distribution. ◁

Example 12.28 (No Treatment for Joe – continued) In the example presented in Table 9.2 (p. 282),

$$E(Y|X) = \beta_0 + \beta_1 X = .6 - .2 \cdot X \tag{12.48}$$

(see Example 12.15) and

$$g(x) = \beta_0 + \beta_1 \cdot x = .6 - .2 \cdot x, \quad \forall x \in \mathbb{R}, \tag{12.49}$$

defines a linear parametrization $g: \mathbb{R} \to \mathbb{R}$ of $E(Y|X)$. Hence, according to Remark 12.11, we may define

$$E(Y|X{=}x) = g(x) = .6 - .2 \cdot x, \quad \forall x \in \mathbb{R}, \tag{12.50}$$

(see Def. 10.33). For $x = 0$, Equation (12.50) yields $E(Y|X{=}0) = .6$, and it yields $E(Y|X{=}1) = .4$ for $x{=}1$. Note that the definition of the conditional expectation values $E(Y|X{=}x)$ for $x \in \mathbb{R} \setminus \{0,1\}$ via Equation (12.50) is arbitrary, because $P_X(X \in \mathbb{R} \setminus \{0,1\}) = 0$. Using any other factorization of $E(Y|X)$ for the definition of the conditional expectation values $E(Y|X{=}x)$ for $x \in \mathbb{R} \setminus \{0,1\}$ would do as well.

Remark 12.26 and Definition 7.26 imply that the function $MSE: \mathbb{R}^2 \to [0, \infty[$ defined by

$$MSE(a_0, a_1) = E\big([Y - (a_0 + a_1 X)]^2\big), \quad \forall (a_0, a_1) \in \mathbb{R}^2, \tag{12.51}$$

has its minimum for $(a_0, a_1) = (.6, -.2)$. Hence, in this example,

$$E(Y|X) = Q_{lin}(Y|X) = .6 - .2 \cdot X,$$

and the function $g: \mathbb{R} \to \mathbb{R}$ defined by Equation (12.49) is a linear parametrization of $E(Y|X)$. ◁

12.5 Uniqueness and Identification of Regression Coefficients

In Example 12.9 we showed that a linear parametrization is not uniquely defined unless additional assumptions hold. Such assumptions are specified in Corollary 12.30, which is uses the following notation and general assumptions:

Notation and Assumptions 12.29
Let the assumptions 12.4 hold. Furthermore, $\boldsymbol{x} := [X_1, \ldots, X_n]'$ is the column vector of $X = (X_1, \ldots, X_n)$, $\boldsymbol{\mu} := [E(X_1), \ldots, E(X_n)]'$ the column vector of the expectations of $X_1, \ldots, X_n$, and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]'$ a column vector of n real numbers.

Assuming finite second moments of Y and $X_1, \ldots, X_n$,

$$\boldsymbol{\Sigma}_{\boldsymbol{xx}} := E\big[[\boldsymbol{x}-\boldsymbol{\mu}]\,[\boldsymbol{x}-\boldsymbol{\mu}]'\big] = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_2X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_nX_1} & \sigma_{X_nX_2} & \cdots & \sigma_{X_n}^2 \end{bmatrix}$$

denotes the covariance matrix of $X = (X_1, \ldots, X_n)$ (see section 7.4.1). Furthermore,

$$\boldsymbol{\Sigma}_{\boldsymbol{yx}} := \big[\sigma_{YX_1}, \ldots, \sigma_{YX_n}\big]$$

denotes the row vector of the covariances $Cov(Y, X_i) = \sigma_{YX_i}$, $i = 1, \ldots, n$, and $\boldsymbol{\Sigma}_{\boldsymbol{xy}} := \boldsymbol{\Sigma}_{\boldsymbol{yx}}'$ the column vector of these covariances. Remember that the notation $X = (X_1, \ldots, X_n)$ refers to an *n-variate random variable*, whereas $\boldsymbol{x} = [X_1, \ldots, X_n]$ denotes the *row vector* of the random variables $X_1, \ldots, X_n$.

The following corollary immediately follows from Theorem 7.28 and Corollary 12.25. It shows how to compute (identify) the regression coefficients of a linear parametrization of $E(Y|\mathscr{C})$ and it specifies the conditions under which a linear parametrization of $E(Y|\mathscr{C})$ is uniquely defined.

Corollary 12.30 (Identification of Parameters)
Let the assumptions 12.29 hold. If there is a version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ with

$$E(Y|\mathscr{C}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i, \tag{12.52}$$

then

$$\beta_0 = E(Y) - \boldsymbol{\beta}'\boldsymbol{\mu}. \tag{12.53}$$

If, in addition, Y and $X_1, \ldots, X_n$ have finite second moments and the inverse $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ exists, then

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\,\boldsymbol{\Sigma}_{\boldsymbol{xy}}, \tag{12.54}$$

and the linear parametrization $g\colon \mathbb{R}^n \to \mathbb{R}$ of $E(Y|\mathscr{C}) = g(X)$ with

$$g(x) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{12.55}$$

is uniquely defined.

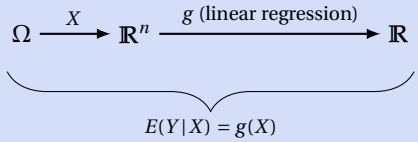


Figure 12.1. $E(Y|X)$ as the composition of X and the linear regression g

12.6 Linear Regression

The uniqueness property formulated in Corollary 12.30 allows us to define a linear regression as follows:

Definition 12.31 (Linear Regression)
Let the assumptions 12.29 hold and suppose that there is a version $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ such that Equation (12.52) holds. Furthermore, assume that Y and $X_1,\dots,X_n$ have finite second moments and that the inverse $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ exists. Then the function $g\colon \mathbb{R}^n \to \mathbb{R}$ defined by Equation (12.55) is called the linear regression of Y on X (with respect to P).

Figure 12.1 shows the conditional expectation $E(Y|X)$ as the composition of X and the linear regression g. Hence, while the conditional expectation $E(Y|X)$ is a function with domain Ω and codomain $\mathbb{R}$, the linear regression g is a function with domain $\mathbb{R}^n$ and codomain $\mathbb{R}$.

Remark 12.32 (Simple and Multiple Linear Regression) If $n \geq 2$, then the linear regression of Y on X is also called the *multiple linear regression* of Y on X. The coefficients $\beta_0, \beta_1, \dots, \beta_n$ are called *regression coefficients*, and β_0 the *intercept*. If $n = 1$, then a linear regression of Y on X is also called a *simple linear regression* of Y on X with *slope* β_1. ◁

Remark 12.33 (Simple Regression as a Special Case) If $n = 1$ and we define $X := X_1$, then Equation (12.52) can be written

$$E(Y|\mathscr{C}) = \beta_0 + \beta_1 X. \tag{12.56}$$

If Equation (12.56) holds, then there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ such that

$$E(Y|\mathscr{C}) = E(Y|X) = \beta_0 + \beta_1 X = Q_{lin}(Y|X). \tag{12.57}$$

Therefore, we obtain the same results for the regression coefficients that have already been described in Theorem 7.13. In particular, Equation (12.53) yields

$$\beta_0 = E(Y) - \beta_1 E(X), \tag{12.58}$$

and (12.54) implies

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}. \tag{12.59}$$

If $n = 1$ and $X = X_1$, then $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1} = \left[\frac{1}{Var(X)}\right]$ and the existence of the inverse $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ is equivalent to $Var(X) > 0$ (see Th. 7.13 and Rem. 7.24). ◁

In the following theorem, we consider the implications of assuming that $Z = (Y, X_1, \ldots, X_n)$ has an $(n+1)$-variate normal distribution (see Def. 8.50).

Theorem 12.34 (Linear Parametrization and Normal Distribution)
Let the assumptions 12.29 hold and let $Z = (Y, X_1, \ldots, X_n)$ be an $(n+1)$-variate real-valued random variable on $(\Omega, \mathcal{A}, P)$ with $Z \sim \mathcal{N}_{\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_{zz}}$ Furthermore, assume that the inverse $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ exists. Then there is a version $E(Y|X) \in \mathcal{E}(Y|X)$ with

$$E(Y|X) = \beta_0 + \boldsymbol{\beta}' \boldsymbol{x}$$

such that Equations (12.53) *and* (12.54) *hold for β_0 and $\boldsymbol{\beta}$, respectively.*

Existence is proved by Rao (1973, pp. 523, Eq. 8a.2.16) and Corollary 12.30 implies that Equations (12.53) and (12.54) hold for β_0 and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]'$.

12.7 Parametrizations of a Discrete Conditional Expectation

Now we consider two parametrizations of a discrete conditional expectation $E(Y|Z)$. In the first one, we assume that the possible values of Z are $z_1, \ldots, z_n$. In this parametrization the coefficients are

$$\beta_0 = 0, \quad \beta_1 = E(Y|Z{=}z_1), \quad \ldots, \quad \beta_n = E(Y|Z{=}z_n).$$

Theorem 12.35 (Means as Coefficients)
Let the assumptions 12.4 hold, assume that $Z\colon (\Omega, \mathcal{A}, P) \to (\Omega'_Z, \mathcal{A}'_Z)$ is a discrete random variable, and $P(Z \in \{z_1, \ldots, z_n\}) = 1$ with $P(Z{=}z_i) > 0$, for all $i = 1, \ldots, n$. Furthermore, define $X_i := 1_{Z=z_i}$, $i = 1, \ldots, n$, and $X := (X_1, \ldots, X_n)$. Then there is a version $E(Y|Z) \in \mathcal{E}(Y|Z)$ with

$$E(Y|Z) = \sum_{i=1}^{n} \beta_i \cdot 1_{Z=z_i} = \sum_{i=1}^{n} \beta_i \cdot X_i, \tag{12.60}$$

where

$$\beta_i = E(Y|Z{=}z_i), \quad \forall\, i = 1, \ldots, n. \tag{12.61}$$

The function $g\colon \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$g(x) \;=\; 0 + \sum_{i=1}^{n} \beta_i \cdot x_i, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{12.62}$$

is a linear parametrization of $E(Y|Z)$. *If* $Z(\Omega) = \{z_1, \ldots, z_n\}$, *then* $V = V^*$ *for all* $V, V^* \in \mathscr{E}(Y|Z)$.

(Proof p. 365)

In the following theorem we present another linear parametrization of $E(Y|Z)$, generalizing Example 12.15. For convenience, now we assume that the possible values of Z are $z_0, z_1, \ldots, z_n$. In this parametrization the coefficients are

$$\begin{aligned} \beta_0 &= E(Y|Z{=}z_0), \\ \beta_1 &= E(Y|Z{=}z_1) - E(Y|Z{=}z_0), \\ &\vdots \\ \beta_n &= E(Y|Z{=}z_n) - E(Y|Z{=}z_0). \end{aligned}$$

Theorem 12.36 (Differences Between Means as Coefficients)
Let the assumptions 12.4 hold, assume that $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ *is discrete,* $P(Z \in \{z_0, z_1, \ldots, z_n\}) = 1$ *with* $P(Z{=}z_i) > 0$, *for all* $i = 0, 1, \ldots, n$. *Furthermore, define* $X_i := 1_{Z=z_i}$, $i = 1, \ldots, n$, *and* $X := (X_1, \ldots, X_n)$. *Then there is a version* $E(Y|Z) \in \mathscr{E}(Y|Z)$ *with*

$$E(Y|Z) \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i \cdot 1_{Z=z_i} \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_i, \tag{12.63}$$

where

$$\beta_0 \;=\; E(Y|Z{=}z_0) \tag{12.64}$$

and

$$\beta_i \;=\; E(Y|Z{=}z_i) - E(Y|Z{=}z_0), \quad \forall\, i = 1, \ldots, n. \tag{12.65}$$

The function $g\colon \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$g(x) \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i \cdot x_i, \quad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{12.66}$$

is a linear parametrization of $E(Y|Z)$. *If* $Z(\Omega) = \{z_0, z_1, \ldots, z_n\}$, *then* $V = V^*$ *for all* $V, V^* \in \mathscr{E}(Y|Z)$.

(Proof p. 365)

Now we treat a lemma on the covariance matrix of the indicators $1_{Z=z_i}$ for the values $z_0, z_1, \ldots, z_n$ of a discrete random variable Z. Among other things this lemma helps to prove that the covariance matrix of the indicators $1_{Z=z_1}, \ldots, 1_{Z=z_n}$ is regular. This implies that the inverse of this covariance matrix exists so that Corollary 12.30 can be applied.

Lemma 12.37 (Covariance Matrix of Indicators)
Let $Z\colon (\Omega, \mathscr{A}, P) \to (\Omega'_Z, \mathscr{A}'_Z)$ be discrete with $P(Z \in \{z_0, z_1, \ldots, z_n\}) = 1$ and $p_i := P(Z = z_i) > 0$, for all $i = 0, 1, \ldots, n$. Then:

(i) The second moments of the indicators $1_{Z=z_1}, \ldots, 1_{Z=z_n}$ are finite, and for all $i, j = 0, 1, \ldots, n$,

$$\sigma_{ij} := Cov(1_{Z=z_i}, 1_{Z=z_j}) = \begin{cases} p_i \cdot (1 - p_i), & \text{if } i = j \\ -p_i \cdot p_j, & \text{if } i \neq j. \end{cases} \tag{12.67}$$

(ii) For all $i = 1, \ldots, n$,

$$\sigma_{ii} > \sum_{j=1,\, j\neq i}^{n} |\sigma_{ij}|. \tag{12.68}$$

(Proof p. 366)

Remark 12.38 (Strict Diagonal Dominance) Proposition (ii) of Lemma 12.37 implies that the covariance matrix of the indicators $1_{Z=z_1}, \ldots, 1_{Z=z_n}$ is *strictly diagonally dominant*, i. e., it satisfies

$$|\sigma_{ii}| > \sum_{j=1,\, j\neq i}^{n} |\sigma_{ij}|, \quad \forall\, i = 1, \ldots, n. \tag{12.69}$$

◁

Remark 12.39 (Regularity of the Covariance Matrix of Indicators) According to Corollary 5.6.17 of Horn and Johnson (1991, p. 302), (12.69) implies that the covariance matrix of the indicators $1_{Z=z_1}, \ldots, 1_{Z=z_n}$ is regular [see Cor. 12.30]. In contrast, the covariance matrix of the indicators $1_{Z=z_0}, 1_{Z=z_1}, \ldots, 1_{Z=z_n}$ *is not* strictly diagonally dominant and it is *not regular*. The reason is that $1_{Z=z_0} = 1 - \sum_{i=1}^{n} 1_{Z=z_i}$, i. e., $1_{Z=z_0}$ is a linear combination of $1_{Z=z_1}, \ldots, 1_{Z=z_n}$, which implies that the covariance matrix of $1_{Z=z_0}, 1_{Z=z_1}, \ldots, 1_{Z=z_n}$ is not regular (see Exercise 12-3). ◁

Example 12.40 (Tom, Jim, and Kate – continued) Table 5.1 displays an example in which the treatment variable X has three values. The conditional expectation $E(Y|X)$ has three values as well, namely the following conditional probabilities:

$$P(Y=1|X=0) = \frac{(10+15+8)/99}{(10+10+5+15+12+8)/99} = .55,$$

$$P(Y=1|X=1) = \frac{(6+5+3)/99}{(2+6+3+5+5+3)/99} = .58\overline{3},$$

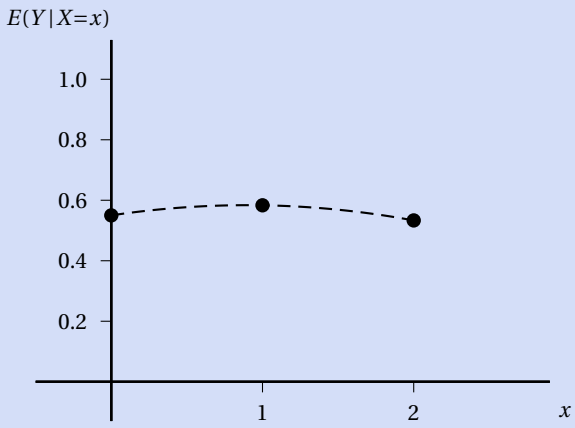


Figure 12.2. A regression with a parametrization that is linear in (X, X^2) but not in X

and

$$P(Y{=}1\,|\,X{=}2) \;=\; \frac{(4+3+1)\,/\,99}{(1+4+2+3+4+1)\,/\,99} \;=\; .53\overline{3}.$$

There are several linear parametrizations of this conditional expectation. For example, we can use the linear parametrization specified in Equation (12.33). In this case, the coefficient β_2 would inform us whether or not the conditional expectation does not only have a parametrization that is linear in (X, X^2), but also linear in X. In this example, the function $g\colon \mathbb{R} \to \mathbb{R}$ with $g(x) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$, is not a linear parametrization of $E(Y\,|\,X)$, and this is illustrated in Figure 12.2 (see Exercise 12-4). Note that X and the conditional expectation $E(Y\,|\,X) = P(Y{=}1\,|\,X)$ have only three different values.

We could also use the linear parametrizations specified in Equations (12.62) and (12.66). In the first case, the coefficients are 0 and the conditional probabilities $P(Y{=}1\,|\,X{=}0)$, $P(Y{=}1\,|\,X{=}1)$, $P(Y{=}1\,|\,X{=}2)$, whereas in the second case, the coefficients are $P(Y{=}1\,|\,X{=}0)$, $P(Y{=}1\,|\,X{=}1) - P(Y{=}1\,|\,X{=}0)$, and $P(Y{=}1\,|\,X{=}2) - P(Y{=}1\,|\,X{=}0)$. ◁

12.8 Invariance of Regression Coefficients

Remark 12.41 (Simple vs. Multiple Regression) Let the assumptions 12.4 hold with $n = 2$ and let the second moments of Y, X_1, and X_2 be finite. Assume that the inverse of the covariance matrix of (X_1, X_2) exists and that there is a version $E(Y\,|\,X_1, X_2) \in \mathscr{E}(Y\,|\,X_1, X_2)$ such that

$$E(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{12.70}$$

Furthermore, assume that there is a version $E(Y|X_1) \in \mathscr{E}(Y|X_1)$ with

$$E(Y|X_1) = \alpha_0 + \alpha_1 X_1. \tag{12.71}$$

Then $\alpha_1 = \beta_1$ does not necessarily hold. ◁

In the following theorem we formulate a sufficient condition for $\alpha_1 = \beta_1$ and generalize it to the n-variate case, where $n \geq 2$. That is, instead of $E(Y|X_1)$, we consider $E(Y|X_1,\ldots,X_m)$ and replace $E(Y|X_1,X_2)$ by $E(Y|X_1,\ldots,X_n)$, presuming $m < n$.

Theorem 12.42 (Invariance of Regression Coefficients)
Let the assumptions 12.4 hold, let $m < n$, and suppose that there is a version $E(Y|X_1,\ldots,X_n) \in \mathscr{E}(Y|X_1,\ldots,X_n)$ such that

$$E(Y|X_1,\ldots,X_n) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i + \sum_{i=m+1}^{n} \beta_i X_i. \tag{12.72}$$

If

$$\forall\, i = m+1,\ldots,n\colon \quad \Big(\beta_i = 0 \quad \text{or} \quad E(X_i|X_1,\ldots,X_m) \underset{P}{=} E(X_i)\Big), \tag{12.73}$$

then there is a version $E(Y|X_1,\ldots,X_m) \in \mathscr{E}(Y|X_1,\ldots,X_m)$ such that

$$E(Y|X_1,\ldots,X_m) = \alpha_0 + \sum_{i=1}^{m} \alpha_i X_i \tag{12.74}$$

with

$$\alpha_0 = \beta_0 + \sum_{i=m+1}^{n} \beta_i E(X_i) \tag{12.75}$$

and

$$\alpha_i = \beta_i, \quad \forall\, i = 1,\ldots,m. \tag{12.76}$$

(Proof p. 366)

Note that $E(X_i|X_1,\ldots,X_m) \underset{P}{=} E(X_i)$, for all $i = m+1,\ldots,n$, [see Eq. (12.73)] follows from

$$X_1,\ldots,X_m \underset{P}{\perp\!\!\!\perp} X_i, \quad \forall\, i = m+1,\ldots,n$$

[see Box 10.2 (vi)].

12.9 Proofs

Proof of Theorem 12.35

The existence of a version $E(Y|Z)$ satisfying (12.60) and (12.61) immediately follows from Definition 9.13 and Remark 10.15. Note that

$$\sigma(X) \;=\; \sigma(1_{Z=z_1},\dots,1_{Z=z_n}) \subset \sigma\big(\{\{Z=z_i\}\colon i=1,\dots,n\}\big) \;=\; \sigma(Z),$$

where $X := (X_1,\dots,X_n) = (1_{Z=z_1},\dots,1_{Z=z_n})$. This implies that the function g defined by Equation (12.62) is a linear parametrization of $E(Y|Z)$ with coefficients

$$\beta_0 = 0, \;\; \beta_1 = E(Y|Z{=}z_1), \;\; \dots, \;\; \beta_n = E(Y|Z{=}z_n)$$

(see Def. 12.5). Uniqueness is an immediate implication of Remark 10.16, provided that $Z(\Omega) = \{z_1,\dots,z_n\}$.

Proof of Theorem 12.36

Let $h\colon \Omega'_Z \to \mathbb{R}$ be a function such that $h(Z) \in \mathscr{E}(Y|Z)$. Then

$$\begin{aligned}
h(Z) \;&\underset{P}{=}\; \sum_{i=0}^{n} h(z_i)\cdot 1_{Z=z_i} && \text{[(5.33)]}\\
&\underset{P}{=}\; \sum_{i=0}^{n} E(Y|Z{=}z_i)\cdot 1_{Z=z_i} && \text{[(10.27)]}\\
&\underset{P}{=}\; E(Y|Z{=}z_0)\cdot 1_{Z=z_0} \;+\; \sum_{i=1}^{n} E(Y|Z{=}z_i)\cdot 1_{Z=z_i}\\
&\underset{P}{=}\; E(Y|Z{=}z_0) \;-\; \sum_{i=1}^{n} E(Y|Z{=}z_0)\cdot 1_{Z=z_i} \;+\; \sum_{i=1}^{n} E(Y|Z{=}z_i)\cdot 1_{Z=z_i} && \text{[(5.32)]}\\
&\underset{P}{=}\; E(Y|Z{=}z_0) \;+\; \sum_{i=1}^{n} \big[\,E(Y|Z{=}z_i) - E(Y|Z{=}z_0)\,\big]\cdot 1_{Z=z_i}\\
&\underset{P}{=}\; \beta_0 \;+\; \sum_{i=1}^{n} \beta_i\cdot 1_{Z=z_i},
\end{aligned}$$

where $\beta_0 := E(Y|Z{=}z_0)$ and $\beta_i := E(Y|Z{=}z_i) - E(Y|Z{=}z_0)$, for $i = 1,\dots,n$. Because the function on the right-hand side of the last equation is Z-measurable (see Rem. 2.17), it is an element of $\mathscr{E}(Y|Z)$, i. e.,

$$\beta_0 \;+\; \sum_{i=1}^{n} \beta_i\cdot 1_{Z=z_i} \;\in\; \mathscr{E}(Y|Z).$$

This proves equations (12.63), (12.64), and (12.65). Note that

$$\sigma(X) \;=\; \sigma\big(\{\{Z=z_i\}\colon i=1,\dots,n\}\big) \subset \sigma\big(\{\{Z=z_i\}\colon i=0,1,\dots,n\}\big) \subset \sigma(Z),$$

where $X := (X_1,\dots,X_n) = (1_{Z=z_1},\dots,1_{Z=z_n})$. This implies that g [see Eq. (12.66)] is a linear parametrization of $E(Y|Z)$ (see Def. 12.5). If $Z(\Omega) = \{z_0,z_1,\dots,z_n\}$, then Remark 10.16 implies that $E(Y|Z)$ is the only version in $\mathscr{E}(Y|Z)$.

Proof of Lemma 12.37

(i). For $i = j$, Equation (12.67) immediately follows from Equation (6.29) for the event $A = \{Z{=}z_i\}$. For $i \neq j$, Equation (12.67) follows from Equation (7.13) and the fact that $P(Z{=}z_i, Z{=}z_j) = 0$ if $i \neq j$. Equation (12.67) also shows that the variances and covariances of the indicators $1_{Z=z_1}, \ldots, 1_{Z=z_n}$ are finite, which implies that their second moments are finite as well [see Box 7.1 (i)].

(ii). This proposition can be derived as follows: For all $i = 1, \ldots, n$,

$$
\begin{aligned}
\sigma_{ii} &= p_i \cdot (1 - p_i) && [(12.67) \text{ for } i = j] \\
&= p_i \cdot \sum_{j=0,\, j\neq i}^{n} p_j && \Big[1 - p_i = \sum_{j=0,\, j\neq i}^{n} p_j\Big] \\
&= \sum_{j=1,\, j\neq i}^{n} p_i\, p_j \;+\; p_i \cdot p_0 \\
&= \sum_{j=1,\, j\neq i}^{n} p_i\, p_j \;+\; p_i \cdot p_0 && [p_i > 0, \text{ for all } i = 0, 1, \ldots, n] \\
&> \sum_{j=1,\, j\neq i}^{n} p_i\, p_j && [p_i \cdot p_0 > 0] \\
&= \sum_{j=1,\, j\neq i}^{n} |-p_i\, p_j| && \big[p_i\, p_j = |-p_i\, p_j|\big] \\
&= \sum_{j=1,\, j\neq i}^{n} |\sigma_{ij}|. && [(12.67) \text{ for } i \neq j]
\end{aligned}
$$

Proof of Theorem 12.42

$$
\begin{aligned}
& E(Y \mid X_1, \ldots, X_m) \\
\underset{P}{=}\; & E\big(E(Y \mid X_1, \ldots, X_n) \mid X_1, \ldots, X_m\big) && [m < n, \text{ Box 10.2 (v)}] \\
\underset{P}{=}\; & E\left(\beta_0 + \sum_{i=1}^{m} \beta_i X_i + \sum_{i=m+1}^{n} \beta_i X_i \,\middle|\, X_1, \ldots, X_m\right) && [(12.72)] \\
\underset{P}{=}\; & \beta_0 + E\left(\sum_{i=1}^{m} \beta_i X_i \,\middle|\, X_1, \ldots, X_m\right) + E\left(\sum_{i=m+1}^{n} \beta_i X_i \,\middle|\, X_1, \ldots, X_m\right) && [\text{Box 10.2 (xvi)}] \\
\underset{P}{=}\; & \beta_0 + \sum_{i=1}^{m} \beta_i X_i + \sum_{i=m+1}^{n} \beta_i\, E(X_i \mid X_1, \ldots, X_m) && [\text{Box 10.2 (xiv), (iii), (xvi)}] \\
\underset{P}{=}\; & \beta_0 + \sum_{i=1}^{m} \beta_i X_i + \sum_{i=m+1}^{n} \beta_i\, E(X_i) && [(12.73)] \\
\underset{P}{=}\; & \alpha_0 + \sum_{i=1}^{m} \beta_i X_i. && [(12.75)]
\end{aligned}
$$

12.10 Exercises

▷ **Exercise 12-1** Consider Example 12.20 and show that $\sigma(X) = \sigma(Z)$.

▷ **Exercise 12-2** Show that under the assumptions made in Remark 12.27, there is no linear parametrization g with $g(x) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$, of $E(Y|X)$ with slope $\beta_1 \neq 0$.

▷ **Exercise 12-3** Assume $X_0, X_1, \ldots, X_n \colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ are random variables with finite second moments. Show: If $X_0 = \mathbf{a}'\boldsymbol{x} + b$, where $\boldsymbol{x} := [X_1, \ldots, X_n]'$, $\mathbf{a}' := [a_1, \ldots, a_n] \in \mathbb{R}^n$, and $b \in \mathbb{R}$, then the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$ of $X_0, X_1, \ldots, X_n$ is singular.

▷ **Exercise 12-4** Compute the parameters of the quadratic function displayed in Figure 12.2 using the three conditional probabilities computed in Example 12.40.

Solutions

▷ **Solution 12-1** Because the function $\ln \colon]0,\infty[\to \mathbb{R}$ is continuous, it is $\left(\mathscr{B}\big|_{]0,\infty[}, \mathscr{B}\right)$-measurable (see Klenke, 2008, Th. 1.88, p. 36). For $X = \ln$, the inverse function is $X^{-1} = \exp$. Hence, for all $c > 0$, monotonicity of $X^{-1} = \exp$ implies

$$X^{-1}(]-\infty, \ln c]) \;=\;]0, c].$$

Note that $]-\infty, \ln c] \in \mathscr{B}$. Define $\mathscr{E}'_X := \{]-\infty, b] \colon b > 0\}$. Then

$$\begin{aligned}
\sigma(X) \;&=\; X^{-1}(\mathscr{B}) \\
&=\; X^{-1}\left(\sigma(\mathscr{E}'_X)\right) && [(1.19)] \\
&=\; \sigma\left(X^{-1}(\mathscr{E}'_X)\right) && [(2.12)] \\
&=\; \sigma\left(\{]0, c] \colon c > 0\}\right) \\
&=\; \mathscr{B}\big|_{]0,\infty[} && [(1.15), (1.19)] \\
&=\; \sigma(Z).
\end{aligned}$$

▷ **Solution 12-2** This proposition follows from

$$P[g(X)<0] + P[g(X)>1] \;=\; \begin{cases} P\left(X < -\dfrac{\beta_0}{\beta_1}\right) + P\left(X > -\dfrac{1-\beta_0}{\beta_1}\right) \;>\; 0, & \text{if } \beta_1 > 0 \\ P\left(X > -\dfrac{\beta_0}{\beta_1}\right) + P\left(X < -\dfrac{1-\beta_0}{\beta_1}\right) \;>\; 0, & \text{if } \beta_1 < 0, \end{cases}$$

because $P[g(X)<0] + P[g(X)>1] \;> 0$ is a contradiction to (12.46).

▷ **Solution 12-3**

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{xx}} \;&=\; \begin{bmatrix} \sigma^2_{X_0} & \sigma_{X_0X_1} & \cdots & \sigma_{X_0X_n} \\ \sigma_{X_1X_0} & \sigma^2_{X_1} & \cdots & \sigma_{X_1X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_nX_0} & \sigma_{X_nX_1} & \cdots & \sigma^2_{X_n} \end{bmatrix} \\
&=\; E\left[\begin{pmatrix} \mathbf{a}'\boldsymbol{x} - E(\mathbf{a}'\boldsymbol{x}) \\ \boldsymbol{x} - E(\boldsymbol{x}) \end{pmatrix} \quad \left(\mathbf{a}'\boldsymbol{x} - E(\mathbf{a}'\boldsymbol{x}),\; \boldsymbol{x}' - E(\boldsymbol{x}')\right)\right] && [X_0 = \mathbf{a}'\boldsymbol{x}, (7.33)]
\end{aligned}$$

$$= E\begin{bmatrix} \mathbf{a}'[\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')]\,\mathbf{a} & \mathbf{a}'[\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')] \\ [\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')]\,\mathbf{a} & [\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')] \end{bmatrix} \quad \text{[Box 7.2 (iii), (v)]}$$

$$= \begin{bmatrix} \mathbf{a}'E\big([\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')]\big)\,\mathbf{a} & \mathbf{a}'E\big([\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')]\big) \\ E\big([\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')]\big)\,\mathbf{a} & E\big([\boldsymbol{x}-E(\boldsymbol{x})][\boldsymbol{x}'-E(\boldsymbol{x}')]\big) \end{bmatrix}. \quad \text{[(7.27)]}$$

The first row of this matrix is obtained by multiplying the lower two submatrices by $\mathbf{a}'$ from the left. Hence, the first row of this variance-covariance matrix is a linear combination of its other rows. This implies that $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$ is singular.

▷ **Solution 12-4** For $x_1 = x$ and $x_2 = x^2$, the parametrization $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ [see Eq. (12.33)] yields

$$.55 = \beta_0,$$

if $x = 0$,

$$.58\overline{3} = \beta_0 + \beta_1 + \beta_2,$$

if $x = 1$, and

$$.53\overline{3} = \beta_0 + \beta_1 \cdot 2 + \beta_2 \cdot 4,$$

if $x = 2$. Hence, $\beta_0 = .55$ and solving the last two equations for the remaining two unknowns yields $\beta_1 \approx .0750$ and $\beta_2 \approx -.041\overline{6}$.

Chapter 13
Linear Logistic Regression

In chapter 12 we treated the notions of a linear parametrization of a conditional expectation and of a *linear regression*. In this chapter we turn to the *linear logit regression*, presuming that the regressand, say Y, is an indicator variable. In Remark 10.4 we noted that in this case a version $E(Y|\mathscr{C})$ of the $\mathscr{C}$-conditional expectation of Y is also called a version of the $\mathscr{C}$-conditional probability of the event $\{Y=1\}$ and that it is also denoted by $P(Y=1|\mathscr{C})$. As noted in Remark 12.27, if X is a $\mathscr{C}$-measurable real-valued random variable on $(\Omega,\mathscr{A},P)$ and X is P-almost surely unbounded, then there is no linear parametrization g of $E(Y|\mathscr{C}) = P(Y=1|\mathscr{C}) = g(X)$ with a nonzero slope. However, in this case, there might be a linear logistic parametrization.

We start with the logit transformation, define the logit of a $\mathscr{C}$-conditional probability $P(Y=1|\mathscr{C})$, a linear logistic parametrization and then present a theorem on uniqueness and the identification of the parameters. Finally, the concept of a linear logit regression is defined.

13.1 Logit Transformation of a Conditional Probability

The general assumptions and notation are as follows:

Notation and Assumptions 13.1
Let $Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ be a random variable with $Y \underset{P}{=} 1_{Y=1}$, let $\mathscr{C}\subset\mathscr{A}$ be a σ-algebra, and let $P(Y=1|\mathscr{C})\in\mathscr{P}(Y=1|\mathscr{C})$ with $0<P(Y=1|\mathscr{C})<1$.

Remark 13.2 (The Logit Transformation) A version $P(Y=1|\mathscr{C})$ of the conditional probability is P-almost surely bounded, because

$$P\big(\{\omega\in\Omega\colon 0\le P(Y=1|\mathscr{C})(\omega)\le 1\}\big) \;=\; 1 \tag{13.1}$$

[see Box 10.3 (iii), (iv)]. It can be transformed using the logit transformation, i. e., using the function $\operatorname{logit}\colon\,]0,1[\,\to\mathbb{R}$ defined by

$$\operatorname{logit}(p) \;:=\; \ln\left(\frac{p}{1-p}\right) \;=\; \ln(p) \;-\; \ln(1-p), \quad \forall\, p\in\,]0,1[\,, \tag{13.2}$$

where ln denotes the natural logarithm. If p represents a probability, then $\ln\left(\frac{p}{1-p}\right)$

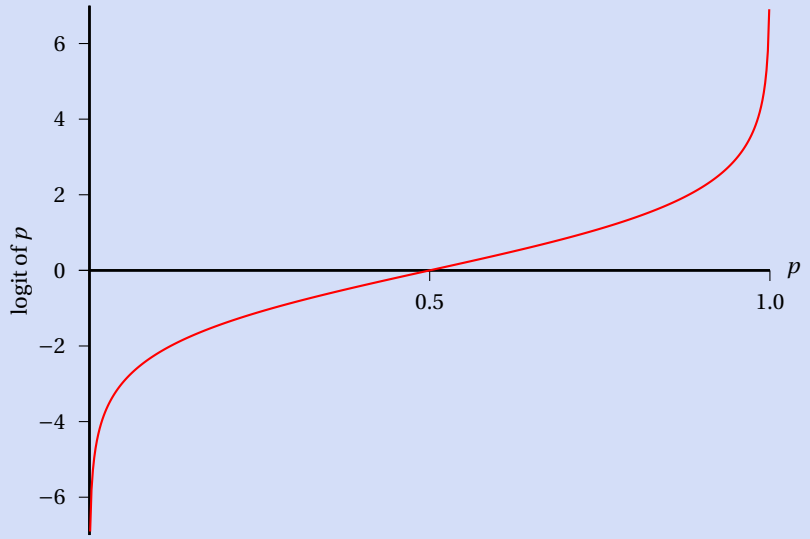


Figure 13.1. Graph of the logit transformation of p

is also called the *log-odds* of p. Figure 13.1 shows the graph of such a logit transformation. In the context of generalized linear models [see, e. g., McCullagh & Nelder, 1989; ?, ?], this function is an example of a *link function*. Some algebra yields

$$p = \frac{\exp[\text{logit}(p)]}{1+\exp[\text{logit}(p)]}, \quad \forall p \in]0,1[\,, \tag{13.3}$$

(see Exercise 13-1). ◁

Definition 13.3 (Logit of $P(Y=1|\mathscr{C})$)
Let the assumptions 13.1 hold. Then

$$\text{logit}\,[P(Y=1|\mathscr{C})] := \ln\left(\frac{P(Y=1|\mathscr{C})}{1-P(Y=1|\mathscr{C})}\right), \tag{13.4}$$

is called the logit of $P(Y=1|\mathscr{C})$.

Remark 13.4 (One-to-One Transformation) Note that logit $[P(Y=1|\mathscr{C})]$ denotes the composition of $P(Y=1|\mathscr{C})$ and the function logit defined by Equation (13.2). Hence, logit $[P(Y=1|\mathscr{C})]$ is a random variable on $(\Omega, \mathscr{A}, P)$. Also note that $P(Y=1|\mathscr{C})$ and its logit contain the same information, i. e., the σ-algebras they generate are identical (see Lemma 13.5). While $P(Y=1|\mathscr{C})$ informs us about the likelihood of the event $\{Y=1\}$ *on the probability scale* with values between 0 and 1, a logit

of $P(Y{=}1|\mathscr{C})$ informs us about this likelihood *on the log-odds scale* with values between $-\infty$ and ∞ [see Eq. (13.2)]. Applying the exponential function on both sides of Equation (13.4) and rearranging terms (see Exercise 13-1) yields

$$P(Y{=}1|\mathscr{C}) \;=\; \frac{\exp\big(\operatorname{logit}[P(Y{=}1|\mathscr{C})]\big)}{1+\exp\big(\operatorname{logit}[P(Y{=}1|\mathscr{C})]\big)}. \tag{13.5}$$

Hence, $P(Y{=}1|\mathscr{C})$ and $\operatorname{logit}[P(Y{=}1|\mathscr{C})]$ are one-to-one transformations of each other. ◁

Lemma 13.5 (σ-Algebra Generated by $P(Y{=}1|\mathscr{C})$ and its Logit)
Let the assumptions 13.1 hold. Then

$$\sigma[P(Y{=}1|\mathscr{C})] \;=\; \sigma\big(\operatorname{logit}[P(Y{=}1|\mathscr{C})]\big). \tag{13.6}$$

(Proof p. 378)

Remark 13.6 (Motivation for Considering the Logit of $P(Y{=}1|\mathscr{C})$) Suppose that X is a real-valued $\mathscr{C}$-measurable random variable on $(\Omega,\mathscr{A},P)$ and that X is not P-almost surely bounded [see Eq. (12.47)]. Then it is still possible to assume that there is a version $P(Y{=}1|\mathscr{C}) \in \mathscr{P}(Y{=}1|\mathscr{C})$ and numbers $\lambda_0,\lambda_1 \in \mathbb{R}$ such that

$$\operatorname{logit}[P(Y{=}1|\mathscr{C})] \;=\; \lambda_0 + \lambda_1 X. \tag{13.7}$$

In contrast, assuming $P(Y{=}1|\mathscr{C}) = \lambda_0 + \lambda_1 X$ for real numbers λ_0,λ_1, $\lambda_1 \neq 0$, would be contradictory if X is not P-almost surely bounded. ◁

13.2 Linear Logistic Parametrization

Notation and Assumptions 13.7
Let $Y\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$ be a random variable with $Y \underset{P}{=} 1_{Y=1}$, let $\mathscr{C} \subset \mathscr{A}$ be a σ-algebra, and let $P(Y{=}1|\mathscr{C}) \in \mathscr{P}(Y{=}1|\mathscr{C})$ with $0 < P(Y{=}1|\mathscr{C}) < 1$. Furthermore, let $X_i\colon (\Omega,\mathscr{A},P) \to (\mathbb{R},\mathscr{B})$, $i = 1,\ldots,n$, be real-valued random variables, define $X := (X_1,\ldots,X_n)\colon (\Omega,\mathscr{A},P) \to (\mathbb{R}^n,\mathscr{B}_n)$, and assume that X is $\mathscr{C}$-measurable.

Using this notation and these assumptions a linear logistic parametrization of a conditional probability $P(Y{=}1|\mathscr{C})$ is now defined as follows:

Definition 13.8 (Linear Logistic Parametrization)
Let the assumptions 13.7 hold. If there are $\lambda_0,\lambda_1,\ldots,\lambda_n \in \mathbb{R}$ and a version $P(Y{=}1|\mathscr{C}) \in \mathscr{P}(Y{=}1|\mathscr{C})$ such that

$$P(Y{=}1|\mathscr{C}) \;=\; \frac{\exp(\lambda_0+\sum_{i=1}^n \lambda_i X_i)}{1+\exp(\lambda_0+\sum_{i=1}^n \lambda_i X_i)}, \tag{13.8}$$

then the function $g\colon \mathbb{R}^n \to [0,1]$ *satisfying*

$$g(x) \;=\; \frac{\exp(\lambda_0+\sum_{i=1}^n \lambda_i x_i)}{1+\exp(\lambda_0+\sum_{i=1}^n \lambda_i x_i)}, \quad \forall x \in \mathbb{R}^n, \tag{13.9}$$

is called a linear logistic parametrization of $P(Y{=}1|\mathscr{C})$.

If $\mathscr{C} = \sigma(X)$, then Equation (13.8) is equivalent to

$$P(Y{=}1|X) \;=\; \frac{\exp(\lambda_0+\sum_{i=1}^n \lambda_i X_i)}{1+\exp(\lambda_0+\sum_{i=1}^n \lambda_i X_i)}. \tag{13.10}$$

Remark 13.9 (Univariate Real-Valued X) If Equation (13.8) holds for $n=1$, then there is a version $P(Y{=}1|X) \in \mathscr{P}(Y{=}1|X)$ such that

$$P(Y{=}1|\mathscr{C}) \;=\; P(Y{=}1|X) \;=\; \frac{\exp(\lambda_0+\lambda_1 X)}{1+\exp(\lambda_0+\beta_1 X)}. \tag{13.11}$$

In Example 17.80 we present a sufficient condition of Equation (13.11) related to the normal distribution. ◁

Remark 13.10 (Conditional Probabilities) If g is a linear logistic parametrization of $E(Y|X)$ satisfying Equation (13.9), then, according to Definition 10.33, we can define

$$\begin{aligned} P(Y{=}1|X{=}x) \;&=\; P(Y{=}1|X_1{=}x_1,\ldots,X_n{=}x_n) \\ &:=\; g(x) \;=\; \frac{\exp(\lambda_0+\sum_{i=1}^n \lambda_i x_i)}{1+\exp(\lambda_0+\sum_{i=1}^n \lambda_i x_i)}, \quad \forall x \in \mathbb{R}^n. \end{aligned} \tag{13.12}$$

This definition is convenient, but note that another factorization g^* of $P(Y{=}1|X)$ might yield other conditional probabilities $P(Y{=}1|X{=}x)$ for values x of X with $P(X{=}x)=0$. However, according to Equation (12.10), if g, g^* are factorizations of two versions $V, V^* \in \mathscr{P}(Y{=}1|X)$, then $g(x) = g^*(x)$, for P_X-almost all $x \in \mathbb{R}^n$ [see Eq. (12.13)]. ◁

Remark 13.11 (Meaning of Coefficients) Figure 13.2 displays the graphs of logistic transformations in which the logits of $P(Y{=}1|X)$ are linear functions $\lambda_0 + \lambda_1 X$. As is easily seen,

$$P\left(Y{=}1 \;\middle|\; X = -\frac{\lambda_0}{\lambda_1}\right) \;=\; \frac{\exp(0)}{1+\exp(0)} \;=\; \frac{1}{1+1} \;=\; \frac{1}{2}.$$

This equation shows that $x = -\frac{\lambda_0}{\lambda_1}$ is the point on the x-axis at which the con-

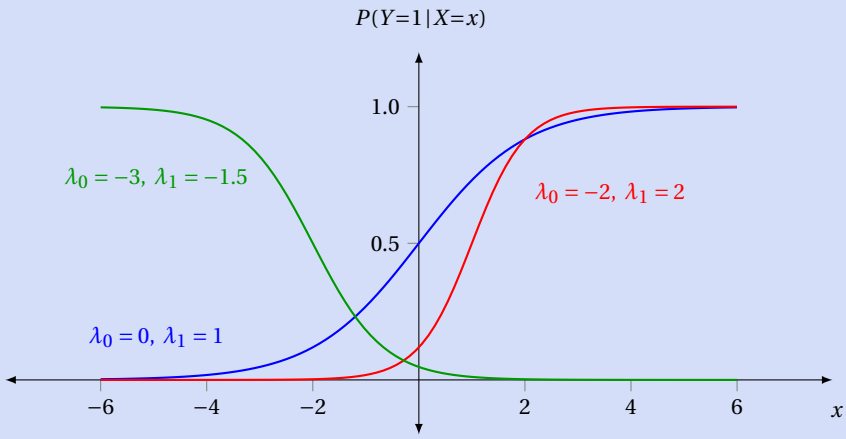


Figure 13.2. Graphs of three logistic functions

ditional probability $P(Y{=}1\,|\,X{=}x)$ is .5. Furthermore, the derivative of the linear parametrization g with respect to x is

$$\frac{d}{dx}\,g(x) \;=\; \frac{d}{dx}\,\frac{\exp(\lambda_0+\lambda_1 x)}{1+\exp(\lambda_0+\lambda_1 x)} \;=\; \frac{\lambda_1\exp(\lambda_0+\lambda_1 x)}{\big(1+\exp(\lambda_0+\lambda_1 x)\big)^2}\,, \tag{13.13}$$

Hence, the derivative (i. e., the slope) of g at $x=-\frac{\lambda_0}{\lambda_1}$ is $\frac{\lambda_1}{4}$ (see Exercise 13-2). ◁

Nagel: Folgenden Abschnitt eingefügt

13.3 A Parametrization of a Discrete Conditional Probability

In Theorem 12.36 we already considered a parametrization of a discrete conditional expectation $E(Y\,|\,Z)$ in which the parameters $\beta_1,\ldots,\beta_n$ are the differences $E(Y\,|\,Z{=}z_i) \;-\; E(Y\,|\,Z{=}z_0)$, $i=1,\ldots,n$. If Y is dichotomous, then there is also a logistic parametrization of the conditional probability $E(Y\,|\,Z) = P(Y{=}1\,|\,Z)$.

Theorem 13.12 (Existence of the Logit Effects)
Let $Y\colon(\Omega,\mathscr{A},P)\to(\mathbb{R},\mathscr{B})$ and $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ be random variables, where $Y \underset{P}{=} 1_{Y=1}$ and Z is discrete with $P(Z\in\{z_0,z_1,\ldots,z_n\}) = 1$ and $P(Z{=}z_i)$, $P(Y{=}1\,|\,Z{=}z_i) > 0$, for all $i=0,1,\ldots,n$. Then $P(Y{=}1|Z)$ is uniquely defined and there are coefficients $\beta_0,\beta_1,\ldots,\beta_n,\lambda_0,\lambda_1,\ldots,\lambda_n\in\mathbb{R}$ such that

$$P(Y{=}1|Z) \;=\; \beta_0 \;+\; \sum_{i=1}^{n}\beta_i\cdot 1_{Z=z_i} \tag{13.14}$$

$$= \frac{\exp\left[\lambda_0 + \sum_{i=1}^n \lambda_i \cdot 1_{Z=z_i}\right]}{1+\exp\left[\lambda_0 + \sum_{i=1}^n \lambda_i \cdot 1_{X=x_i}\right]} \tag{13.15}$$

with

$$\beta_0 = P(Y=1|Z=z_0) \tag{13.16}$$

$$= \frac{\exp(\lambda_0)}{1+\exp(\lambda_0)} \tag{13.17}$$

and

$$\beta_i = P(Y=1|Z=z_0) - P(Y=1|Z=z_i) \tag{13.18}$$

$$= \frac{\exp(\lambda_0+\lambda_i)}{1+\exp(\lambda_0+\lambda_i)} - \frac{\exp(\lambda_0)}{1+\exp(\lambda_0)}. \tag{13.19}$$

(Proof p. 379)

Remark 13.13 (Log Odds) In terms of conditional probabilities, the logit intercept can be written

$$\lambda_0 = \ln\left[\frac{P(Y=1|Z=z_0)}{1-P(Y=1|Z=z_0)}\right]. \tag{13.20}$$

Hence, λ_0 is the log odds of $P(Y=1|Z=z_0)$. Similarly,

$$\lambda_0+\lambda_i = \ln\left[\frac{P(Y=1|Z=z_i)}{1-P(Y=1|Z=z_i)}\right], \quad \forall\, i=1,\ldots,n, \tag{13.21}$$

[see Eqs. (13.16) to (13.19) and (13.2)]. This equation shows that $\lambda_0+\lambda_i$ is the log odds of $P(Y=1|Z=z_i)$. ◁

Remark 13.14 (Log Odds Ratio) Equations (13.20) and (13.21) immediately imply

$$\lambda_i = \ln\left[\frac{P(Y=1|Z=z_i)}{1-P(Y=1|Z=z_i)}\right] - \ln\left[\frac{P(Y=1|Z=z_0)}{1-P(Y=1|Z=z_0)}\right] \tag{13.22}$$

$$= \ln\left[\frac{\frac{P(Y=1|Z=z_i)}{1-P(Y=1|Z=z_i)}}{\frac{P(Y=1|Z=z_0)}{1-P(Y=1|Z=z_0)}}\right], \quad \forall\, i=1,\ldots,n, \tag{13.23}$$

for the logit effect λ_i. Hence, λ_i, $i=1,\ldots,n$, is the difference between the log odds of $P(Y=1|Z=z_i)$ and $P(Y=1|Z=z_0)$, respectively [see Eq. (13.22)]. Equation (13.23) shows that λ_i is the log odds ratio of $P(Y=1|Z=z_i)$ and $P(Y=1|Z=z_0)$. ◁

Remark 13.15 (Odds Ratio) The exponential function of λ_i is

$$\exp(\lambda_i) \;=\; \frac{\dfrac{P(Y=1|Z=z_i)}{1-P(Y=1|Z=z_i)}}{\dfrac{P(Y=1|Z=z_0)}{1-P(Y=1|Z=z_0)}}, \quad \forall\, i=1,\ldots,n. \tag{13.24}$$

This equation shows that the number $\exp(\lambda_i)$ is the odds ratio of $P(Y=1|Z=z_i)$ and $P(Y=1|Z=z_0)$. ◁

Remark 13.16 (Risk Ratio) Another closely related parameter is

$$\kappa_i \;:=\; \frac{P(Y=1|Z=z_i)}{P(Y=1|Z=z_0)}, \quad \forall\, i=1,\ldots,n. \tag{13.25}$$

This parameter is called the *risk ratio of* $P(Y=1|Z=z_i)$ *and* $P(Y=1|Z=z_0)$. ◁

Remark 13.17 (Four Effect Parameters) Hence, under the assumptions of Theorem 13.12, we may consider four different effect parameters: β_i, λ_i, $\exp(\lambda_i)$, and κ_i. They all quantify the effect of x_i compared to x_0 on Y, each one on a different scale. ◁

üüüü Nagel

13.4 Identification of Coefficients of a Linear Logistic Parametrization

The following theorem specifies the conditions under which a linear logit parametrization of $P(Y=1|X)$ is uniquely defined.

Theorem 13.18 (Identification of Coefficients and Uniqueness)
Let the assumptions 13.7 hold and let $\lambda_0, \lambda_1, \ldots, \lambda_n \in \mathbb{R}$. *Furthermore, assume:*

- *(a) there is a version* $P(Y=1|\mathscr{C}) \in \mathscr{P}(Y=1|\mathscr{C})$ *such that Equation* (13.8) *holds,*
- *(b)* $X_1, \ldots, X_n$ *have finite second moments, and*
- *(c) the inverse* $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ *of the covariance matrix of* $X = (X_1, \ldots, X_n)$ *exists.*

Then, using

$$L \;:=\; \operatorname{logit}[P(Y=1|\mathscr{C})] \;=\; \lambda_0 + \boldsymbol{\lambda}'\boldsymbol{x}, \tag{13.26}$$

the following two equations hold:

$$\lambda_0 \;=\; E(L) - \boldsymbol{\lambda}'\boldsymbol{\mu}, \tag{13.27}$$

$$\boldsymbol{\lambda} \;=\; \boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xl}}, \tag{13.28}$$

where $\boldsymbol{\mu} := [E(X_1),\ldots,E(X_n)]'$ *and* $\boldsymbol{\Sigma_{xl}}$ *denotes the column vector of the covariances* $Cov(X_i, L)$*. The coefficients* λ_0 *and* $\boldsymbol{\lambda}$ *are uniquely determined and hence, the linear logistic parametrization* $g\colon \mathbb{R}^n \to [0,1]$ *of* $P(Y{=}1|\mathscr{C})$ *satisfying* (13.9) *is uniquely defined.*

(Proof p. 380)

Remark 13.19 (Identification vs. Estimation) Note that, for a version $P(Y{=}1|\mathscr{C}) \in \mathscr{P}(Y{=}1|\mathscr{C})$, the logit of $P(Y{=}1|\mathscr{C})$ is uniquely defined. This also applies to the expectation vector $\boldsymbol{\mu} = [E(X_1),\ldots,E(X_n)]$, the covariance matrix $\boldsymbol{\Sigma_{xx}}$, and the covariance vector $\boldsymbol{\Sigma_{xl}}$ in Equation (13.28). This means that the coefficients of the linear logistic parametrization are uniquely defined (or 'identified'). Estimation in the logistic case is more difficult as compared to the linear regression, because $L = \text{logit}[P(Y{=}1|\mathscr{C})]$ is a nonobservable random variable (see Rem. 12.2). For methods of estimation, see, e. g., McCullagh and Nelder (1989). ◁

13.5 Linear Logistic Regression and Linear Logit Regression

The uniqueness property formulated in Theorem 13.18 allows us to define a linear logit regression as follows:

Definition 13.20 (Linear Logistic and Linear Logit Regression)
Let the assumptions 13.7 hold and suppose there are an $E(Y|\mathscr{C}) \in \mathscr{E}(Y|\mathscr{C})$ *and* $\lambda_0, \lambda_1, \ldots, \lambda_n \in \mathbb{R}$ *such that Equation* (13.8) *holds. Furthermore, assume that* Y *and* $X_1,\ldots,X_n$ *have finite second moments and that the inverse* $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}$ *of the covariance matrix of* $X = (X_1,\ldots,X_n)$ *exists. Then the function* $g\colon \mathbb{R}^n \to [0,1]$ *defined by Equation* (13.9) *is called the linear logistic regression or the linear inverse logit regression and the function* $f\colon \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$f(x) \;=\; \lambda_0 + \sum_{i=1}^{n} \lambda_i\, x_i, \quad \forall\, x \in \mathbb{R}^n, \tag{13.29}$$

the linear logit regression of Y *on* X*.*

Figure 13.3 shows $P(Y{=}1\,|\,X)$ as the composition of the functions introduced above. According to this figure, $P(Y{=}1\,|\,X)$ is the composition of X and the linear logistic regression g, which itself is the composition of the linear logit regression f and the logistic function h that transforms a logit into a probability.

Remark 13.21 (Simple and Multiple Linear Logistic Regression) If $n \geq 2$, then a linear logistic regression is also called a *multiple linear logistic regression*. If $n = 1$, then it is also called a *simple linear logistic regression.* ◁

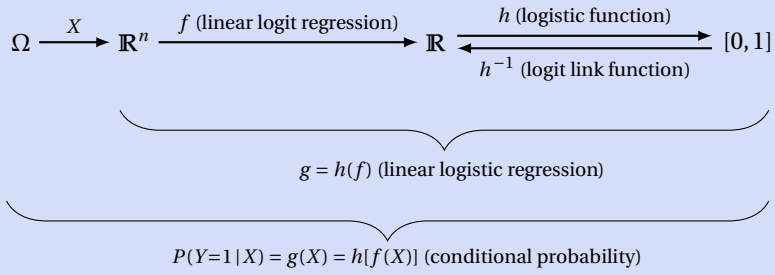


Figure 13.3. $P(Y{=}1\,|\,X)$ as the composition of X, the linear logit regression f, and the logistic function h.

Example 13.22 (Joe and Ann With Random Assignment – continued) Table 9.1 shows the random variables U, X, and Y as well as the conditional expectations $E(Y|X,U)$ and $E(Y|X)$. Considering $E(Y|X) = P(Y{=}1|X)$, we may write

$$P(Y{=}1|X) \;=\; \frac{\exp(\alpha_0+\alpha_1\cdot X)}{1+\exp(\alpha_0+\alpha_1\cdot X)} \;\approx\; \frac{\exp(-.2007+.6061\cdot X)}{1+\exp(-.2007+.6061\cdot X)} \tag{13.30}$$

(see Exercise 13-3). Inserting the two values of X, this equation yields the probabilities

$$P(Y{=}1|X)(\omega) \;=\; P(Y{=}1\,|\,X{=}0) \;=\; .45, \quad \text{for } \omega\in\{X{=}0\}$$

and

$$P(Y{=}1|X)(\omega) \;=\; P(Y{=}1\,|\,X{=}1) \;=\; .60, \quad \text{for } \omega\in\{X{=}1\}.$$

The function $g\colon \mathbb{R}^n \to [0,1]$ defined by

$$g(x) \;\approx\; \frac{\exp(-.2007+.6061\cdot x)}{1+\exp(-.2007+.6061\cdot x)}, \quad \forall\, x\in\mathbb{R},$$

is the linear logistic regression of Y on X.

Considering the conditional expectation $E(Y|X,U) = P(Y{=}1|X,U)$ we can write

$$P(Y{=}1|X,U) \;=\; \frac{\exp\left(\lambda_0+\lambda_1\cdot X+\lambda_2\cdot 1_{U=Ann}+\lambda_3\cdot X\cdot 1_{U=Ann}\right)}{1+\exp\left(\lambda_0+\lambda_1\cdot X+\lambda_2\cdot 1_{U=Ann}+\lambda_3\cdot X\cdot 1_{U=Ann}\right)}$$

$$\approx \frac{\exp\left(.8473+.5390\cdot X-2.2336\cdot 1_{U=Ann}+.4418\cdot X\cdot 1_{U=Ann}\right)}{1+\exp\left(.8473+.5390\cdot X-2.2336\cdot 1_{U=Ann}+.4418\cdot X\cdot 1_{U=Ann}\right)}$$

(see Exercise 13-4) Inserting the two values of X and the two values of $1_{U=Ann}$, this equation yields the four probabilities $P(Y{=}1|X{=}x,U{=}u)$ listed in Table 9.1. The function $g\colon \mathbb{R}^3 \to [0,1]$ defined by

$$g(x) \;\approx\; \frac{\exp\left(.8473+.5390\cdot x_1-2.2336\cdot x_2+.4418\cdot x_3\right)}{1+\exp\left(.8473+.5390\cdot x_1-2.2336\cdot x_2+.4418\cdot x_3\right)}, \quad \forall\, x\in\mathbb{R}^3,$$

is the linear logistic regression of Y on $X = (X_1, X_2, X_3) = (X, 1_{U=Ann}, X \cdot 1_{U=Ann})$.

Rearranging the equation for $P(Y{=}1|X,U)$ yields

$$P(Y{=}1|X,U) \underset{P}{=} \frac{\exp\left((\lambda_0 + \lambda_2 \cdot 1_{U=Ann}) + (\lambda_1 + \lambda_3 \cdot 1_{U=Ann}) \cdot X\right)}{1 + \exp\left((\lambda_0 + \lambda_2 \cdot 1_{U=Ann}) + (\lambda_1 + \lambda_3 \cdot 1_{U=Ann}) \cdot X\right)}$$

$$\approx \frac{\exp\left((.8473 - 2.2336 \cdot 1_{U=Ann}) + (.5390 + .4418 \cdot 1_{U=Ann}) \cdot X\right)}{1 + \exp\left((.8473 - 2.2336 \cdot 1_{U=Ann}) + (.5390 + .4418 \cdot 1_{U=Ann}) \cdot X\right)},$$

showing that the logit is $f_0(U) + f_1(U) \cdot X$ with logit intercept function

$$f_0(U) \;=\; \lambda_0 + \lambda_2 \cdot 1_{U=Ann} \;\approx\; .8473 - 2.2336 \cdot 1_{U=Ann}$$

and logit effect function

$$f_1(U) \;=\; \lambda_1 + \lambda_3 \cdot 1_{U=Ann} \;=\; .5390 + .4418 \cdot 1_{U=Ann}.$$

Note that

$$\begin{aligned} &\alpha_1 \approx .6061 \\ \neq\; &E[f_1(U)] \approx .5390 + .4418 \cdot E(1_{U=Ann}) \approx .7599. \end{aligned}$$

Hence, although X and U are independent, the slope α_1 of the logit in the logistic parametrization of $E(Y|X) = P(Y{=}1|X)$ is *not* equal to the expectation of the logit effect function $f_1(U)$ of the logit in the logistic parametrization of $E(Y|X,U) = P(Y{=}1|X,U)$.

From a methodological point of view this means that random assignment of a unit to one of two treatment conditions—which creates independence of a treatment variable X and the person variable U—does not imply that the slope α_1 of the logit in the logistic parametrization of $E(Y|X) = P(Y{=}1|X)$ can be interpreted as an average effect of treatment variable on Y [or on the logit of the linear logit parametrization of $E(Y|X)$]. In examples in which $f_1(U) = \lambda_1$ is a constant, this implies that $\alpha_1 = \lambda_1$ does not follow from independence of X and U. In contrast, compare the corresponding invariance property formulated in Theorem 12.42 for a linear parametrization. ◁

Remark 13.23 (Existence of a Linear Logistic Regression) Here is a sufficient condition for the existence of a linear logistic regression. Let $X\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a continuous random variable, let $A \in \mathscr{A}$ with $0 < P(A) < 1$, and let $Y = 1_A$. Furthermore, assume that X has a normal distribution with respect to each of the two conditional-probability measures $P^{Y=y}$, $y = 0, 1$, and that $Var(X \mid Y{=}0) = Var(X \mid Y{=}1)$. Then $P(Y{=}1|X)$ has a linear logistic parametrization (for a proof see Examples 17.80 to 17.82). ◁

13.6 Proofs

Proof of Lemma 13.5

Let $V\colon (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$ be a version in $\mathscr{P}(Y{=}1|\mathscr{C})$ with values in $]0,1[$. Furthermore, let $g\colon]0,1[\to \mathbb{R}$ be a continuous and strictly monotone function with $g\big(]0,1[\big) = \mathbb{R}$. Then

the inverse function $g^{-1}\colon \mathbb{R} \to]0,1[$ exists (see Ellis & Gulick, 2006, section 7.1) and it is continuous (Ellis & Gulick, 2006, Th. 7.4, p. 436) and strictly monotone with $g^{-1}(\mathbb{R}) =]0,1[$. Hence,

$$
\begin{aligned}
&\sigma[g(V)] \\
&= (g \circ V)^{-1}(\mathscr{B}) && [(2.14)] \\
&= V^{-1} \circ g^{-1}(\mathscr{B}) && [(2.29)] \\
&= V^{-1} \circ g^{-1}\left[\sigma\left(\{]-\infty, b]\colon b \in \mathbb{R}\}\right)\right] && [(1.19)] \\
&= \sigma\left[V^{-1} \circ g^{-1}\left(\{]-\infty, b]\colon b \in \mathbb{R}\}\right)\right] && [(2.12)] \\
&= \sigma\left[V^{-1}\left(\{]-\infty, g^{-1}(b)]\colon b \in \mathbb{R}\}\right)\right] && [\text{monotonicity, continuity}] \\
&= \sigma\left[V^{-1}\left(\{]-\infty, c]\colon c \in]0,1[\}\right)\right] && [\text{codomain of } g] \\
&= \sigma\left[V^{-1}\left(\{]-\infty, b]\colon b \in \mathbb{R}\}\right)\right] \quad \left[b \le 0\colon V^{-1}\left(]-\infty, b]\right) = \emptyset,\ b \ge 1\colon V^{-1}\left(]-\infty, b]\right) = \mathbb{R}\right] \\
&= \sigma\left[V^{-1}\left(\{]-\infty, b]\colon b \in \mathbb{R}\}\right)\right] && [(2.12)] \\
&= V^{-1}(\mathscr{B}) && [(1.19)] \\
&= \sigma(V). && [(2.14)]
\end{aligned}
$$

The result $\sigma[g(V)] = \sigma(V)$ can now be applied to $g = \text{logit}$ [see Equation (13.2)].

Proof of Theorem 13.12

By definition, $P(Y{=}1|Z) = E(1_{Y=1}|Z)$. Hence, the existence of coefficients $\beta_0, \beta_1, \ldots, \beta_n$ and a version $P(Y{=}1|Z) \in \mathscr{P}(Y{=}1|Z)$ satisfying Equations (13.14), (13.16), and (13.18) has already been proved in Theorem 12.36. In order to show that there are $\lambda_0, \lambda_1, \ldots, \lambda_n$ satisfying Equation (13.15) we define

$$\lambda_0 := \text{logit}[P(Y{=}1|Z{=}z_0)], \tag{13.31}$$

[see Eq. (13.2)] and

$$\lambda_i := \text{logit}[P(Y{=}1|Z{=}z_i)] - \text{logit}[P(Y{=}1|Z{=}z_0)], \quad \forall\, i = 1, \ldots, n. \tag{13.32}$$

These definitions and Equation (13.3) then yield

$$P(Y{=}1|Z{=}z_0) = \frac{\exp\left(\text{logit}[P(Y{=}1|Z{=}z_0)]\right)}{1 + \exp\left(\text{logit}[P(Y{=}1|Z{=}z_0)]\right)} = \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)},$$

and

$$P(Y{=}1|Z{=}z_i) = \frac{\exp\left(\text{logit}[P(Y{=}1|Z{=}z_i)]\right)}{1 + \exp\left(\text{logit}[P(Y{=}1|Z{=}z_i)]\right)} = \frac{\exp(\lambda_0 + \lambda_i)}{1 + \exp(\lambda_0 + \lambda_i)}, \quad \forall\, i = 1, \ldots, n.$$

Hence, Equation (13.16) implies

$$\beta_0 = P(Y{=}1|Z{=}z_0) = \frac{\exp\left(\lambda_0\right)}{1 + \exp\left(\lambda_0\right)},$$

and Equation (13.18) yields

$$
\begin{aligned}
\beta_i &= P(Y{=}1|Z{=}z_i) - P(Y{=}1|Z{=}z_0) \\
&= \frac{\exp(\lambda_0 + \lambda_i)}{1 + \exp(\lambda_0 + \lambda_i)} - \frac{\exp(\lambda_0)}{1 + \exp(\lambda_0)}, \quad \forall\, i = 1, \ldots, n.
\end{aligned}
$$

Proof of Theorem 13.18

Denote $\boldsymbol{x} := [X_1, \ldots, X_n]'$, $\boldsymbol{\mu} := [E(X_1), \ldots, E(X_n)]'$, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]'$, as well as

$$L := \text{logit}[P(Y{=}1|\mathscr{C})] = \lambda_0 + \boldsymbol{\lambda}' \boldsymbol{x}$$

[see Eqs. (13.26) and (13.5)]. Taking the expectation on both sides, using the definition of $\boldsymbol{\mu}$ and rearranging yields

$$\lambda_0 = E(L) - \boldsymbol{\lambda}' \boldsymbol{\mu}.$$

Furthermore, consider the n-dimensional covariance vector

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{z}} &= \boldsymbol{\Sigma}_{\boldsymbol{x}, \lambda_0 + \boldsymbol{\lambda}'\boldsymbol{x}} && \text{[(13.26)]} \\ &= \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\lambda}. && \text{[Box 7.3 (ii), (iii)]} \end{aligned}$$

Multiplying both sides by $\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1}$ yields

$$\boldsymbol{\lambda} = \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{l}}.$$

This equation also shows that the vector $\boldsymbol{\lambda}$ is uniquely defined, and this implies that λ_0 is uniquely defined as well. Uniqueness of λ_0 and $\boldsymbol{\lambda}$ implies that the linear logit parametrization g satisfying Equation (13.9) is uniquely defined as well.

13.7 Exercises

▷ **Exercise 13-1** Prove Equation (13.5).

▷ **Exercise 13-2** Calculate the derivative (13.13).

▷ **Exercise 13-3** Consider Example 13.22 and compute the coefficients of the linear logit parametrization of $E(Y|X)$.

▷ **Exercise 13-4** Consider Example 13.22 and compute the coefficients of the linear logit parametrization of $E(Y|X,U)$.

Solutions

▷ **Solution 13-1**

$$\exp[\text{logit}(p)] = \exp\left[\ln\left(\frac{p}{1-p}\right)\right] = \frac{p}{1-p}.$$

Hence,

$$p = \frac{\frac{p}{1-p}}{\frac{1}{1-p}} = \frac{\frac{p}{1-p}}{\frac{1-p+p}{1-p}} = \frac{\frac{p}{1-p}}{\frac{1-p}{1-p} + \frac{p}{1-p}} = \frac{\frac{p}{1-p}}{1 + \frac{p}{1-p}} = \frac{\exp[\text{logit}(p)]}{1 + \exp[\text{logit}(p)]}.$$

▷ **Solution 13-2** The chain rule and the quotient rule of differential calculus yield

$$\begin{aligned}\frac{d}{dx}g(x) &= \frac{d}{dx}\,\frac{\exp(\lambda_0+\lambda_1 x)}{1+\exp(\lambda_0+\lambda_1 x)}\\ &= \frac{\lambda_1\exp(\lambda_0+\lambda_1 x)\big(1+\exp(\lambda_0+\lambda_1 x)\big)-\exp(\lambda_0+\lambda_1 x)\cdot\lambda_1\exp(\lambda_0+\lambda_1 x)}{\big(1+\exp(\lambda_0+\lambda_1 x)\big)^2}\\ &= \frac{\lambda_1\exp(\lambda_0+\lambda_1 x)\big(1+\exp(\lambda_0+\lambda_1 x)-\exp(\lambda_0+\lambda_1 x)\big)}{\big(1+\exp(\lambda_0+\lambda_1 x)\big)^2}\\ &= \frac{\lambda_1\,\exp(\lambda_0+\lambda_1 x)}{\big(1+\exp(\lambda_0+\lambda_1 x)\big)^2}.\end{aligned}$$

▷ **Solution 13-3** Inserting the value $x = 0$ in the equation

$$\text{logit}[P(Y{=}1|X{=}x)] \;=\; \ln\left(\frac{P(Y{=}1|X{=}x)}{1-P(Y{=}1|X{=}x)}\right) \;=\; \alpha_0+\alpha_1 x$$

[see Eqs. (13.2) and (13.12)] yields

$$\text{logit}[P(Y{=}1|X{=}0)] \;=\; \ln\left(\frac{.45}{1-.45}\right) \;\approx\; -.2007 \;\approx\; \alpha_0,$$

and inserting the value $x = 1$ yields

$$\text{logit}[P(Y{=}1|X{=}1)] \;=\; \ln\left(\frac{.6}{1-.6}\right) \;\approx\; .4055 \;\approx\; \alpha_0+\alpha_1.$$

[In R these values are obtained by `qlogis(.45)` and `qlogis(.60)`, respectively.] Solving the last equation yields $\alpha_1 \approx .4055-(-.2007) = .6061$.

▷ **Solution 13-4** For $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$, the equation

$$\text{logit}[P(Y{=}1|X{=}0,U{=}Ann)] \;=\; \ln\left(\frac{P(Y{=}1|X{=}0,U{=}Ann)}{1-P(Y{=}1|X{=}0,U{=}Ann)}\right) \;=\; \lambda_0+\lambda_1 x_1+\lambda_2 x_2+\lambda_3 x_3$$

[see Eqs. (13.2) and (13.12)] yields

$$\ln\left(\frac{P(Y{=}1|X{=}0,U{=}Joe)}{1-P(Y{=}1|X{=}0,U{=}Joe)}\right) \;=\; \ln\left(\frac{.7}{1-.7}\right) \;\approx\; -.8473 \;=\; \lambda_0,$$

for $x_1 = 1$, $x_2 = 0$, and $x_3 = 0$ it yields

$$\ln\left(\frac{P(Y{=}1|X{=}1,U{=}Joe)}{1-P(Y{=}1|X{=}1,U{=}Joe)}\right) \;=\; \ln\left(\frac{.8}{1-.8}\right) \;\approx\; 1.3863 \;\approx\; \lambda_0+\lambda_1,$$

for $x_1 = 0$, $x_2 = 1$, and $x_3 = 0$ it yields

$$\ln\left(\frac{P(Y{=}1|X{=}0,U{=}Ann)}{1-P(Y{=}1|X{=}0,U{=}Ann)}\right) \;=\; \ln\left(\frac{.2}{1-.2}\right) \;\approx\; -1.3863 \;\approx\; \lambda_0+\lambda_2,$$

and for $x_1 = 1$, $x_2 = 1$, and $x_3 = 1$ it yields

$$\ln\left(\frac{P(Y{=}1|X{=}1,U{=}Ann)}{1-P(Y{=}1|X{=}1,U{=}Ann)}\right) \;=\; \ln\left(\frac{.4}{1-.4}\right) \;\approx\; -.4055 \;\approx\; \lambda_0+\lambda_1+\lambda_2+\lambda_3.$$

[In R these values are obtained by `qlogis(.7)` to `qlogis(.4)`.] Solving the second equation yields $\lambda_1 \approx .5390$, solving the third equation yields $\lambda_2 \approx -2.2336$, and solving the last one yields $\lambda_3 \approx .4418$.

Chapter 14

Conditional Expectation With Respect to a Conditional-Probability Measure

In chapter 10 we treated the concept of a $\mathscr{C}$-conditional expectation $E(Y|\mathscr{C})$ with respect to a probability measure P on a measurable space $(\Omega,\mathscr{A})$. In this chapter we introduce the concept of a $\mathscr{C}$-conditional expectation $E^B(Y|\mathscr{C})$ of Y with respect to the *conditional-probability measure* P^B on $(\Omega,\mathscr{A})$. A special case with $\mathscr{C}=\sigma(X)$ is the X-conditional expectation of Y with respect to P^B, which is also denoted by $E^B(Y|X)$. If $B=\{Z=z\}$ is the event that a random variable Z on $(\Omega,\mathscr{A},P)$ takes on the value z and $P(Z=z)>0$, then we use the notation $E^{Z=z}(Y|X)$ and call it a version of the *X-conditional expectation of Y with respect to $P^{Z=z}$*.

In empirical applications the conditional expectation $E^{Z=z}(Y|X)$ can be used to describe how the conditional expectation values of Y depend on the values x of X given that Z takes on the value z. The dependency of Y on X described by $E^{Z=z}(Y|X)$ may not only differ for different values z_1 and z_2 of Z, but it may also differ from the dependency described by the X-conditional expectation $E(Y|X)$ of Y with respect to P. If, for instance, X denotes a treatment variable and $Z=sex$ with values m (male) and f (female), then $E^{Z=m}(Y|X)$ and $E^{Z=f}(Y|X)$ refer to the X-conditional expectation of Y for males and females, respectively. In a data sample, these are the conditional expectations estimated using only the y-values and x-values obtained within the male and female subsamples, respectively. In contrast, $E^{X=x}(Y|Z)$ refers to the Z-conditional expectation of Y given treatment x, and this is the conditional expectation estimated in the analysis of experimental or quasi-experimental data using only the y-values and z-values obtained in treatment condition x. If the treatment variable X is dichotomous with values 0 (control) and 1 (treatment), then the difference $g_1(Z):=E^{X=1}(Y|Z)-E^{X=0}(Y|Z)$ is the Z-conditional effect function of X. The values $g_1(z)$ are the effects of X on Y given the value z of Z.

14.1 Introductory Examples

Example 14.1 (Joe and Ann With Random Assignment – continued) Table 14.1 displays the random variables U, X, Y, and, among other things, the conditional expectations $E(Y|X)$ and $E(Y|X,U)$. In this example, the conditional expectation $E(Y|X,U)$ is uniquely defined and satisfies

Table 14.1. Joe and Ann With Random Assignment: Conditional Expectations With Respect to $P^{X=x}$

Outcomes ω				Observables			Conditional expectations							
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$P(X=1\|U)$	$E^{X=0}(Y\|U)$	$E^{X=1}(Y\|U)$	Effect function $g_1(U)$	$P^{X=0}(\{\omega\})$	$P^{X=1}(\{\omega\})$
(*Joe*,	*no*,	−)	.09	*Joe*	0	0	.7	.45	.4	.7	.8	.1	.15	0
(*Joe*,	*no*,	+)	.21	*Joe*	0	1	.7	.45	.4	.7	.8	.1	.35	0
(*Joe*,	*yes*,	−)	.04	*Joe*	1	0	.8	.6	.4	.7	.8	.1	0	.1
(*Joe*,	*yes*,	+)	.16	*Joe*	1	1	.8	.6	.4	.7	.8	.1	0	.4
(*Ann*,	*no*,	−)	.24	*Ann*	0	0	.2	.45	.4	.2	.4	.2	.4	0
(*Ann*,	*no*,	+)	.06	*Ann*	0	1	.2	.45	.4	.2	.4	.2	.1	0
(*Ann*,	*yes*,	−)	.12	*Ann*	1	0	.4	.6	.4	.2	.4	.2	0	.3
(*Ann*,	*yes*,	+)	.08	*Ann*	1	1	.4	.6	.4	.2	.4	.2	0	.2

$$\begin{aligned} E(Y|X,U) &= .2 + .5\cdot 1_{U=Joe} + .2\cdot X - .1\cdot 1_{U=Joe}\cdot X \\ &= (.2+.5\cdot 1_{U=Joe}) + (.2-.1\cdot 1_{U=Joe})\cdot X \\ &= g_0(U) + g_1(U)\cdot X. \end{aligned} \tag{14.1}$$

In this equation, $g_0(U) = .2+.5\cdot 1_{U=Joe}$ is the U-conditional *intercept function* and

$$g_1(U) = .2 - .1\cdot 1_{U=Joe}$$

is the U-conditional *effect function*. The effect function $g_1(U)$ is a random variable on $(\Omega,\mathscr{A},P)$. It is the composition of the person variable U and a function g_1 assigning the person-specific effect to each value u of U. In this example, the treatment effect for $u = Joe$ is $g_1(Joe) = .2-.1\cdot 1 = .1$, and for $u = Ann$ it is $g_1(Ann) = .2-.1\cdot 0 = .2$.

In Theorem 15.3 we show that Equation (14.1) implies

$$g_0(U) = E^{X=0}(Y|U) \tag{14.2}$$

and

$$g_1(U) = E^{X=1}(Y|U) - E^{X=0}(Y|U), \tag{14.3}$$

where $E^{X=x}(Y|U)$, $x = 0,1$, denotes the U-conditional expectation of Y with respect to the conditional-probability measure $P^{X=x}$.

Furthermore, because $E(1_{U=Joe}) = P(U=Joe) = .5$ [see Eq. (6.4)], the expectation of the effect function is

Table 14.2. Joe and Ann With Self-Selection: Conditional Expectations With Respect to $P^{X=x}$

Outcomes ω				Observables			Conditional expectations							
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$P(X=1\|U)$	$E^{X=0}(Y\|U)$	$E^{X=1}(Y\|U)$	$g_1(U)$	$P^{X=0}(\{\omega\})$	$P^{X=1}(\{\omega\})$
(*Joe*,	*no*,	−)	.144	*Joe*	0	0	.7	.60	.04	.7	.8	.1	.24	0
(*Joe*,	*no*,	+)	.336	*Joe*	0	1	.7	.60	.04	.7	.8	.1	.56	0
(*Joe*,	*yes*,	−)	.004	*Joe*	1	0	.8	.42	.04	.7	.8	.1	0	.01
(*Joe*,	*yes*,	+)	.016	*Joe*	1	1	.8	.42	.04	.7	.8	.1	0	.04
(*Ann*,	*no*,	−)	.096	*Ann*	0	0	.2	.60	.76	.2	.4	.2	.16	0
(*Ann*,	*no*,	+)	.024	*Ann*	0	1	.2	.60	.76	.2	.4	.2	.04	0
(*Ann*,	*yes*,	−)	.228	*Ann*	1	0	.4	.42	.76	.2	.4	.2	0	.57
(*Ann*,	*yes*,	+)	.152	*Ann*	1	1	.4	.42	.76	.2	.4	.2	0	.38

$$E[g_1(U)] = E(.2-.1\cdot 1_{U=Joe}) = .2-.1\cdot E(1_{U=Joe}) = .2-.1\cdot .5 = .15. \tag{14.4}$$

For simplicity, this expectation is also called the *average treatment effect*. In this example, $E[g_1(U)]$ is also the slope of X in the equation

$$E(Y|X) = .45+.15\cdot X. \tag{14.5}$$

From a methodological point of view, note that in this example the function $g_1(U)$ and its expectation $E[g_1(U)]$ have a causal interpretation as a U-conditional effect function and the average effect of the treatment. Furthermore, the slope of X in Equation (14.5) is also identical to the average causal effect of the treatment variable X. As shown in Corollary 15.21, this follows from independence of X and U [see the column headed $P(X=1|U)$ in Table 14.1]. In empirical applications, independence of X and U is created by randomly assigning the drawn person to treatment x *with identical probabilities for all persons*. ◁

Example 14.2 (Joe and Ann With Self-Selection – continued) Table 14.2 shows another example with Joe and Ann. In this example, the drawn person is *not randomly assigned* to one of the treatment conditions. Instead, assignment is by self-selection into one of the treatment conditions with the person-specific probabilities displayed in the column headed $P(X=1|U)$. Equation (14.1) still holds for $E(Y|X,U)$. Therefore, the conditional treatment effect for $u = Joe$ is again $g_1(Joe) = .1$, and for $u = Ann$ it is $g_1(Ann) = .2$. Furthermore, because

$P(U{=}Joe) = .5$ still holds, the average treatment effect is again $E[g_1(U)] = .15$. In contrast to Example 14.1, $E[g_1(U)]$ is *not identical* to the slope of X in the equation

$$E(Y|X) = .60 - .18 \cdot X. \tag{14.6}$$

Obviously, now the slope $-.18$ has no causal interpretation. This slope would be extremely misleading if used for the evaluation of the treatment. While the function $g_1(U)$ and its expectation $E[g_1(U)]$ can still be causally interpreted as a U-conditional effect function and the average effect of the treatment, respectively, the slope of X in Equation (14.6) does not have a causal meaning. In this example, X and U are *not* independent because the conditional probabilities $P(X{=}1\,|\,U{=}u)$ *do depend* on the values u of the person variable U. ◁

Example 14.3 (No Treatment for Joe – continued) Table 14.3 displays a third example with Joe and Ann. Now the conditional expectations $E(Y|X)$ and $E(Y|X,U)$ are still well-defined (see Exercises 14-1 to 14-3). However, in this example, there are infinitely many versions of this conditional expectation. In Table 9.2 we already specified a version $V \in \mathscr{E}(Y|X,U)$ with

$$V(\omega) = 0, \quad \text{if } \omega \in \{(Joe, yes, -), (Joe, yes, +)\},$$

and in Example 10.19 we noted that assigning any other real number instead would do as well. For instance, assigning

$$V^*(\omega) = 9, \quad \text{if } \omega \in \{(Joe, yes, -), (Joe, yes, +)\},$$

we define a second version $V^* \in \mathscr{E}(Y|X,U)$, provided, of course, that the other values are assigned as in Table 9.2. In Example 10.19 we also noted that two versions V and V^* of $E(Y|X,U)$ are P-equivalent.

In this example, there are also infinitely many versions of the U-conditional effect function. In contrast to different versions of $E(Y|X,U)$, different versions of the U-conditional effect function are not necessarily P-equivalent. For instance, if we consider the version $V \in \mathscr{E}(Y|X,U)$ specified above, then the associated effect function $g_1(U)$ has the two values

$$g_1(Joe) = 0 - .696 = -.696 \qquad \text{and} \qquad g_1(Ann) = .40 - .20 = .20,$$

whereas considering the version $V^* \in \mathscr{E}(Y|X,U)$ specified above, the associated effect function $g_1^*(U)$ has the two values

$$g_1^*(Joe) = 9 - .696 = 8.304 \qquad \text{and} \qquad g_1^*(Ann) = .40 - .20 = .20.$$

Furthermore, the expectations of different versions of the effect function also differ from each other. The expectation of $g_1(U)$ is

$$E[g_1(U)] = -.696 \cdot .5 + .20 \cdot -5 = -.448,$$

whereas the expectation of $g_1^*(U)$ is

$$E[g_1^*(U)] = 8.304 \cdot .5 + .20 \cdot -5 = 4.05.$$

◁

Table 14.3. No Treatment for Joe: Conditional Expectations With Respect to $P^{X=x}$

Outcomes ω				Observables			Conditional expectations						
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$P(X=1\|U)$	$E^{X=0}(Y\|U)$	$E^{X=1}(Y\|U)$	$p^{X=0}(\{\omega\})$ (rounded)	$p^{X=1}(\{\omega\})$
(*Joe*,	*no*,	−)	.152	*Joe*	0	0	.696	.6	0	.696	9	.245	0
(*Joe*,	*no*,	+)	.348	*Joe*	0	1	.696	.6	0	.696	9	.561	0
(*Joe*,	*yes*,	−)	0	*Joe*	1	0	9	.4	0	.696	9	0	0
(*Joe*,	*yes*,	+)	0	*Joe*	1	1	9	.4	0	.696	9	0	0
(*Ann*,	*no*,	−)	.096	*Ann*	0	0	.2	.6	.76	.2	.4	.155	0
(*Ann*,	*no*,	+)	.024	*Ann*	0	1	.2	.6	.76	.2	.4	.039	0
(*Ann*,	*yes*,	−)	.228	*Ann*	1	0	.4	.4	.76	.2	.4	0	.6
(*Ann*,	*yes*,	+)	.152	*Ann*	1	1	.4	.4	.76	.2	.4	0	.4

Remark 14.4 (Methodological Conclusions) Obviously, the notions effect function, conditional effects, and average effects are crucial for the evaluation of treatments, interventions, and expositions. This does not only apply to U-conditional effects, but also to conditioning on other variables, say Z, such as $Z = \textit{gender}$, $Z = \textit{severity of symptoms}$, $Z = \textit{educational status}$, etc. However, although a conditional expectation $E(Y|X,Z)$ is uniquely defined up to P-equivalence, considering a dichotomous treatment variable X with values 0 and 1 neither guarantees that the effect function $g_1(Z)$ is uniquely defined up to P-equivalence nor that the expectation of $g_1(Z)$ is identical for different versions of the effect function. This suggests that we need to learn more about the effect function and its components $E^{X=x}(Y|U)$ [see Eq. (14.3)]. In more general terms, we need to learn more about a conditional expectation with respect to a conditional-probability measure. ◁

14.2 Assumptions and Definitions

In section 4.2 we considered a probability space $(\Omega, \mathscr{A}, P)$ and an event $B \in \mathscr{A}$ with $P(B) > 0$. According to Theorem 4.23, the function $P^B\colon \mathscr{A} \to [0,1]$ defined by

$$P^B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \forall\, A \in \mathscr{A}, \tag{14.7}$$

is a probability measure on $(\Omega, \mathscr{A})$, called the *B-conditional-probability measure*. We also noted that $(\Omega, \mathscr{A}, P^B)$ is a probability space sharing the measurable space $(\Omega, \mathscr{A})$ with the original probability space $(\Omega, \mathscr{A}, P)$.

In section 9.1 we used the conditional-probability measure P^B in order to introduce the *B-conditional expectation*

$$E(Y|B) \;=\; E^B(Y) \;=\; \int Y \, dP^B, \tag{14.8}$$

assuming $P(B) > 0$ [see Eq. (9.2)] and that the expectation of Y with respect to P^B exists (see Def. 6.1), i. e., that Y is quasi-integrable with respect to P^B (see Def. 3.28). Hence, $E(Y|B)$ denotes the B-conditional expectation of Y and, by definition, it is identical to the expectation $E^B(Y)$ of Y with respect to P^B. If Z is a random variable on $(\Omega, \mathscr{A}, P)$ and $B = \{Z{=}z\} = \{\omega \in \Omega\colon Z(\omega) = z\}$ with $P(Z{=}z) > 0$, then we use the notation $E^{Z=z}(Y)$ and

$$E(Y|Z{=}z) \;=\; E^{Z=z}(Y) \tag{14.9}$$

instead of $E^B(Y)$ as well as $P^{Z=z}$ instead of P^B. Note, however, that $E(Y|Z{=}z)$ is also defined if $P(Z{=}z) = 0$. [For the definition see Eq. (10.27) and for uniqueness see Rem. 10.28.]

14.2.1 Conditional Expectation With Respect to a Conditional-Probability Measure

In this section, we often refer to the following assumptions and notation:

Notation and Assumptions 14.5
$Y\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is a random variable, $B \in \mathscr{A}$ with $P(B) > 0$, $\mathscr{C} \subset \mathscr{A}$ a σ-algebra, and P^B is defined by Equation (14.7)*. Furthermore, Y is nonnegative or such that $E^B(Y)$ is finite.*

Remark 14.6 (Finite Expectation With Respect to P^B) If Y is a random variable with finite expectation $E(Y)$, then $E^B(Y)$ is finite, too (see Exercise 14-4). In contrast, finiteness of $E^B(Y)$ does not imply that $E(Y)$ is finite. ◁

Reading the following definition, note that V, X, Y are random variables on the probability space $(\Omega, \mathscr{A}, P)$ if and only if they are random variables on $(\Omega, \mathscr{A}, P^B)$, provided, of course, that $P(B) > 0$ so that P^B is defined (see Exercise 14-5). Also remember that $\sigma(V) = V^{-1}(\overline{\mathscr{B}})$ denotes the σ-algebra generated by the random variable $V\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ (see section 2.3.2).

Definition 14.7 ($\mathscr{C}$-Conditional Expectation With Respect to P^B)
Let the assumptions 14.5 hold. A random variable $V\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is called a version of the $\mathscr{C}$-conditional expectation of Y with respect to P^B, if the following two conditions hold:

(a) $\sigma(V) \subset \mathscr{C}$.
(b) $E^B(1_C \cdot V) = E^B(1_C \cdot Y), \quad \forall\, C \in \mathscr{C}$.

If V satisfies (a) and (b), then we also use the notation $E^B(Y|\mathscr{C}) := V$.

According to Equation (6.1), condition (b) of this definition is equivalent to

$$\int 1_C \cdot V \, dP^B = \int 1_C \cdot Y dP^B, \quad \forall C \in \mathscr{C}. \tag{14.10}$$

This equation shows more clearly how the measure P^B is involved in the definition of $E^B(Y|\mathscr{C})$.

Remark 14.8 (Existence and Uniqueness) Obviously, the definition of $E^B(Y|\mathscr{C})$ is analog to the definition of a version of a $\mathscr{C}$-conditional expectation $E(Y|\mathscr{C})$ (see Def. 10.2). The only difference is that, instead of referring to P, now we refer to the conditional-probability measure P^B defined by Equation (14.7). Hence, Theorem 10.9 (i) implies that, under the assumptions of Definition 14.7, there exists at least one version $E^B(Y|\mathscr{C})$. Furthermore, Theorem 10.9 (ii) implies that two versions V and V^* of the $\mathscr{C}$-conditional expectation of Y with respect to P^B are P^B-equivalent. In other words, $E^B(Y|\mathscr{C})$ is P^B-unique (see Rem. 10.11). ◁

Remark 14.9 (Set of all Versions of the $\mathscr{C}$-Conditional P^B-Expectation) The notation $E^B(Y|\mathscr{C})$ refers to a *version* of the $\mathscr{C}$-conditional expectation of Y with respect to P^B. In contrast, *the set* of all random variables V on $(\Omega, \mathscr{A}, P)$ that satisfy conditions (a) and (b) of Definition 14.7 is denoted by $\mathscr{E}^B(Y|\mathscr{C})$. Hence, we can write $E^B(Y|\mathscr{C}) \in \mathscr{E}^B(Y|\mathscr{C})$ (see also Exercise 14-6). ◁

Remark 14.10 ($E^B(Y|\mathscr{C})$ is not Necessarily P-Unique) Although $E^B(Y|\mathscr{C})$ is P^B-unique, it is *not necessarily P-unique* (see Rem. 10.11). In section 14.4 we present necessary and sufficient conditions for P-uniqueness of $E^B(Y|\mathscr{C})$, a property that has important implications (see, e. g., section 14.4.4 and Box 14.1). ◁

Remark 14.11 (Properties of $E^B(Y|\mathscr{C})$) Because $E^B(Y|\mathscr{C})$ is a conditional expectation, the properties that have been treated in detail in chapters 10 to 11 analogously also apply to $E^B(Y|\mathscr{C})$. We simply have to exchange the probability measure P by P^B, the expectation $E(\cdot)$ by $E^B(\cdot)$, the variance $Var(\cdot)$ by $Var^B(\cdot)$, and the covariance $Cov(\cdot,\cdot)$ by $Cov^B(\cdot,\cdot)$. ◁

Remark 14.12 (Existence of a Real-Valued Version $E^B(Y|\mathscr{C})$) Box 10.1 (x) immediately yields: If the assumptions 14.5 hold and $E^B(Y)$ is finite, then there is a *real-valued* version $V \in \mathscr{E}^B(Y|\mathscr{C})$. ◁

Remark 14.13 ($\mathscr{C}$-Conditional Probability With Respect to P^B) Let the assumptions 14.5 hold and let $A \in \mathscr{A}$. Then we call

$$P^B(A|\mathscr{C}) := E^B(1_A|\mathscr{C}) \tag{14.11}$$

a version of the $\mathscr{C}$-conditional probability of A with respect to P^B. Correspondingly, $\mathscr{P}^B(A|\mathscr{C})$ denotes the set of all these versions $P^B(A|\mathscr{C})$. ◁

Now we adapt notation and terminology of a $\mathscr{C}$-conditional expectation with respect to a conditional-probability measure to the case in which the σ-algebra $\mathscr{C}$ is generated by a random variable.

Definition 14.14 (X-Conditional Expectation With Respect to P^B)
Let the assumptions 14.5 hold and assume that $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ is a random variable. Then we define:

$$E^B(Y|X) \;:=\; E^B\big(Y\,\big|\,\sigma(X)\big), \tag{14.12}$$

and call it a version of the X-conditional expectation of Y with respect to P^B.

Correspondingly, we use $\mathscr{E}^B(Y|X)$ to denote the set of all versions of the X-conditional expectation of Y with respect to P^B.

Remark 14.15 (X-Conditional Probability of an Event With Respect to P^B) If $A\in\mathscr{A}$, then

$$P^B(A|X) \;:=\; E^B(1_A|X) \tag{14.13}$$

is called *a version of the X-conditional expectation of A with respect to P^B*, and $\mathscr{P}^B(A|X)$ denotes the set of all these versions. ◁

If B is the event $\{Z{=}z\}=\{\omega\in\Omega\colon Z(\omega)=z\}$ that a random variable Z takes on the value z, then we adapt the notation and the terminology correspondingly.

Notation and Assumptions 14.16
$Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ and $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ are random variables and $\mathscr{C}\subset\mathscr{A}$ is a σ-algebra. Furthermore, $z\in\Omega'_Z$ with $P(Z{=}z)>0$ and $P^{Z=z}:=P^B$ with $B=\{Z{=}z\}$, and Y is nonnegative or such that $E^{Z=z}(Y):=E^B(Y)$ is finite.

Remark 14.17 (Conditional Expectation With Respect to $P^{Z=z}$) Let the assumptions 14.16 hold. Then we define

$$E^{Z=z}(Y|\mathscr{C}) \;:=\; E^{\{Z=z\}}(Y|\mathscr{C}) \tag{14.14}$$

and call it a version of the *$\mathscr{C}$-conditional expectation of Y with respect to $P^{Z=z}$*. The measure $P^{Z=z}$ is defined by Equation (14.7) with $B=\{Z{=}z\}$. Correspondingly, $\mathscr{E}^{Z=z}(Y|\mathscr{C})$ denotes the set of all versions of the $\mathscr{C}$-conditional expectation of Y with respect to $P^{Z=z}$. ◁

Remark 14.18 ($\mathscr{C}$-Conditional Probability With Respect to $P^{Z=z}$) Correspondingly, for $A\in\mathscr{A}$ we define $P^{Z=z}(A|\mathscr{C}):=E^{Z=z}(1_A|\mathscr{C})$, a version of the $\mathscr{C}$-conditional probability of the event A with respect to the measure $P^{Z=z}$, and we use $\mathscr{P}^{Z=z}(A|\mathscr{C})$ to denote the family of all versions of the $\mathscr{C}$-conditional probability of the event A with respect to the measure $P^{Z=z}$. ◁

In the next definition we additionally consider a random variable X and use it such that $\sigma(X)$ takes the role of the σ-algebra $\mathscr{C}$.

Notation and Assumptions 14.19
$X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$, $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$, and $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ are random variables. Furthermore, $z\in\Omega'_Z$ with $P(Z{=}z)>0$, and $P^{Z=z}:=P^B$ with $B=\{Z{=}z\}$. Finally, Y is nonnegative or such that $E^{Z=z}(Y)$ is finite.

Under these assumptions and using this notation we define an X-conditional expectation with respect to a $(Z{=}z)$-conditional probability measure as follows:

Definition 14.20 (X-Conditional Expectation With Respect to $P^{Z=z}$)
Let the assumptions 14.19 hold. Then

$$E^{Z=z}(Y\,|\,X)\;:=\;E^{\{Z=z\}}\big(Y\,\big|\,\sigma(X)\big), \tag{14.15}$$

is called a version of the X-conditional expectation of Y with respect to $P^{Z=z}$.

Again note the distinction between a version $E^{Z=z}(Y\,|\,X)$ and $\mathscr{E}^{Z=z}(Y\,|\,X)$, the family of all versions of the X-conditional expectation of Y with respect to $P^{Z=z}$. Of course, what has been said in Remark 14.11 about the properties of a $\mathscr{C}$-conditional expectation with respect to P^B applies to $E^{Z=z}(Y\,|\,X)$ as well.

Remark 14.21 (X-Conditional Probability With Respect to $P^{Z=z}$) Correspondingly, for $A\in\mathscr{A}$ we define $P^{Z=z}(A\,|\,X):=E^{Z=z}(1_A\,|\,X)$, a version of the X-conditional probability of the event A with respect to the measure $P^{Z=z}$, and we use $\mathscr{P}^{Z=z}(A\,|\,X)$ to denote the family of all these versions. ◁

Remark 14.22 (Rules of Computation) The rules of computation for conditional expectations $E(Y\,|\,\mathscr{C})$ with respect to a probability measure P analogously hold for conditional expectation $E^B(Y\,|\,\mathscr{C})$ with respect to a conditional-probability measure P^B. For example, according to Rule (iv) of Box 10.1,

$$E^B[E^B(Y\,|\,\mathscr{C})]\;=\;E^B(Y). \tag{14.16}$$

Similarly, according to Rule (iii) of Box 10.1,

$$E^B(\alpha\cdot Y\,|\,\mathscr{C})\;\underset{P^B}{=}\;\alpha\cdot E^B(Y\,|\,\mathscr{C}),\quad\alpha\in\mathbb{R}. \tag{14.17}$$

We simply have to exchange the notation $E(\cdot)$, which refers to the measure P, by $E^B(\cdot)$ referring to the conditional-probability measure P^B and, of course, exchange P by P^B. ◁

In the following theorem we extend Equation (14.17) showing how to deal with a $\mathscr{C}$-conditional expectation with respect to $P^{Z=z}$ of $f(Z)\cdot Y$.

Theorem 14.23 (Regressand $f(Z)\cdot Y$)
Let the assumptions 14.16 hold. If $f\colon(\Omega'_Z,\mathcal{A}'_Z)\to(\overline{\mathbb{R}},\overline{\mathcal{B}})$ is a measurable function and $f(z)\in\mathbb{R}$, then

$$E^{Z=z}[f(Z)\cdot Y|\mathscr{C}]\underset{P^{Z=z}}{=}f(z)\cdot E^{Z=z}(Y|\mathscr{C}). \tag{14.18}$$

(Proof p. 409)

Remark 14.24 (Two Special Cases) For the constant $Y=1$, Rule (i) of Box 10.1 yields $E^{Z=z}(Y|\mathscr{C})\underset{P^{Z=z}}{=}1$. Therefore, Equation (14.18) implies

$$E^{Z=z}[f(Z)|\mathscr{C}]\underset{P^{Z=z}}{=}f(z) \tag{14.19}$$

Another special case of Equation (14.18) is

$$E^{Z=z}[f(Z)]\;=\;f(z), \tag{14.20}$$

which follows from Remark 10.5 and (14.19) for $\mathscr{C}=\{\Omega,\emptyset\}$. ◁

Remark 14.25 (Two Probability Spaces) There are also some properties of a conditional expectation $E^B(Y|\mathscr{C})$ that are related to the fact that two probability spaces, $(\Omega,\mathcal{A},P)$ and $(\Omega,\mathcal{A},P^B)$, are involved. By definition, a version of the conditional expectation $E^B(Y|\mathscr{C})$ with respect to P^B is a random variable on the probability space $(\Omega,\mathcal{A},P^B)$. Therefore, it is also a random variable on $(\Omega,\mathcal{A},P)$. However, different elements of $\mathscr{E}^B(Y|\mathscr{C})$ are not necessarily P-equivalent; they are necessarily equivalent only with respect to P^B. Hence, if $V,V^*\in\mathscr{E}^B(Y|\mathscr{C})$, then the expectations $E(V)$ and $E(V^*)$ with respect to P may differ from each other, whereas $E^B(V)$ and $E^B(V^*)$ are necessarily identical. These issues are treated in detail in section 14.4.2. ◁

14.2.2 Partial Conditional Expectation

Now we introduce the concept of a *partial conditional expectation* using a factorization of a version $g(X,Z)=E(Y|X,Z)\in\mathscr{E}(Y|X,Z)$. We show how this concept is related to a conditional expectation with respect to a conditional probability measure. In Definition 14.26 we refer to the functions $g_z\colon\Omega'_X\to\overline{\mathbb{R}}$ that, for all $z\in\Omega'_Z$, are defined by

$$g_z(x)\;=\;g(x,z),\quad\forall\,x\in\Omega'_X. \tag{14.21}$$

Referring to the concept of an $(X{=}x,Z{=}z)$-conditional expectation value introduced in Definition 10.33, we can write

$$g_z(x)\;=\;g(x,z)\;=\;E(Y|X{=}x,Z{=}z),\quad\forall\,(x,z)\in\Omega'_X\times\Omega'_Z. \tag{14.22}$$

In Equations (14.21) to (14.23) we do not assume $P(Z{=}z)>0$.

Definition 14.26 (Partial Conditional Expectation)
Let $X\colon (\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$, $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$, and $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ be random variables and assume that Y is nonnegative or with finite expectation $E(Y)$. Furthermore, let $g(X,Z)=E(Y|X,Z)\in\mathscr{E}(Y|X,Z)$ and, for $z\in\Omega'_Z$, let the function g_z, be defined by Equation (14.21)*. Then the function $E(Y|X,Z=z)\colon\Omega\to\overline{\mathbb{R}}$ defined by*

$$E(Y|X,Z=z) := g_z(X) \tag{14.23}$$

is called a version of the partial $(X,Z=z)$-conditional expectation of Y (with respect to P).

To emphasize, for each $z\in\Omega'_Z$, the function $E(Y|X,Z=z)$ denotes the composition of X and g_z. Hence, for each $z\in\Omega'_Z$ it is a random variable on $(\Omega,\mathscr{A},P)$ that is X-measurable (see Lemma 2.52). In Theorem 14.29 we show that $E(Y|X,Z=z)$ is a version of the conditional expectation of Y on X with respect to $P^{Z=z}$, provided that $P(Z=z)>0$.

Remark 14.27 (Partial Conditional Probability) If Y is dichotomous with values 0 and 1, then we also use the notation $E(Y|X,Z=z):=P(Y=1|X,Z=z)$ and call it the partial $(X,Z=z)$-conditional probability of the event $\{Y=1\}$ — or simply of $Y=1$ — (with respect to P). ◁

Remark 14.28 (Discrete Z) Under the assumptions of Definition 14.26, suppose that $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ is discrete and $Z(\Omega)\subset\Omega'_Z$ is finite or countable with $\{z\}\in\mathscr{A}'_Z$ for all $z\in Z(\Omega)$. Then

$$E(Y|X,Z) = \sum_{z\in Z(\Omega)} E(Y|X,Z=z)\cdot 1_{Z=z} \tag{14.24}$$

holds *for the specific version* $E(Y|X,Z)$ that is used in Definition 14.26 (see Exercise 14-7). Furthermore,

$$V \underset{P}{=} \sum_{z\in Z(\Omega)} E(Y|X,Z=z)\cdot 1_{Z=z}, \quad \forall\, V\in\mathscr{E}(Y|X,Z). \tag{14.25}$$

◁

Theorem 14.29 (Relationship Between $E(Y|X,Z=z)$ and $E^{Z=z}(Y|X)$)
Let the assumptions of Definition 14.26 hold and suppose that $z\in\Omega'_Z$ with $P(Z=z)>0$. Then

$$E(Y|X,Z=z)\in\mathscr{E}^{Z=z}(Y|X), \tag{14.26}$$

and therefore

$$E(Y|X,Z=z) \underset{P^{Z=z}}{=} E^{Z=z}(Y|X), \quad \forall\, E^{Z=z}(Y|X)\in\mathscr{E}^{Z=z}(Y|X). \tag{14.27}$$

(Proof p. 409)

Remark 14.30 (An Immediate Implication) If the assumptions of Theorem 14.29 hold and if Z is discrete with $P(Z \in \Omega_0') = 1$ and $P(Z{=}z) > 0$ for all $z \in \Omega_0'$, then

$$V \underset{P}{=} \sum_{z \in \Omega_0'} E^{Z=z}(Y|X) \cdot \mathbf{1}_{Z=z}, \quad \forall V \in \mathscr{E}(Y|X,Z). \tag{14.28}$$

◁

14.2.3 Examples

Example 14.31 (Joe and Ann With Self-Selection–continued) We continue Example 14.2 illustrating how to compute $E^{X=0}(Y|U)$ and $E^{X=1}(Y|U)$, which, in this example, are both uniquely defined. First we compute the probabilities of the elementary events with respect to the measures $P^{X=0}$ and $P^{X=1}$, and then specify $E^{X=0}(Y|U)$ and $E^{X=1}(Y|U)$.

Enumerating the eight elementary events $\{\omega_1\}$ to $\{\omega_8\}$ from top to bottom of the first column of Table 14.2, the probabilities of these elementary events with respect to $P^{X=0}$ can be computed as follows:

$$\begin{aligned} P^{X=0}(\{\omega_1\}) &= P^{X=0}[\{(\mathit{Joe}, \mathit{no}, -)\}] = \frac{P[\{(\mathit{Joe}, \mathit{no}, -)\} \cap \{X{=}0\}]}{P(X{=}0)} \\ &= \frac{P[\{(\mathit{Joe}, \mathit{no}, -)\}]}{P(X{=}0)} = \frac{.144}{.144 + .336 + .096 + .024} = .24. \end{aligned}$$

For the elementary event $\{\omega_2\}$ we obtain

$$\begin{aligned} P^{X=0}(\{\omega_2\}) &= P^{X=0}[\{(\mathit{Joe}, \mathit{no}, +)\}] = \frac{P[\{(\mathit{Joe}, \mathit{no}, +)\} \cap \{X{=}0\}]}{P(X{=}0)} \\ &= \frac{P[\{(\mathit{Joe}, \mathit{no}, +)\}]}{P(X{=}0)} = \frac{.336}{.144 + .336 + .096 + .024} = .56. \end{aligned}$$

For $\{\omega_5\}$ we obtain

$$\begin{aligned} P^{X=0}(\{\omega_5\}) &= P^{X=0}[\{(\mathit{Ann}, \mathit{no}, -)\}] = \frac{P[\{(\mathit{Ann}, \mathit{no}, -)\} \cap \{X{=}0\}]}{P(X{=}0)} \\ &= \frac{P[\{(\mathit{Ann}, \mathit{no}, -)\}]}{P(X{=}0)} = \frac{.096}{.144 + .336 + .096 + .024} = .16, \end{aligned}$$

and for $\{\omega_6\}$,

$$\begin{aligned} P^{X=0}(\{\omega_6\}) &= P^{X=0}[\{(\mathit{Ann}, \mathit{no}, +)\}] = \frac{P[\{(\mathit{Ann}, \mathit{no}, +)\} \cap \{X{=}0\}]}{P(X{=}0)} \\ &= \frac{P[\{(\mathit{Ann}, \mathit{no}, +)\}]}{P(X{=}0)} = \frac{.024}{.144 + .336 + .096 + .024} = .04. \end{aligned}$$

The probabilities of the other four elementary events with respect to $P^{X=0}$ are 0 (see also the last but one column of Table 14.2). The probabilities of the eight elementary events with respect to $P^{X=1}$ are computed analogously (see the last column of Table 14.2).

Now we specify the U-conditional expectation of Y with respect to $P^{X=0}$. Because Y is an indicator variable with values 0 and 1, the conditional expectation $E^{X=0}(Y|U)$ can also be denoted by $P^{X=0}(Y=1|U)$ [see Eq. (14.13)]. It has two different values, one for $U(\omega) = \mathit{Joe}$ and one for $U(\omega) = \mathit{Ann}$. These values can be computed as follows:

$$P^{X=0}(Y=1|U=\mathit{Joe}) = \frac{P^{X=0}(Y=1, U=\mathit{Joe})}{P^{X=0}(U=\mathit{Joe})} = \frac{.56}{.24+.56} = .7.$$

and

$$P^{X=0}(Y=1|U=\mathit{Ann}) = \frac{P^{X=0}(Y=1, U=\mathit{Ann})}{P^{X=0}(U=\mathit{Ann})} = \frac{.04}{.04+.16} = .2.$$

The results of the corresponding computations for $P^{X=1}(Y=1|U)$ are displayed in the last but two column of Table 14.2).

Note that, in this example, there is only one single element of $\mathscr{E}^{X=0}(Y|U)$ and one single element of $\mathscr{E}^{X=1}(Y|U)$. In contrast, in Example 14.32 there is one single element of $\mathscr{E}^{X=0}(Y|U)$, but an infinite number of different elements of $\mathscr{E}^{X=1}(Y|U)$. ◁

Example 14.32 (No Treatment for Joe – continued) We continue Example 14.3. In this example, $E^{X=0}(Y|U)$ is uniquely defined (see Exercise 14-8). In contrast, this is not true for $E^{X=1}(Y|U)$. The probabilities of the elementary events with respect to the measures $P^{X=x}$ are computed analogously to Example 14.31. The results are displayed in the last two columns of Table 14.3.

Now we specify a version of the U-conditional expectation of Y with respect to the measure $P^{X=1}$. Because Y is an indicator variable with values 0 and 1, a version of this conditional expectation can also be denoted by $P^{X=1}(Y=1|U)$ [see Eq. (14.13)]. It has two different values, one for $U(\omega) = \mathit{Joe}$ and one for $U(\omega) = \mathit{Ann}$. The latter is

$$P^{X=1}(Y=1|U=\mathit{Ann}) = \frac{P^{X=1}(Y=1, U=\mathit{Ann})}{P^{X=1}(U=\mathit{Ann})} = \frac{.4}{.6+.4} = .4.$$

In contrast, the conditional probability $P^{X=1}(Y=1|U=\mathit{Joe})$ is undefined, because $P^{X=1}(U=\mathit{Joe}) = 0$. Therefore, we can choose any real number as the value of $E^{X=1}(Y|U)$ for $\omega \in \{U=\mathit{Joe}\} = \{\omega_1, \ldots, \omega_4\}$. For example,

$$V_1(\omega) = \begin{cases} 9, & \text{if } U(\omega) = \mathit{Joe}, \\ .4, & \text{if } U(\omega) = \mathit{Ann}, \end{cases}$$

[see Eq. (14.38)] defines a first element of $\mathscr{E}^{X=1}(Y|U)$ (see Table 14.3), and

$$V_1^*(\omega) = \begin{cases} 0, & \text{if } U(\omega) = \mathit{Joe}, \\ .4, & \text{if } U(\omega) = \mathit{Ann}, \end{cases}$$

is a second element of the set $\mathscr{E}^{X=1}(Y|U)$. Obviously, V_1 and V_1^* are $P^{X=1}$-equivalent, because $P^{X=1}(A_1) = 0$, where

$$A_1 := \{\omega \in \Omega\colon V_1(\omega) \neq V_1^*(\omega)\} = \{\omega_1, \ldots, \omega_4\}$$
$$= \{(\textit{Joe}, \textit{no}, -), (\textit{Joe}, \textit{no}, +), (\textit{Joe}, \textit{yes}, -), (\textit{Joe}, \textit{yes}, +)\}.$$

This probability can be computed by

$$P^{X=1}(A_1) = \sum_{\omega \in A_1} P^{X=1}(\{\omega\}) = 0+0+0+0 = 0$$

(see the last column of Table 14.3). In contrast,

$$P(A_1) = \sum_{\omega \in A_1} P(\{\omega\}) = .152+.348+0+0 = .5.$$

Hence, the two versions V_1 and V_1^* of the U-conditional $P^{X=1}$-expectation of Y are $P^{X=1}$-equivalent, i. e., they are equivalent with respect to the measure $P^{X=1}$. However, the two versions are *not equivalent* with respect to the measure P. (This issue will be treated in more detail in section 14.4.)

Note that the values of the conditional expectations $E^{X=x}(Y|U)$ are defined for *all* elements $\omega \in \Omega$ and that these conditional expectations are random variables on all three probability spaces $(\Omega, \mathscr{A}, P)$, $(\Omega, \mathscr{A}, P^{X=0})$, and $(\Omega, \mathscr{A}, P^{X=1})$. Furthermore, the values of each of the conditional expectations $E^{X=x}(Y|U)$, $x = 0, 1$, only depend on the drawn person. This illustrates that they are measurable with respect to U [see Def. 14.7 (b)]. ◁

14.3 Factorization

Let the assumptions 14.5 hold and let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable. Because $E^B(Y|X)$ is measurable with respect to X, Lemma 2.52 implies that there is a measurable function $g_B\colon (\Omega'_X, \mathscr{A}'_X) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ such that

$$E^B(Y|X) = g_B(X) \tag{14.29}$$

is the composition of X and g_B. A function g_B satisfying Equation (14.29) is called a *factorization* of the version $E^B(Y|X) \in \mathscr{E}^B(Y|X)$ (see also section 10.4.4).

14.3.1 Conditional Expectation Value With Respect to P^B

The values of a factorization g_B of $E^B(Y|X)$ are called $(X{=}x)$*-conditional expectation values* with respect to P^B and they are denoted by

$$E^B(Y|X{=}x) := g_B(x), \quad \forall\, x \in \Omega'_X, \tag{14.30}$$

(see Def. 10.33). Instead of $E^B(1_A|X{=}x)$ we also use the notation $P^B(A|X{=}x)$, provided that $A \in \mathscr{A}$.

Correspondingly, under the assumptions 14.5,

$$E^{Z=z}(Y|X{=}x) := g_{Z=z}(x), \quad \forall\, x \in \Omega'_X \tag{14.31}$$

and $P^{Z=z}(A|X=x) := E^{Z=z}(1_A|X=x)$ for $A \in \mathscr{A}$. Note that $g_{Z=z}$ is not necessarily identical to the function g_z defined by Equation (14.21) (see Rem. 14.33 for more details).

Remark 14.33 (Relationship Between the Functions g_z and $g_{Z=z}$) The relationship between the function g_z defined by Equation (14.21) and the function $g_{Z=z}$ defined by Equation (14.31) is as follows:

(i) $g_z \underset{P_X^{Z=z}}{=} g_{Z=z}$ if $P(Z=z) > 0$

(ii) $g_z = g_{Z=z}$ if $z \in \Omega'_Z$ with $P(Z=z, X=x) > 0$ for all $x \in \Omega'_X$

(see Exercise 14-9). If $P(Z=z) = 0$, then $g_{Z=z}$ is not defined. ◁

Remark 14.34 (Relationship Between Factorizations) Suppose that the assumptions 14.19 hold, where Y is nonnegative or with finite expectation $E(Y)$. Then Theorem 14.29 implies

$$\begin{aligned} E^{Z=z}(Y|X=x) &= E(Y|X=x, Z=z) \\ &= E(Y|X=x, 1_{Z=z}=1), \quad \text{for } P_X^{Z=z}\text{-a.a. } x \in \Omega'_X, \end{aligned} \tag{14.32}$$

where $E^{Z=z}(Y|X=x)$ and $E(Y|X=x, Z=z)$ are the conditional expectation values defined by (14.30), (14.31), and (10.27), respectively (see Exercise 14-10). Note that $B = \{1_B = 1\}$. Therefore, for $B \in \mathscr{A}$ with $P(B) > 0$, $Z = 1_B$, and $z = 1$, the first of these two equations yields

$$E^B(Y|X=x) = E(Y|X=x, 1_B=1), \quad \text{for } P_X^B\text{-a.a. } x \in \Omega'_X. \tag{14.33}$$

◁

14.3.2 Uniqueness of Factorizations

Remark 14.35 (Uniqueness of Factorizations) For a fixed version $E^B(Y|X)$, the factorization g_B of $E^B(Y|X)$ is uniquely defined, provided that Ω'_X is identical to the image $X(\Omega) = \{\omega \in \Omega\colon X(\omega) \in \Omega'_X\}$. If $\Omega'_X \neq X(\Omega)$, then there can be different factorizations of a single version $E^B(Y|X)$ (see Rem. 10.28).

If g_B and g_B^* are factorizations of two versions V and V^* of $\mathscr{E}^B(Y|X)$, respectively, then, according to Corollary 10.29,

$$g_B \underset{P_X^B}{=} g_B^*, \tag{14.34}$$

i. e., g_B and g_B^* are P_X^B-equivalent, where P_X^B denotes the probability measure on $(\Omega'_X, \mathscr{A}'_X)$ defined by

$$P_X^B(A') = P^B(X \in A'), \quad \forall\ A' \in \mathscr{A}'_X. \tag{14.35}$$

Hence, because $P^B(X \in A')$ is just another notation for $P^B[X^{-1}(A')]$, P_X^B is the image measure of P^B under X (see Th. 2.78 and Def. 2.79). ◁

Remark 14.36 (Family of Factorizations) Note the distinction between (a) a *factorization*, i. e., a random variable g_B on the probability space $(\Omega'_X, \mathscr{A}'_X, P^B_X)$, which, for a version $E^B(Y|X) \in \mathscr{E}^B(Y|X)$ satisfies Equation (14.29), and (b) the *family of factorizations*, i. e., the family of random variables g_B that satisfy

$$V \underset{P^B}{=} g_B(X), \quad \forall\, V \in \mathscr{E}^B(Y|X)\,. \tag{14.36}$$

◁

Remark 14.37 (Values of the X-Conditional Expectation With Respect to P^B) If $P(B) > 0$ and $E^B(Y|X) \in \mathscr{E}^B(Y|X)$, then, for all $x \in \Omega'_X$,

$$E^B(Y|X)(\omega) \;=\; E^B(Y|X{=}x), \quad \forall\, \omega \in \{X{=}x\} \tag{14.37}$$

(see Rem. 10.37). This also implies that the value of $E^B(Y|X)$ is constant on all sets $\{X{=}x\}$, $x \in \Omega'_X$. Correspondingly, if $P(Z{=}z) > 0$, then, for all $x \in \Omega'_X$,

$$E^{Z=z}(Y|X)(\omega) \;=\; E^{Z=z}(Y|X{=}x), \quad \forall\, \omega \in \{X{=}x\}. \tag{14.38}$$

In other words, whenever the random variable X takes on the value x, and this is the case if $\omega \in \{X{=}x\}$, then the random variable $E^{Z=z}(Y|X)$ takes on the value $E^{Z=z}(Y|X{=}x)$. Note that Equation (14.38) also holds if Ω is finite or countable and some $\omega \in \{X{=}x\}$ have probability $P^{Z=z}(\{\omega\}) = 0$. [As an example, consider the values of $E^{X=1}(Y|U)$ for $\omega \in \{(\textit{Joe}, \textit{yes}, -), (\textit{Joe}, \textit{yes}, +)\}$ in Example 14.32.] ◁

14.4 Uniqueness

14.4.1 A Necessary and Sufficient Condition of Uniqueness

In Theorem 10.17 we presented a necessary and sufficient condition for uniqueness of a conditional expectation. In the following corollary we translate this result to a conditional expectation with respect to P^B.

Corollary 14.38 (Uniqueness of $E^B(Y|\mathscr{C})$)
Let the assumptions 14.5 hold, let $\mathscr{E}$ be a finite or countable partition of Ω, and assume $\mathscr{C} = \sigma(\mathscr{E})$. Then $V = V^*$ for all $V, V^* \in \mathscr{E}^B(Y|\mathscr{C})$ if and only if

$$P^B(A) \;>\; 0, \quad \forall\, A \in \mathscr{E}. \tag{14.39}$$

Remark 14.39 (Values of $E^B(Y|\mathscr{C})$) In other words, under the assumptions of Corollary 14.38, the conditional expectation $E^B(Y|\mathscr{C})$ is uniquely defined if and only if (14.39) holds. Furthermore, if (14.39) holds, then

$$E^B(Y|\mathscr{C}) \;=\; \sum_{A \in \mathscr{E}} E^B(Y|A) \cdot \mathbb{1}_A \tag{14.40}$$

[see Eq. (10.14)] and

$$\forall A \in \mathscr{E}: \quad E^B(Y|\mathscr{C})(\omega) = E^B(Y|A), \quad \text{if } \omega \in A \tag{14.41}$$

[see Eq. (14.41)]. The last equation shows that $E^B(Y|\mathscr{C})$ describes how the conditional expectation values $E^B(Y|A)$ depend on the events $A \in \mathscr{E}$. ◁

The corresponding result for the X-conditional expectation $E^B(Y|X)$ of Y with respect to P^B is as follows:

Corollary 14.40 (Uniqueness of $E^B(Y|X)$)
Let the assumptions 14.5 hold. Furthermore, let $X\colon (\Omega,\mathscr{A},P) \to (\Omega'_X,\mathscr{A}'_X)$ be a random variable such that Ω'_X is countable, $X(\Omega) = \Omega'_X$, and $\mathscr{A}'_X = \mathscr{P}(\Omega'_X)$. Then $V = V^*$ for all $V, V^* \in \mathscr{E}^B(Y|X)$ if and only if

$$P^B(X{=}x) > 0, \quad \forall\, x \in X(\Omega). \tag{14.42}$$

Remark 14.41 (Values of $E^B(Y|X)$) Hence, under the assumptions of Corollary 14.40, $E^B(Y|X)$ is uniquely defined if and only if (14.42) holds. And, if (14.42) holds, then

$$E^B(Y|X) = \sum_{x \in X(\Omega)} E^B(Y|X{=}x) \cdot 1_{X=x} \tag{14.43}$$

and

$$\forall x \in X(\Omega): \quad E^B(Y|X)(\omega) = E^B(Y|X{=}x), \quad \text{if } \omega \in \{X{=}x\}. \tag{14.44}$$

This equation shows that the conditional expectation $E^B(Y|X)$ describes how the conditional expectation values $E^B(Y|X{=}x)$ depend on the values x of X. ◁

Example 14.42 (Joe and Ann) Applying Corollary 14.40 to the introductory Examples 14.1 and 14.2 yields that the conditional expectations $E^{X=0}(Y|U)$ and $E^{X=1}(Y|U)$ are uniquely defined. Furthermore, in these two examples, the conditional expectation values $E^{X=x}(Y|U{=}u)$ are also uniquely defined, which follows from the fact that $P^{X=x}(U{=}u) > 0$ for all pairs (x,u) of values of X and U. In contrast, in Example 14.3, only $E^{X=0}(Y|U)$ is uniquely defined, but $E^{X=1}(Y|U)$ is not. In fact, in this example, $E^{X=1}(Y|U)$ is even not P-unique, although it is P^B-unique. This issue is dealt with in section 14.4.2. ◁

14.4.2 Uniqueness w.r.t. P and Other Probability Measures

A conditional expectation $E^B(Y|\mathscr{C})$ is always P^B-unique (see Rem. 14.8). However, we may also ask if $E^B(Y|\mathscr{C})$ is Q-unique, where Q is *any* probability measure on $(\Omega,\mathscr{A})$. This includes $Q = P$, but also $Q = P^C$, where $C \in \mathscr{A}$ and $C \neq B$ (see Remarks 5.13 and 5.16). If $E^B(Y|\mathscr{C})$ is Q-unique, then two versions $V, V^* \in \mathscr{E}^B(Y|\mathscr{C})$ do not only have identical distributions, expectations, variances, etc. with respect to P^B, but also with respect to Q (see Cors. 5.20 and 6.17). The following example shows why this is of interest.

Example 14.43 (Pre-Post Design) Suppose that X is an indicator variable with values 0 (control) and 1 (treatment), Y represents *life satisfaction after treatment*, and Z *life satisfaction before treatment*. Then P-uniqueness of $E^{X=x}(Y|Z)$ is crucial if we consider

$$E(Y|X,Z) \underset{P}{=} g_0(Z) + g_1(Z)\cdot X$$

with the Z-conditional-effect function

$$g_1(Z) \underset{P}{=} E^{X=1}(Y|Z) - E^{X=0}(Y|Z),$$

or its expectation

$$E[g_1(Z)] = E\big[E^{X=1}(Y|Z)\big] - E\big[E^{X=0}(Y|Z)\big].$$

Furthermore, for $x=0$ and $x=1$, we may also consider the $(X{=}x)$-conditional expectation values of $g_1(Z)$

$$E[g_1(Z)\,|\,X{=}x] = E\big[E^{X=1}(Y|Z)\,\big|\,X{=}x\big] - E\big[E^{X=0}(Y|Z)\,\big|\,X{=}x\big],$$

the *average effect of X on Y* given $x=0$ (control) given $x=1$ (treatment), respectively. Considering $E[g_1(Z)|X{=}x]$, where $x{=}0$ or $x{=}1$, it is crucial that $E^{X=0}(Y|Z)$ and $E^{X=1}(Y|Z)$ are unique with respect to the measure $P^{X=x}$ for the following reason: For $x=0$ or $x=1$, if $P(X{=}x)>0$ and $E^{X=1}(Y|Z)$ is $P^{X=x}$-unique, then the conditional expectation value $E\big[E^{X=1}(Y|Z)\,\big|\,X{=}x\big]$ is identical for different versions $E^{X=1}(Y|Z)\in\mathscr{E}^{X=1}(Y|Z)$. Correspondingly, for $x=0$ or $x=1$, if $P(X{=}x)>0$ and $E^{X=0}(Y|Z)$ is $P^{X=x}$-unique, then $E\big[E^{X=0}(Y|Z)\,\big|\,X{=}x\big]$ is identical for different versions $E^{X=0}(Y|Z)\in\mathscr{E}^{X=0}(Y|Z)$ (see Exercise 14-11). ◁

14.4.3 Necessary and Sufficient Conditions of P-Uniqueness

Now we present conditions that are equivalent to P-uniqueness of $E^B(Y|\mathscr{C})$. Note that, in this theorem, we do not refer to the expectation $E^B(V)$ of V with respect to the measure P^B, but to the expectation $E(V)$ of V with respect to P.

Theorem 14.44 (Conditions Equivalent to P-Uniqueness of $E^B(Y|\mathscr{C})$)
Let the assumptions 14.5 hold. Then the following propositions are equivalent to each other:

(a) $E^B(Y|\mathscr{C})$ *is P-unique.*

(b) $P \underset{\mathscr{C}}{\ll} P^B$.

(c) $P(B|\mathscr{C}) \underset{P}{\geq} 0$.

If there is a version $V\in\mathscr{E}^B(Y|\mathscr{C})$ *such that* $E(V)$ *is finite, then (a) to (c) are also equivalent to*

(d) $\forall\, V,V^*\in\mathscr{E}^B(Y|\mathscr{C})\colon\ E(V)=E(V^*)$.

(Proof p. 410)

Remark 14.45 (Sufficient Conditions for Finiteness of $E(V)$) Remember, the expectation of a random variable Y exists if $\int Y^+ dP$ or $\int Y^- dP$ are finite (see Def. 3.28). Hence, the expectation $E(V)$ of a version $V \in \mathscr{E}^B(Y|\mathscr{C})$ exists and is finite, e. g., if one of the following conditions holds:

(a) $\mathscr{C}$ is a finite set and $E^B(Y)$ is finite (see Exercise 14-12),

(b) $E^B(Y|\mathscr{C})$ has only a finite number of real values (see Rem. 6.5),

(c) $E^B(Y|\mathscr{C})$ is P-almost surely bounded on both sides, i. e.,
$\exists\, \alpha \in \mathbb{R} : -\alpha \underset{P}{\leq} E^B(Y|\mathscr{C}) \underset{P}{\leq} \alpha$ [see Eq. (3.50)], or

(d) Y is P-almost surely bounded on both sides, i. e.,
$\exists\, \alpha \in \mathbb{R} : -\alpha \underset{P}{\leq} Y \underset{P}{\leq} \alpha$ [see Box 10.3 (vi) (viii), and (b)].

A special case of (d) is $0 \underset{P}{\leq} Y \underset{P}{\leq} \alpha$, for $0 < \alpha \in \mathbb{R}$. Another one is $Y = 1_A$, if $A \in \mathscr{A}$. ◁

In the following corollary we translate Theorem 14.44 to the special case of an X-conditional expectation $E^B(Y|X)$ with respect to P^B and apply Lemma 5.25.

Corollary 14.46 (P-Uniqueness of $E^B(Y|X)$)
Let the assumptions 14.5 hold and let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable. Then the following propositions are equivalent to each other:

(a) $E^B(Y|X)$ is P-unique.

(b) $P \underset{\sigma(X)}{\ll} P^B$.

(c) $P(B|X) \underset{P}{\geqslant} 0$.

(d) $P_X \underset{\mathscr{A}'_X}{\ll} P^B_X$.

If there is a version $V \in \mathscr{E}^B(Y|X)$ such that $E(V)$ is finite, then (a) to (d) are also equivalent to

(e) $\forall\, V, V^* \in \mathscr{E}^B(Y|X)\colon\ E(V) = E(V^*)$.

Remark 14.47 (Absolute Continuity if X is Discrete) Under the assumptions of Corollary 14.46: If X is discrete and $\{x\} \in \mathscr{A}'_X$ for all $x \in \Omega'_X$, then $P_X \underset{\mathscr{A}'_X}{\ll} P^B_X$ [see 14.46 (d)] is equivalent to

$$\forall\, x \in \Omega'_X : \quad P^B(X{=}x) = 0 \quad \Rightarrow \quad P(X{=}x) = 0 \tag{14.45}$$

(see Exercise 14-13). ◁

Example 14.48 (Joe and Ann With Random Assignment – continued) In the example displayed in Table 14.1 (p. 384), $P^{X=x}(U{=}u) > 0$ for all pairs of values of X and U. This implies that the conditional expectations $E^{X=x}(Y|U)$ are uniquely defined for both values of X (see Cor. 14.40), which implies that they are P-unique. Furthermore, the expectations

$$E[E^{X=x}(Y|U)] = \sum_u E(Y|X{=}x, U{=}u) \cdot P(U{=}u), \quad x = 0, 1, \tag{14.46}$$

are finite. According to Remark 14.45 (b), this follows from the fact that the conditional expectation values $E(Y|X=x,U=u) = E^{X=x}(Y|U=u)$ are finite. Finally, $E(V) = E(V^*) = E[E^{X=x}(Y|U)]$, for all $V, V^* \in \mathscr{E}^{X=x}(Y|U)$ [see Cor. 14.46 (e)]. ◁

Example 14.49 (No Treatment for Joe – continued) Continuing Example 14.32, consider the event

$$\{X=1\} = \big\{(Joe, yes, -), (Joe, yes, -), (Ann, yes, -), (Ann, yes, +)\big\},$$

that the *drawn person is treated* and the event

$$\{U=Joe\} = \big\{(Joe, no, -), (Joe, no, +), (Joe, yes, -), (Joe, yes, +)\big\},$$

that *Joe is drawn*. In this example, we already computed the $\{X=1\}$-conditional probability $P^{X=1}(U=Joe) = 0$ and the (unconditional) probability $P(U=Joe) = .50$. Because $\{U=Joe\} \in \sigma(U)$, in this example, it is not true that $P^{X=1}(A) = 0$ implies $P(A) = 0$ for all $A \in \sigma(U)$. Therefore, $P \ll^{\sigma(U)} P^{X=1}$ does not hold (see Def. 3.70). Hence, Corollary 14.46 implies that the conditional expectation $E^{X=1}(Y|U)$ is not P-unique.

In Table 14.3 (p. 387), the values of $E^{X=1}(Y|U)$ are not uniquely defined for all four $\omega \in \{U=Joe\}$, because, instead of 9, we could have assigned *any* real number to these four possible outcomes ω. Because $P^{X=1}(U=Joe) = 0$, the conditional expectation $E^{X=1}(Y|U)$ is well-defined. It is $P^{X=1}$-unique. However, because $E^{X=1}(Y|U)$ is not P-unique, in this example, $E(V_1) = E(V_1^*)$ does *not* hold for all $V_1, V_1^* \in \mathscr{E}^{X=1}(Y|U)$. This has already been illustrated in Example 14.3. ◁

14.4.4 Properties Related to P-Uniqueness

Box 14.1 summarizes some important properties related to P-uniqueness (for proofs see Exercise 14-14), some of which have already been treated and illustrated in section 14.4.3. In the following remarks we comment some of the implications of P-uniqueness.

Remark 14.50 (Implications of P-Uniqueness) Suppose that Y is nonnegative or with finite expectation $E(Y)$, that $E^B(Y|\mathscr{C})$ is P-unique, and $C \in \mathscr{A}$ with $P(C) > 0$. Then, according to property (v) of Box 14.1, $E^B(Y|\mathscr{C})$ is also P^C-unique, and according to property (vi) of Box 14.1, the distributions P_V^C with respect to P^C of all versions $V \in \mathscr{E}^B(Y|\mathscr{C})$ are identical (see Cor. 5.20). This implies, e. g., that the expectation $E^C[E^B(Y|\mathscr{C})]$ of $E^B(Y|\mathscr{C})$ with respect to the conditional-probability measure P^C is identical for all versions $V \in \mathscr{E}^B(Y|\mathscr{C})$ [see Box 14.1 (vii)]. The same applies to its variance $Var^C[E^B(Y|\mathscr{C})]$ [see Rem. 6.27 and Box 6.2 (v)] as well as to its covariance $Cov^C[E^B(Y|\mathscr{C}), W]$ with another random variable $W\colon (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$ [see Box 7.1 (x)], provided that this variance and this covariance with respect to P^C exist. ◁

Remark 14.51 (The Special Case $C = \Omega$) A special case is $C = \Omega$. Hence, under P-uniqueness of $E^B(Y|\mathscr{C})$, the following equations hold for all $V, V^* \in \mathscr{E}^B(Y|\mathscr{C})$:

Box 14.1 P**-Uniqueness of** $E^B(Y|\mathscr{C})$

Let $(\Omega,\mathscr{A},P)$ be a probability space, let $\mathscr{C},\mathscr{D}\subset\mathscr{A}$ be σ-algebras, let $B,C\in\mathscr{A}$ with $P(B)$, $P(C)>0$, and let $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$ be a random variable that is nonnegative or with finite expectation $E(Y)$. Then:

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;:\Leftrightarrow\; \forall\, V,V^*\in\mathscr{E}^B(Y|\mathscr{C})\colon V\underset{P}{=}V^* \tag{i}$$

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;\Leftrightarrow\; P\underset{\mathscr{C}}{\ll}P^B \tag{ii}$$

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;\Leftrightarrow\; P(B\,|\,\mathscr{C})\underset{P}{>}0 \tag{iii}$$

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;\Rightarrow\; E^B(Y|\mathscr{D}) \text{ is } P\text{-unique}, \quad \text{if } \mathscr{D}\subset\mathscr{C} \tag{iv}$$

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;\Rightarrow\; E^B(Y|\mathscr{C}) \text{ is } P^C\text{-unique} \tag{v}$$

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;\Rightarrow\; \forall\, V,V^*\in\mathscr{E}^B(Y|\mathscr{C})\colon P^C_V=P^C_{V^*} \tag{vi}$$

$$E^B(Y|\mathscr{C}) \text{ is } P\text{-unique} \;\Rightarrow\; \forall\, V,V^*\in\mathscr{E}^B(Y|\mathscr{C})\colon E^C(V)=E^C(V^*). \tag{vii}$$

If $\alpha,\beta\in\mathbb{R}$ and $E^B(Y|\mathscr{C})$ or $E^C(Y|\mathscr{C})$ is real-valued, then

$$E^B(Y|\mathscr{C}),\, E^C(Y|\mathscr{C}) \text{ are } P\text{-unique} \;\Rightarrow\; \alpha\,E^B(Y|\mathscr{C})+\beta\,E^C(Y|\mathscr{C}) \text{ is } P\text{-unique}. \tag{viii}$$

If $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ is a random variable, then:

$$\forall\, x\in\Omega'_X\colon\; P^B(X{=}x)>0 \;\Rightarrow\; E^B(Y|X) \text{ is } P\text{-unique}. \tag{ix}$$

If g_B is a factorization of $E^B(Y|X)$, then

$$E^B(Y|X) \text{ is } P\text{-unique} \;\Leftrightarrow\; g_B \text{ is } P_X\text{-unique}. \tag{x}$$

$$P_V \;=\; P_{V^*}, \tag{14.47}$$

$$E(V) \;=\; E(V^*), \tag{14.48}$$

$$Var(V) \;=\; Var(V^*), \tag{14.49}$$

$$Cov(V,W) \;=\; Cov(V^*,W), \tag{14.50}$$

provided that these expectations, variances, and covariances with respect to the measure P exist. ◁

Remark 14.52 (Another Implication of P**-Uniqueness)** Let the assumptions 14.5 hold. Furthermore, let $\mathscr{D}\subset\mathscr{A}$ be a σ-algebra, $\mathscr{C}\subset\mathscr{D}$, $C\in\mathscr{A}$ with $P(C)>0$, and suppose that $E^B(Y|\mathscr{C})$ is P^C-unique. Then

$$E^C[E^B(Y|\mathscr{C})\,|\,\mathscr{D}]\underset{P^C}{=}E^B(Y|\mathscr{C}), \tag{14.51}$$

which immediately follows from Rule (vii) of Box 10.1, because, by definition, $E^B(Y|\mathscr{C})$ is $\mathscr{C}$-measurable, and because we assume $\mathscr{C}\subset\mathscr{D}$. For the special case $C=\Omega$, this yields

$$E[E^B(Y|\mathscr{C})|\mathscr{D}] \underset{P}{=} E^B(Y|\mathscr{C}). \tag{14.52}$$

◁

Remark 14.53 (Expectation of a Linear Combination) An immediate implication of Box 14.1 (viii) is:

$$E\big[\alpha\cdot E^B(Y|\mathscr{C}) + \beta\cdot E^C(Y|\mathscr{C})\big] = \alpha\cdot E\big[E^B(Y|\mathscr{C})\big] + \beta\cdot E\big[E^C(Y|\mathscr{C})\big], \tag{14.53}$$

provided that $E^B(Y|\mathscr{C})$ and $E^C(Y|\mathscr{C})$ are P-unique, that $E^B(Y|\mathscr{C})$ or $E^C(Y|\mathscr{C})$ is real-valued, and the expectation $E\big[E^B(Y|\mathscr{C})\big]$ or $E\big[E^C(Y|\mathscr{C})\big]$ is finite. Under these assumptions, for all real-valued versions $V_B \in \mathscr{E}^B(Y|\mathscr{C})$ and $V_C \in \mathscr{E}^C(Y|\mathscr{C})$,

$$E(\alpha\cdot V_B + \beta\cdot V_C) = \alpha\cdot E(V_B) + \beta\cdot E(V_C).$$

◁

Example 14.54 (Joe and Ann With Self-Selection – continued) Consider again Table 14.2 (p. 385). In this example, the U-conditional-effect function $g_1(U) = E^{X=1}(Y|U) - E^{X=0}(Y|U)$ is a uniquely defined random variable on $(\Omega, \mathscr{A}, P)$, because $E^{X=0}(Y|U)$ as well as $E^{X=1}(Y|U)$ are uniquely defined, i. e., each of the sets $\mathscr{E}^{X=0}(Y|U)$ and $\mathscr{E}^{X=1}(Y|U)$ has only one single element. This implies that the average treatment effect

$$E[g_1(U)] = E\big[E^{X=1}(Y|U) - E^{X=0}(Y|U)\big]$$

is uniquely defined as well. Note that the assumptions of Box 14.1 (viii) are less restrictive, because they allow that each of the sets $\mathscr{E}^{X=0}(Y|U)$ and $\mathscr{E}^{X=1}(Y|U)$ has more than one element. The requirement of Box 14.1 (viii) is not uniqueness but only P-uniqueness.

◁

Example 14.55 (No Treatment for Joe – continued) In Table 14.3 (p. 387) the set $\mathscr{E}^{X=0}(Y|U)$ has only one single element. However, $\mathscr{E}^{X=1}(Y|U)$ has infinitely many elements: Replacing the value 9 by any other real number yields a new element of $\mathscr{E}^{X=1}(Y|U)$. More important, it is *not true* that all these elements are pairwise P-equivalent. Hence, in this example, the U-conditional-effect function $g_1(U) = E^{X=1}(Y|U) - E^{X=0}(Y|U)$ is *not P-unique*. In other words, it *not true* that all elements of the set

$$\big\{V_1 - V_0 \colon V_0 \in \mathscr{E}^{X=0}(Y|U),\ V_1 \in \mathscr{E}^{X=1}(Y|U)\big\}$$

are pairwise P-equivalent. Therefore

$$E(V_1 - V_0) = E(V_1^* - V_0^*), \quad \forall\, V_0, V_0^* \in \mathscr{E}^{X=0}(Y|U) \text{ and } \forall\, V_1, V_1^* \in \mathscr{E}^{X=1}(Y|U)$$

does not hold (see Rem. 14.53), and this means that there is no average treatment effect in this example. ◁

In the following corollary we extend Theorem 14.29, adding another assumption. Remember, if assumptions 14.19 hold, then according to Theorem 14.29,

$$E^{Z=z}(Y|X) \underset{P^{Z=z}}{=} E(Y|X, Z=z), \tag{14.54}$$

referring to the partial conditional expectation $E(Y|X, Z=z)$ [see Def. 14.26 and Eq. (14.32)].

Corollary 14.56 (Implications of P^C-Uniqueness of $E^{Z=z}(Y|X)$)
Let the assumptions 14.19 hold, where Y is nonnegative of with finite expectation $E(Y)$, and let $C \in \mathscr{A}$ with $P(C) > 0$. If $E^{Z=z}(Y|X)$ is P^C-unique, then

$$E^{Z=z}(Y|X) \underset{P^C}{=} E(Y|X, Z=z), \tag{14.55}$$

and

$$E^{Z=z}(Y|X=x) = E(Y|X=x, Z=z), \quad \text{for } P_X^C\text{-a.a. } x \in \Omega'_X. \tag{14.56}$$

(Proof p. 411)

Remark 14.57 (Implications of P-Uniqueness of $E^{Z=z}(Y|X)$) For $C = \Omega$, this corollary yields: If $E^{Z=z}(Y|X)$ is P-unique, then

$$E^{Z=z}(Y|X) \underset{P}{=} E(Y|X, Z=z), \tag{14.57}$$

and

$$E^{Z=z}(Y|X=x) = E(Y|X=x, Z=z), \quad \text{for } P_X\text{-a.a. } x \in \Omega'_X. \tag{14.58}$$

◁

Now we consider the family of factorizations g_B of $E^B(Y|X)$, which are defined by Equation (14.29). Because each element of $\mathscr{E}^B(Y|X)$ has at least one factorization g_B, there is a family of factorizations, which are random variables on the probability space $(\Omega'_X, \mathscr{A}'_X, P_X)$.

The next corollary immediately follows from Corollary 5.21 (i) if P^C takes the role of P and P_X^C the role of P_X.

Corollary 14.58 (P^C-Uniqueness and P_X^C-Uniqueness)
Let the assumptions 14.5 hold, let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega'_X, \mathscr{A}'_X)$ be a random variable, let $V = g_B(X)$ and $V^* = g_B^*(X)$ be two elements of $\mathscr{E}^B(Y|X)$, and let $C \in \mathscr{A}$ with $P(C) > 0$. Then

$$V \underset{P^C}{=} V^* \quad \Leftrightarrow \quad g_B \underset{P_X^C}{=} g_B^*. \tag{14.59}$$

Remark 14.59 (P-Uniqueness and P_X-Uniqueness) For $C = \Omega$, Corollary 14.58 yields: If $V = g_B(X)$ and $V^* = g_B^*(X)$ are two elements of $\mathscr{E}^B(Y|X)$, then

$$g_B(X) \underset{P}{=} g_B^*(X) \quad \Leftrightarrow \quad g_B \underset{P_X}{=} g_B^*. \tag{14.60}$$

Note that both sides of (14.60) are not only equivalent to each other but also to

$$g_B(x) = g_B^*(x), \quad \text{for } P_X\text{-a.a. } x \in \Omega_X', \tag{14.61}$$

as well as to the propositions on the left-hand side and the right-hand side of (x) in Box 14.1. ◁

Remark 14.60 (Some Formulas for the Expectation of $E^B(Y|X)$) Suppose that the assumptions 14.5 hold, where Y is nonnegative or with finite expectation $E(Y)$, and let $X\colon (\Omega, \mathscr{A}, P) \to (\Omega_X', \mathscr{A}_X')$ be a random variable. If $E^B(Y|X)$ is P-unique, then, according to Equations (6.13) and (14.33),

$$E[E^B(Y|X)] = \int E^B(Y|X{=}x)\ P_X(dx) = \int E(Y|X{=}x, 1_B{=}1)\ P_X(dx). \tag{14.62}$$

Furthermore, if X is discrete and $P_X(\Omega_0') = 1$, then

$$\begin{aligned} E[E^B(Y|X)] &= \int E^B(Y|X{=}x)\ P_X(dx) && [(6.13)] \\ &= \sum_{x \in \Omega_0'} E^B(Y|X{=}x) \cdot P(X{=}x) && [(6.15)] \\ &= \sum_{x \in \Omega_0'} E(Y|X{=}x, 1_B{=}1) \cdot P(X{=}x). && [(14.33)] \end{aligned} \tag{14.63}$$

◁

Remark 14.61 (Some Formulas for the Expectation of $E^{Z=z}(Y|X)$) Correspondingly, let the assumptions 14.19 hold, where Y is nonnegative or with finite expectation $E(Y)$. If $E^{Z=z}(Y|X)$ is P-unique, then for $B = \{Z{=}z\}$, Equations (14.62) and (14.58) yield

$$E[E^{Z=z}(Y|X)] = \int E^{Z=z}(Y|X{=}x)\ P_X(dx) = \int E(Y|X{=}x, Z{=}z)\ P_X(dx), \tag{14.64}$$

and, if X is discrete and $P_X(\Omega_0') = 1$, then Equations (14.63) and (14.58) yield

$$\begin{aligned} E[E^{Z=z}(Y|X)] &= \int E^{Z=z}(Y|X{=}x)\ P_X(dx) \\ &= \sum_{x \in \Omega_0'} E(Y|X{=}x, Z{=}z) \cdot P(X{=}x). \end{aligned} \tag{14.65}$$

◁

14.5 Conditional Mean Independence With Respect to $P^{Z=z}$

According to the following theorem, a numerical random variable Y on $(\Omega, \mathscr{A}, P)$ that is nonnegative or with finite expectation $E(Y)$ is $\mathscr{C}$-conditionally mean independent from Z with respect to $P^{Z=z}$. For simplicity, we use the notation

$$E^{Z=z}(Y|\mathscr{C},\mathscr{D}) := E^{Z=z}\big(Y\big|\,\sigma(\mathscr{C}\cup\mathscr{D})\big), \tag{14.66}$$

and

$$E^{Z=z}(Y|\mathscr{C},Z) := E^{Z=z}\big(Y\big|\,\sigma[\mathscr{C}\cup\sigma(Z)]\big). \tag{14.67}$$

Theorem 14.62 (Conditional Mean Independence)
Let the assumptions 14.5 hold, where Y is nonnegative or with finite expectation $E(Y)$. Furthermore, let $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ be a random variable, and suppose that $z\in\Omega'_Z$ with $P(Z=z)>0$. Then

$$E^{Z=z}(Y|\mathscr{C}),\, E(Y|\mathscr{C},Z) \in \mathscr{E}^{Z=z}(Y|\mathscr{C},Z),$$

which implies

$$E^{Z=z}(Y|\mathscr{C},Z) \underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathscr{C}) \underset{P^{Z=z}}{=} E(Y|\mathscr{C},Z). \tag{14.68}$$

(Proof p. 411)

Remark 14.63 (A Caveat) If $\sigma(\mathscr{C},Z)\neq\mathscr{C}$, then $E^{Z=z}(Y|\mathscr{C},Z)$ and $E(Y|\mathscr{C},Z)$ are not necessarily elements of $\mathscr{E}^{Z=z}(Y|\mathscr{C})$ [see Def. 14.7 (a)]. Nevertheless,

$$E^{Z=z}(Y|\mathscr{C}) \underset{P^{Z=z}}{=} E(Y|\mathscr{C},Z).$$

According to Box 6.1 (ix) and Box 10.2 (iv), this implies

$$E^{Z=z}(Y) = E^{Z=z}\big[E^{Z=z}(Y|\mathscr{C})\big] = E^{Z=z}\big[E(Y|\mathscr{C},Z)\big], \tag{14.69}$$

and, for $\mathscr{C}=\sigma(X)$,

$$\begin{aligned} E^{Z=z}(Y) &= E^{Z=z}\big[E^{Z=z}(Y|X)\big] && \text{[Box 10.2 (iv)]}\\ &= E^{Z=z}\big[E(Y|X,Z=z)\big] && \text{[(14.27)]}\\ &= E^{Z=z}\big[E(Y|X,Z)\big]. && \text{[(14.27), (14.68)]} \end{aligned} \tag{14.70}$$

Note that we still presume that assumptions of Theorem 14.62 hold. ◁

Remark 14.64 (Two Implications Concerning Mean Independence) Let the assumptions 14.19 hold, where Y is nonnegative or with finite expectation $E^{Z=z}(Y)$. Then

$$E^{Z=z}(Y|\mathscr{C},X) \underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathscr{C}), \quad \text{if } \sigma(X)\subset\sigma(Z) \tag{14.71}$$

(see Exercise 14-15). Furthermore, considering a σ-algebra $\mathscr{D}$ and assuming that Y is nonnegative or with finite expectation $E(Y)$,

$$E^{Z=z}\big[E(Y|\mathscr{C},Z)\,|\mathscr{D}\big] \underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathscr{D}), \quad \text{if } \mathscr{D}\subset\sigma(\mathscr{C},Z) \tag{14.72}$$

(see Exercise 14-16). ◁

In the following theorem we generalize the propositions of Remark 14.61.

Theorem 14.65 (Expectation of $E^{Z=z}(Y|X,W)$ With Respect to $P^{W=w}$)
Let the assumptions 14.19 hold, where Y is nonnegative or with finite expectation $E(Y)$, let $W\colon (\Omega,\mathcal{A},P)\to(\Omega'_W,\mathcal{A}'_W)$ be a random variable, let $w\in\Omega'_W$ with $P(W=w)>0$, and assume that $E^{Z=z}(Y|X,W)$ is $P^{W=w}$-unique. Then

$$\begin{aligned} E^{W=w}\big[E^{Z=z}(Y|X,W)\big] &= \int E^{Z=z}(Y|X=x,W=w)\; P_X^{W=w}(dx) \\ &= \int E(Y|X=x,W=w,Z=z)\; P_X^{W=w}(dx). \end{aligned} \tag{14.73}$$

(Proof p. 412)

Remark 14.66 (Discrete X) If the assumptions of Theorem 14.65 hold, X is discrete, and $P_X(\Omega'_0)=1$, then Equations (14.73), (6.15), (14.58), and (9.13) yield

$$\begin{aligned} E^{W=w}[E^{Z=z}(Y|X,W)] &= \int E^{Z=z}(Y|X=x,W=w)\; P_X^{W=w}(dx) \\ &= \sum_{x\in\Omega'_0} E(Y|X=x,W=w,Z=z)\cdot P^{W=w}(X=x) \\ &= \sum_{x\in\Omega'_0} E(Y|X=x,W=w,Z=z)\cdot P(X=x|W=w). \end{aligned} \tag{14.74}$$

◁

In the following theorem we study an implication of conditional mean independence [see Def. 10.44 (ii)] on conditional expectations with respect to $P^{Z=z}$.

Theorem 14.67 (An Implication of Conditional Mean Independence)
Let the assumptions 14.5 hold, where Y is nonnegative or with finite expectation $E(Y)$, and let $\mathcal{C},\mathcal{D}\subset\mathcal{A}$ be σ-algebras. Then

$$E(Y|\mathcal{C},\mathcal{D},Z)\underset{P}{=}E(Y|\mathcal{C},Z)\quad\Rightarrow\quad E^{Z=z}(Y|\mathcal{C},\mathcal{D})\underset{P^{Z=z}}{=}E^{Z=z}(Y|\mathcal{C}). \tag{14.75}$$

(Proof p. 413)

Remark 14.68 (An Implication of Conditional Mean Independence) Let the assumptions of Theorem 14.67 hold. If X and W are two random variables on the probability space $(\Omega,\mathcal{A},P)$, then

$$E(Y|X,W,Z)\underset{P}{=}E(Y|X,Z)\quad\Rightarrow\quad E^{Z=z}(Y|X,W)\underset{P^{Z=z}}{=}E^{Z=z}(Y|X). \tag{14.76}$$

Furthermore, for $X=\alpha$, $\alpha\in\Omega'_X$, we can conclude that $E(Y|W,Z)\underset{P}{=}E(Y|Z)$ implies

$$E^{Z=z}(Y|W)\underset{P^{Z=z}}{=}E^{Z=z}(Y). \tag{14.77}$$

◁

14.6 Proofs

Proof of Theorem 14.23

Because $f(Z) \underset{P^{Z=z}}{=} f(z)$ (see Rem. 9.1),

$$\begin{aligned} E^{Z=z}\big[f(Z)\cdot Y\,\big|\mathscr{C}\big] &\underset{P^{Z=z}}{=} E^{Z=z}\big[f(z)\cdot Y\,\big|\mathscr{C}\big] && \text{[Box 10.1 (ix)]} \\ &\underset{P^{Z=z}}{=} f(z)\cdot E^{Z=z}(Y|\mathscr{C}). && \text{[Box 10.1 (xiv)]} \end{aligned}$$

Proof of Theorem 14.29

If the assumptions of Definition 14.26 hold, $g(X,Z)$ is a composition of $(X,Z)\colon (\Omega,\mathscr{A},P) \to (\Omega'_X\times\Omega'_Z,\mathscr{A}'_X\otimes\mathscr{A}'_Z)$ and a function $g\colon \Omega'_X\times\Omega'_Z\to\overline{\mathbb{R}}$, and g_z is defined by (14.21), then

$$1_{Z=z}\cdot g(X,Z) \;=\; 1_{Z=z}\cdot g_z(X). \tag{14.78}$$

This is easily seen as follows: For all $\omega\in\Omega$,

$$\begin{aligned} 1_{Z=z}(\omega)\cdot g[X(\omega),Z(\omega)] &= \begin{cases} 0, & \text{if } 1_{Z=z}(\omega)=0 \\ g_z[X(\omega)], & \text{if } 1_{Z=z}(\omega)=1 \end{cases} \\ &= \begin{cases} 1_{Z=z}(\omega)\cdot g_z[X(\omega)], & \text{if } 1_{Z=z}(\omega)=0 \\ 1_{Z=z}(\omega)\cdot g_z[X(\omega)], & \text{if } 1_{Z=z}(\omega)=1 \end{cases} \\ &= 1_{Z=z}(\omega)\cdot g_z[X(\omega)]. \end{aligned}$$

Because $g\colon \Omega'_X\times\Omega'_Z\to\overline{\mathbb{R}}$ is $(\mathscr{A}'_X\otimes\mathscr{A}'_Z,\overline{\mathscr{B}})$-measurable, the function g_z defined in (14.21) is $(\mathscr{A}'_X,\overline{\mathscr{B}})$ measurable (see Bauer, 2001, Lemma 23.5. p. 138). Hence, condition (a) of Definition 14.7 holds. Furthermore, for all $C\in\sigma(X)$, we can conclude $C\cap\{Z{=}z\}\in\sigma(X,Z)$, and therefore, for all $C\in\sigma(X)$,

$$\begin{aligned} \int 1_C\; E(Y|X,Z{=}z)\; dP^{Z=z} &= \frac{1}{P(Z{=}z)}\int 1_C\cdot 1_{Z=z}\; E(Y|X,Z{=}z)\; dP && \text{[(9.15)]} \\ &= \frac{1}{P(Z{=}z)}\int 1_C\cdot 1_{Z=z}\, g_z(X)\; dP && \text{[(14.23)]} \\ &= \frac{1}{P(Z{=}z)}\int 1_C\cdot 1_{Z=z}\, g(X,Z)\; dP && \text{[(14.78)]} \\ &= \frac{1}{P(Z{=}z)}\int 1_C\cdot 1_{Z=z}\; E(Y|X,Z)\; dP && \text{[(10.22)]} \\ &= \frac{1}{P(Z{=}z)}\int 1_C\cdot 1_{Z=z}\, Y\; dP && \text{[Def. 10.2 (b)]} \\ &= \int 1_C\cdot Y\; dP^{Z=z}. && \text{[(9.15)]} \end{aligned}$$

This shows that condition (b) of Definition 14.7 holds as well.

Proof of Theorem 14.44

(b) ⇒ (a). Remember that (b) is defined by

$$\forall C \in \mathscr{C}: \quad P^B(C) = 0 \quad \Rightarrow \quad P(C) = 0.$$

If $V, V^* \in \mathscr{E}^B(Y|\mathscr{C})$, then $\{V \neq V^*\} \in \mathscr{C}$, $P^B(\{V \neq V^*\}) = 0$, and (b) implies $P(\{V \neq V^*\}) = 0$.

(a) ⇒ (b). This proposition is proved by contraposition, i. e., we show ¬ (b) ⇒ ¬ (a). Assume that there is an $A \in \mathscr{C}$ with $P^B(A) = 0$ and $P(A) > 0$, and let $V \in \mathscr{E}^B(Y|\mathscr{C})$ with $V(\omega) = 0$ for all $\omega \in A$. [Note that if $V' \in \mathscr{E}^B(Y|\mathscr{C})$ and $A \in \mathscr{C}$ with $P^B(A) = 0$, then $V := V' \cdot 1_{A^c} \in \mathscr{E}^B(Y|\mathscr{C})$.] Then $V^* = V + 1_A$ is also $\mathscr{C}$-measurable and, for all $C \in \mathscr{C}$,

$$\begin{aligned} \int 1_C V^* \, dP^B &= \int 1_C \cdot (V + 1_A) \, dP^B \\ &= \int 1_C V \, dP^B + \int 1_C 1_A \, dP^B \\ &= \int 1_C V \, dP^B, \end{aligned}$$

because $\int 1_C \cdot 1_A \, dP^B = P^B(C \cap A) = 0$. Hence, according to Definition 14.7, $V^* = V + 1_A \in \mathscr{E}^B(Y|\mathscr{C})$. However, $P(\{V \neq V^*\}) = P(A) > 0$.

(c) ⇒ (b). Let $P(B|\mathscr{C}) \underset{P}{\geq} 0$. This implies $P(A) = 0$, where

$$A = \{\omega \in \Omega: E(1_B|\mathscr{C})(\omega) = 0\}.$$

However, if $P(A) = 0$, then, according to Rule (vii) of Box 4.1,

$$P(A \cup C) = P(C), \quad \forall C \in \mathscr{C}. \tag{14.79}$$

Now, let $C \in \mathscr{C}$ with $P^B(C) = 0$. This yields

$$\begin{aligned} \int E(1_B|\mathscr{C}) \cdot 1_C \, dP &= \int 1_B \cdot 1_C \, dP && \text{[Def. 10.2 (b)]} \\ &= P(B \cap C) = P^B(C) \cdot P(B) = 0. \end{aligned}$$

Because $E(1_B|\mathscr{C}) \cdot 1_C \underset{P}{\geq} 0$ [see Box 10.3 (iii)], Lemma 3.44 yields $E(1_B|\mathscr{C}) \cdot 1_C \underset{P}{=} 0$, which is equivalent to

$$\begin{aligned} 1 &= P\big(\{\omega \in \Omega: E(1_B|\mathscr{C})(\omega) \cdot 1_C(\omega) = 0\}\big) \\ &= P\big(\{\omega \in \Omega: \omega \in A \text{ or } \omega \in C^c\}\big) = P(A \cup C^c) \\ &= P(C^c). && \text{[(14.79)]} \end{aligned}$$

Hence, $P(C) = 0$, which shows that $P(B|\mathscr{C}) \underset{P}{\geq} 0$ implies (b).

¬ (c) ⇒ ¬ (b). Again, let $A = \{\omega \in \Omega: E(1_B|\mathscr{C})(\omega) = 0\}$ and assume $P(A) > 0$. Now, $E(1_B|\mathscr{C})(\omega) = 0$ for all $\omega \in A$ implies $E(1_B|\mathscr{C}) \cdot 1_A = 0$. Therefore,

$$\begin{aligned} 0 &= E[E(1_B|\mathscr{C}) \cdot 1_A] && \text{[Box 6.1 (i)]} \\ &= E[E(1_B \cdot 1_A|\mathscr{C})] && [A \in \mathscr{C}, \text{ Box 10.1 (xiv)}] \\ &= E(1_B \cdot 1_A) && \text{[10.1 (iv)]} \\ &= E(1_{A \cap B}) && [1_B \cdot 1_A = 1_{A \cap B}]. \\ &= P(A \cap B). && \text{[(6.4)]} \end{aligned}$$

Because $P(B) > 0$, the equation $P(A \cap B) = P^B(A) \cdot P(B) = 0$ implies $P^B(A) = 0$. Hence, $A \in \mathscr{C}$ with $P^B(A) = P(A \cap B)/P(B) = 0$ and $P(A) > 0$.

(b) $\Rightarrow$ (d). As has been shown above, (b) is equivalent to P-uniqueness of $E^B(Y|\mathscr{C})$, and according to Remark 10.8 and Box 6.1 (ix), P-uniqueness of $E^B(Y|\mathscr{C})$ implies (d).

(d) $\Rightarrow$ (b). This proposition is proved by contraposition, i. e., we show $\neg$ (b) $\Rightarrow \neg$ (d). Assume that there is an $A \in \mathscr{C}$ with $P^B(A) = 0$ and $P(A) > 0$, and let $V \in \mathscr{E}^B(Y|\mathscr{C})$ be real-valued [see Box 10.1 (x)] and have a finite expectation $E(V)$. Then $V^* = V + 1_A$ is also $\mathscr{C}$-measurable and, for all $C \in \mathscr{C}$,

$$\begin{aligned}\int 1_C\, V^*\, dP^B &= \int 1_C \cdot (V + 1_A)\, dP^B \\ &= \int 1_C\, V\, dP^B + \int 1_C\, 1_A\, dP^B \\ &= \int 1_C\, V\, dP^B,\end{aligned}$$

because $\int 1_C \cdot 1_A\, dP^B = P^B(C \cap A) = 0$. Therefore, according to Definition 14.7, $V^* = V + 1_A \in \mathscr{E}^B(Y|\mathscr{C})$ and $P(\{V \neq V^*\}) = P(A) > 0$. Now,

$$\begin{aligned}E(V^*) &= \int V^*\, dP && [(6.1)] \\ &= \int (V + 1_A)\, dP && [V^* := V + 1_A] \\ &= \int V\, dP + \int 1_A\, dP && [(3.34)] \\ &= E(V) + P(A). && [(6.1), (6.4)]\end{aligned}$$

Because $E(V)$ is finite and $P(A) > 0$ it follows that $E(V) \neq E(V^*)$.

Proof of Corollary 14.56

If $E^{Z=z}(Y|X)$ is P^C-unique, then all pairs of elements of $\mathscr{E}^{Z=z}(Y|X)$ are P^C-equivalent. According to Theorem 14.29, $E(Y|X, Z{=}z)$ is an element of $\mathscr{E}^{Z=z}(Y|X)$, which implies Equation (14.55). Finally, according to Corollary 5.21, Equation (14.56) is equivalent to (14.55).

Proof of Theorem 14.62

First, we show that $E(Y|\mathscr{C}, Z) \in \mathscr{E}^{Z=z}(Y|\mathscr{C}, Z)$. By Definition 10.2 (a), $\sigma[E(Y|\mathscr{C}, Z)] \subset \sigma(\mathscr{C}, Z)$, which implies that condition (a) of Definition 14.7 holds. In order to show condition (b) of Definition 14.7 note that, for all $C \in \sigma(\mathscr{C}, Z)$, $\{Z{=}z\}, \{Z{=}z\} \cap C \in \sigma(\mathscr{C}, Z)$ and $1_{\{Z=z\}\cap C} = 1_{Z=z} 1_C$. Hence, for all $C \in \sigma(\mathscr{C}, Z)$,

$$\begin{aligned}E^{Z=z}[1_C\, E(Y|\mathscr{C}, Z)] &= \frac{1}{P(Z{=}z)}\, E[1_{Z=z} 1_C \cdot E(Y|\mathscr{C}, Z)] && [(9.11)] \\ &= \frac{1}{P(Z{=}z)}\, E(1_{Z=z} 1_C \cdot Y) && [\text{Def. 10.2 (b)}] \\ &= E^{Z=z}(1_C \cdot Y). && [(9.11)]\end{aligned}$$

Now we show that $E^{Z=z}(Y|\mathscr{C}) \in \mathscr{E}^{Z=z}(Y|\mathscr{C}, Z)$. By Definition 14.7 (a) and the definition of $\sigma(\mathscr{C}, Z)$,

$$\sigma[E^{Z=z}(Y|\mathscr{C})] \subset \mathscr{C} \subset \sigma(\mathscr{C}, Z),$$

and hence, $\sigma[E^{Z=z}(Y|\mathscr{C}, Z)] \subset \sigma(\mathscr{C}, Z)$.

In order to prove condition (b) of Definition 14.7, note that, for the traces (see Example 1.10),

$$\sigma(\mathscr{C}, Z)\big|_{\{Z=z\}} = \mathscr{C}\big|_{\{Z=z\}}$$

[see Eqs. (1.16) and (1.16) with $\mathscr{E} = \sigma(Z)$ and $A := \{Z{=}z\}$]. Hence, for all $C \in \sigma(\mathscr{C}, Z)$, there is an $A_C \in \mathscr{C}$ such that

$$\{Z{=}z\} \cap C = \{Z{=}z\} \cap A_C. \tag{14.80}$$

Therefore, for all $C \in \sigma(\mathscr{C}, Z)$,

$$\begin{aligned}
E^{Z=z}[1_C \cdot E^{Z=z}(Y|\mathscr{C})] &= \frac{1}{P(Z{=}z)} E[1_{Z=z} 1_C \cdot E^{Z=z}(Y|\mathscr{C})] && [(9.11)]\\
&= \frac{1}{P(Z{=}z)} E[1_{Z=z} 1_{A_C} \cdot E^{Z=z}(Y|\mathscr{C})] && [(14.80)]\\
&= E^{Z=z}[1_{A_C} \cdot E^{Z=z}(Y|\mathscr{C})] && [(9.11)]\\
&= E^{Z=z}(1_{A_C} \cdot Y) && [\text{Def. 14.7 (b)}]\\
&= \frac{1}{P(Z{=}z)} E(1_{Z=z} 1_{A_C} \cdot Y) && [(9.11)]\\
&= E^{Z=z}(1_{Z=z} 1_{A_C} \cdot Y) && [1_{Z=z} = 1_{Z=z} \cdot 1_{Z=z}]\\
&= E^{Z=z}(1_C \cdot Y). && [(14.80)]
\end{aligned}$$

Proof of Theorem 14.65

Let $h\colon \Omega'_X \times \Omega'_W \to \overline{\mathbb{R}}$ be nonnegative or with finite expectation with respect to the measure $P^{W=w}_{X,W}$. Then

$$\begin{aligned}
&\int h(x,w')\, P^{W=w}_{X,W}[d(x,w')]\\
&= \int h(X,W)\, 1_{W=w}\, dP^{W=w} && [(6.13), (9.7)]\\
&= \int h(X,w)\, dP^{W=w} && [h(X,W)\cdot 1_{W=w} = h(X,w)\cdot 1_{W=w}, (9.7)]\\
&= \int h(x,w)\, P^{W=w}_X(dx). && [(6.18)]
\end{aligned} \tag{14.81}$$

Furthermore, let $z \in \Omega'_Z$ with $P(Z{=}z) > 0$ and $w \in \Omega'_W$ with $P(W{=}w) > 0$. If W takes the role of Z and $P^{Z=z}$ the role of P in Equation (14.70), then the last equation of (14.70) yields

$$E^{W=w}\big[E^{Z=z}(Y\,|\,X,W)\big] = E^{W=w}\big[E^{Z=z}(Y\,|\,X,W{=}w)\big].$$

Hence,

$$\begin{aligned}
E^{W=w}\big[E^{Z=z}(Y\,|\,X,W)\big] &= E^{W=w}\big[E^{Z=z}(Y\,|\,X,W{=}w)\big]\\
&= \int E^{Z=z}(Y\,|\,X,W{=}w)\, dP^{W=w} && [(6.2)]
\end{aligned}$$

$$
\begin{aligned}
&= \int E^{Z=z}(Y \mid X{=}x, W{=}w)\, P_X^{W=w}(dx) && \text{[(6.13)]}\\
&= \int E^{Z=z}(Y \mid X{=}x, W{=}w')\, P_{X,W}^{W=w}[d(x,w')] && \text{[(14.81)]}\\
&= \int E(Y \mid X{=}x, Z{=}z, W{=}w')\, P_{X,W}^{W=w}[d(x,w')] && \text{[(14.56)]}\\
&= \int E(Y \mid X{=}x, Z{=}z, W{=}w)\, P_X^{W=w}(dx). && \text{[(14.81)]}
\end{aligned}
$$

Proof of Theorem 14.67

Obviously, $E^{Z=z}(Y|\mathscr{C})$ is $\sigma(\mathscr{C},\mathscr{D})$-measurable. Furthermore,

$$
\begin{aligned}
E^{Z=z}(Y|\mathscr{C},\mathscr{D}) &\underset{P^{Z=z}}{=} E(Y|\mathscr{C},\mathscr{D},Z) && \text{[Eqs. (14.68)]}\\
&\underset{P^{Z=z}}{=} E(Y|\mathscr{C},Z) && [E(Y|\mathscr{C},\mathscr{D},Z) \underset{P}{=} E(Y|\mathscr{C},Z), \text{ Cor. 5.18}]\\
&\underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathscr{C}). && \text{[Eqs. (14.68)]}
\end{aligned}
$$

14.7 Exercises

▷ **Exercise 14-1** Compute the values of the conditional expectation $E(Y|X)$ in the example presented in Table 14.3 (p. 387).

▷ **Exercise 14-2** Compute the values of the conditional expectation $E(Y|X,U)$ in the example presented in Table 14.3 (p. 387).

▷ **Exercise 14-3** Consider Table 14.3. Why do the values of the conditional expectation $E(Y|X,U)$ have to be identical for $\omega_3 = (\textit{Joe}, \textit{yes}, -)$ and $\omega_4 = (\textit{Joe}, \textit{yes}, +)$?

▷ **Exercise 14-4** Show that if $Y\colon (\Omega,\mathscr{A},P) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ is a random variable with finite expectation $E(Y)$, then $E^B(Y)$ is finite.

▷ **Exercise 14-5** Let $(\Omega,\mathscr{A},P)$, $(\Omega,\mathscr{A},Q)$ be probability spaces, and consider the random variable $X\colon (\Omega,\mathscr{A},P) \to (\Omega'_X,\mathscr{A}'_X)$. Show that X is a random variable on $(\Omega,\mathscr{A},P)$ if and only if it is also a random variable on $(\Omega,\mathscr{A},Q)$.

▷ **Exercise 14-6** Show that the assumptions 14.5 imply $\mathscr{E}^B(1_B \cdot Y|\mathscr{C}) \subset \mathscr{E}^B(Y|\mathscr{C})$.

▷ **Exercise 14-7** Prove Equation (14.24).

▷ **Exercise 14-8** Compute the values of the conditional expectation $E^{X=0}(Y \mid U)$ in the example presented in Table 14.3 (p. 387).

▷ **Exercise 14-9** Prove propositions (i) and (ii) of Remark 14.33.

▷ **Exercise 14-10** Show that Theorem 14.29 implies Equation (14.32).

▷ **Exercise 14-11** Prove: For all $x \in X(\Omega) = \{0,1\}$: If $P(X{=}x) > 0$ and $E^{X=0}(Y|Z)$ is $P^{X=x}$-unique, then $E(V_0|X{=}x) = E(V_0^*|X{=}x)$ for all versions $V_0, V_0^* \in \mathscr{E}^{X=0}(Y|Z)$.

▷ **Exercise 14-12** Show that the expectation $E(V)$ of a version $V \in \mathscr{E}^B(Y|\mathscr{C})$ exists and is real-valued if $\mathscr{C}$ is a finite set and $E^B(Y)$ is finite.

▷ **Exercise 14-13** Prove Remark 14.47 for X being discrete.

▷ **Exercise 14-14** Prove the propositions summarized in Box 14.1.

▷ **Exercise 14-15** Prove Equation (14.71).

▷ **Exercise 14-16** Prove Equation (14.72).

▷ **Exercise 14-17** Table 14.3 (p. 387) contains a first element, say V, of $\mathscr{E}(Y|X,U)$. Define a second element of $\mathscr{E}(Y|X,U)$ and show that the two elements are P-equivalent. Repeat this task for the conditional expectation $E^{X=1}(Y|U)$ and show that the two elements are $P^{X=1}$-equivalent.

Solutions

▷ **Solution 14-1** The values of the conditional expectation $E(Y|X)$ are the two conditional expectation values $E(Y|X{=}0)$ and $E(Y|X{=}1)$. Because $E(Y|X{=}x) = P(Y{=}1|X{=}x)$, they can be computed as follows:

$$P(Y{=}1|X{=}0) = \frac{P(Y{=}1,X{=}0)}{P(X{=}0)} = \frac{.348+.024}{.152+.348+.096+.024} = .60,$$

and

$$P(Y{=}1|X{=}1) = \frac{P(Y{=}1,X{=}1)}{P(X{=}1)} = \frac{0+.152}{0+0+.228+.152} = .40.$$

▷ **Solution 14-2** The values of $E(Y|X,U)$ are the four conditional expectation values $E(Y|X{=}x,U{=}u)$. Because $E(Y|X{=}x,U{=}u) = P(Y{=}1|X{=}x,U{=}u)$, they can be computed as follows:

$$P(Y{=}1|X{=}0,U{=}\mathit{Joe}) = \frac{P(Y{=}1,X{=}0,U{=}\mathit{Joe})}{P(X{=}0,U{=}\mathit{Joe})} = \frac{.348}{.152+.348} = .696,$$

$$P(Y{=}1|X{=}0,U{=}\mathit{Ann}) = \frac{P(Y{=}1,X{=}0,U{=}\mathit{Ann})}{P(X{=}0,U{=}\mathit{Ann})} = \frac{.024}{.096+.024} = .20,$$

$$P(Y{=}1|X{=}1,U{=}\mathit{Ann}) = \frac{P(Y{=}1,X{=}1,U{=}\mathit{Ann})}{P(X{=}1,U{=}\mathit{Ann})} = \frac{.152}{.228+.152} = .40.$$

The conditional expectation value $E(Y|X{=}1,U{=}\mathit{Joe}) = P(Y{=}1|X{=}1,U{=}\mathit{Joe})$ is undefined, because $P(X{=}1,U{=}\mathit{Joe}) = 0$. Choosing any number (such as 9) as a value of $E(Y|X,U)$ for $\omega_3 = (\mathit{Joe},\mathit{yes},-)$ and $\omega_4 = (\mathit{Joe},\mathit{yes},+)$ yields an element of $\mathscr{E}(Y|X,U)$. Different elements of $\mathscr{E}(Y|X,U)$ are identical almost surely with respect to the measure P.

▷ **Solution 14-3** If $E(Y|X,U)(\omega_3) \neq E(Y|X,U)(\omega_4)$, then $E(Y|X,U)$ would not be measurable with respect to (X,U). Measurability of a random variable V with respect to (X,U) requires that V takes on only one single value for all $\omega \in (X,U)^{-1}(\{(x,u)\})$ (see Cor. 2.53).

▷ **Solution 14-4** According to Equation (9.11),

$$E^B(Y) \;=\; \frac{1}{P(B)}\,E(1_B\,Y).$$

If Y has a finite expectation with respect to P, then $\int Y^+dP<\infty$ and $\int Y^-dP<\infty$. Because $0\le 1_BY^+\le Y^+$ and $0\le 1_BY^-\le Y^-$, Equation (3.24) yields $\int 1_BY^+dP<\infty$ and $\int 1_BY^-dP<\infty$, and therefore $-\infty<\int 1_BYdP<\infty$. Therefore, if $-\infty<E(Y)<\infty$, then $-\infty<E(1_BY)<\infty$.

▷ **Solution 14-5** The definition of a random variable $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ only requires that there is a probability measure, here denoted P, on a measurable space $(\Omega,\mathscr{A})$ and that X is measurable with respect to $\mathscr{A}$, i. e., $X^{-1}(A')\in\mathscr{A}$, for all $A'\in\mathscr{A}'_X$. Hence, if X is a random variable on $(\Omega,\mathscr{A},P)$, then it is also a random variable on $(\Omega,\mathscr{A},Q)$, whenever P and Q are probability measures on $(\Omega,\mathscr{A})$.

▷ **Solution 14-6** According to condition (a) of Definition 14.7, $E^B(1_B\cdot Y|\mathscr{C})$ is measurable with respect to $\mathscr{C}$. Condition (b) holds as well, because, for all $C\in\mathscr{C}$,

$$\begin{aligned}
E^B[1_C\cdot E^B(1_B\cdot Y|\mathscr{C})] \;&=\; E^B[E^B(1_C\cdot 1_B\cdot Y|\mathscr{C})] && [C\in\mathscr{C}\text{, Box 10.2 (xiv)}]\\
&=\; E^B(1_C\cdot 1_B\cdot Y) && [\text{Box 10.2 (iv)}]\\
&=\; \frac{1}{P(B)}\cdot E(1_C\cdot 1_B\cdot 1_B\cdot Y) && [(9.7)]\\
&=\; \frac{1}{P(B)}\cdot E(1_C\cdot 1_B\cdot Y) && [1_B\cdot 1_B=1_B]\\
&=\; E^B(1_C\cdot Y). && [(9.7)]
\end{aligned}$$

▷ **Solution 14-7**

$$\begin{aligned}
\forall\,\omega\in\Omega\colon\qquad &\left(\sum_{z\in Z(\Omega)}E(Y|X,Z{=}z)\cdot 1_{Z=z}\right)(\omega)\\
&=\left(\sum_{z\in Z(\Omega)}g_z(X)\cdot 1_{Z=z}\right)(\omega) && [(14.23)]\\
&=\left(\sum_{z\in Z(\Omega)}g(X,Z)\cdot 1_{Z=z}\right)(\omega) && [(14.78)]\\
&=\sum_{z\in Z(\Omega)}g(X,Z)(\omega)\cdot 1_{Z=z}(\omega) && [(2.31)]\\
&=\sum_{z\in Z(\Omega)}g[X(\omega),Z(\omega)]\cdot 1_{Z=z}(\omega) && [(2.31)]\\
&=g[X(\omega),Z(\omega)] && \left[\sum_{z\in Z(\Omega)}1_{Z=z}(\omega)=1\right]\\
&=g[(X,Z)(\omega)] && [(5.18)]\\
&=g(X,Z)(\omega) && [(2.26)]\\
&=E(Y|X,Z)(\omega). && [\text{Def. 14.26}]
\end{aligned}$$

▷ **Solution 14-8** The values of the conditional expectation $E^{X=0}(Y|U)$ are the two conditional expectation values $E^{X=0}(Y|U{=}Joe)$ and $E^{X=0}(Y|U{=}Ann)$ with respect to the measure $P^{X=0}$. Because $E^{X=0}(Y|U{=}u)=P^{X=0}(Y{=}1|U{=}u)$, they can be computed as follows:

$$P^{X=0}(Y{=}1\,|\,U{=}Joe) \;=\; \frac{P^{X=0}(Y{=}1, U{=}Joe)}{P^{X=0}(U{=}Joe)} \;\approx\; \frac{.561}{.561+.245} \;\approx\; .696$$

and

$$P^{X=0}(Y{=}1\,|\,U{=}Ann) \;=\; \frac{P^{X=0}(Y{=}1, U{=}Ann)}{P^{X=0}(U{=}Ann)} \;\approx\; \frac{.039}{.039+.155} \;\approx\; .2.$$

▷ **Solution 14-9** (i)

If $g(X,Z) \in \mathscr{E}(Y\,|\,X,Z)$ and $g_{Z=z}(X) \in \mathscr{E}^{Z=z}(Y\,|\,X{=}x)$, then for all $C \in \sigma(X)$,

$$\begin{aligned}
&\int 1_C \cdot g_z(X)\, d\,P^{Z=z} \\
&= \frac{1}{P(Z{=}z)} \int 1_C \cdot 1_{Z=z}\, g_z(X)\, d\,P && \text{[(14.21), (9.11)]} \\
&= \frac{1}{P(Z{=}z)} \int 1_C \cdot 1_{Z=z}\, g(X,Z)\, d\,P && \text{[(14.78)]} \\
&= \frac{1}{P(Z{=}z)} \int 1_C \cdot 1_{Z=z}\, Y\, d\,P && \text{[}C \cap \{Z{=}z\} \in \sigma(X,Z)\text{, Def. 10.2 (b)]} \\
&= \int 1_C\, Y\, d\,P^{Z=z} && \text{[(9.11)]} \\
&= \int 1_C\, g_{Z=z}(X)\, d\,P^{Z=z}. && \text{[Def. 14.7 (b)]}
\end{aligned}$$

Hence, $g_Z(X) \underset{P^{Z=z}}{=} g_{Z=z}(X)$ (see Th. 3.48) or, equivalently, $g_Z(x) = g_{Z=z}(x)$, for $P_X^{Z=z}$-a.a. $x \in \Omega'_X$.

(ii) If $P(X{=}x, Z{=}z) > 0$ for all $x \in \Omega'_X$, then $P(Z{=}z) > 0$ [see Box 4.1 (v)] and

$$P_X^{Z=z}(\{x\}) \;=\; P^{Z=z}(X{=}x) \;=\; \frac{P(X{=}x, Z{=}z)}{P(Z{=}z)} > 0, \quad \forall\, x \in \Omega'_X.$$

Hence, if $P(X{=}x, Z{=}z) > 0$, then Remark 2.71 and $g_Z(X) \underset{P^{Z=z}}{=} g_{Z=z}(X)$ imply $g_Z(x) = g_{Z=z}(x)$ for all $x \in \Omega'_X$.

▷ **Solution 14-10** Note that $\{Z{=}z\} = \{1_{Z=z}{=}1\}$. Hence,

$$\begin{aligned}
E(Y\,|\,X, Z{=}z) &\underset{P^{Z=z}}{=} E^{Z=z}(Y\,|\,X) && \text{[(14.27)]} \\
&\underset{P^{Z=z}}{=} E^{1_{Z=z}=1}(Y\,|\,X) && \text{[(14.15)]} \\
&\underset{P^{Z=z}}{=} E(Y\,|\,X, 1_{Z=z}{=}1). && \text{[(14.27)]}
\end{aligned}$$

For $E(Y\,|\,X,Z) = g(X,Z) \in \mathscr{E}(Y\,|\,X,Z)$,

$$\begin{aligned}
E(Y\,|\,X{=}x, Z{=}z) &= g(x,z) && \text{[(10.27)]} \\
&= g_z(x) && \text{[(14.21)]} \\
&= g_{Z=z}(x), \quad \text{for } P_X^{Z=z}\text{-a.a. } x \in \Omega'_X && \text{[Rem. 14.33 (i)]} \\
&= E^{Z=z}(Y\,|\,X{=}x). && \text{[(14.31)]}
\end{aligned}$$

▷ **Solution 14-11** This proposition follows from the definition of uniqueness of a conditional expectation with respect to a probability measure $P^{X=x}$, equivalence of two random variables with respect to a probability measure $P^{X=x}$, Corollaries 5.20 and 6.17, and Equations (9.5) and (9.6) with $B = \{X{=}x\}$.

▷ **Solution 14-12** If $\mathscr{C} = \{A_1, \dots, A_n\}$ is finite, then there is a finite partition $\{B_1, \dots, B_m\}$ of Ω with $\mathscr{C} = \sigma(\{B_1, \dots, B_m\})$ (see Rem. 1.21). Then, according to Lemma 2.19,

$$V \;=\; \sum_{j=1}^{m} \alpha_j 1_{B_j}, \qquad \text{where } \alpha_j \;=\; \begin{cases} 0, & \text{if } P^B(B_j) = 0, \\ \int 1_{B_j} \cdot Y \, dP^B, & \text{otherwise} \end{cases}$$

hold for all versions $V \in \mathscr{E}^B(Y|\mathscr{C})$. Hence, if $E^B(Y) = \int Y \, dP^B$ is finite, then all α_j, $j = 1, \dots, m$, can be chosen as elements of $\mathbb{R}$, and for such a choice

$$E(V) \;=\; \sum_{j=1}^{m} \alpha_j \cdot P(B_j)$$

is finite as well.

▷ **Solution 14-13**

(d) $\Rightarrow$ (14.45). By definition, condition (d) of Corollary 14.46 is equivalent to

$$\forall\, A' \in \mathscr{A}'_X\colon \quad P^B_X(A') \;=\; 0 \quad \Rightarrow \quad P_X(A') \;=\; 0.$$

If X is discrete and $\{x\} \in \mathscr{A}'_X$ for all $x \in \Omega'_X$, then we can choose $A' = \{x\}$ for all $x \in \Omega'_X$, and this yields

$$\forall\, x \in \Omega'_X\colon \quad P^B_X(\{x\}) \;=\; 0 \quad \Rightarrow \quad P_X(\{x\}) \;=\; 0,$$

which is (14.45) in a different notation [see Eq. (5.30) and Rem. 5.53].

(14.45) $\Rightarrow$ (d). Assume that (14.45) holds and let $A' \in \mathscr{A}'_X$ with

$$P^B_X(A') \;=\; P^B(\{X \in A'\}) \;=\; P(B \cap \{X \in A'\}) \;=\; 0$$

[see Eqs. (5.2), (5.3), and (14.7)]. Because

$$0 \;=\; P(B \cap \{X \in A'\}) \;=\; \sum_{\substack{x \in A' \\ P(X=x)>0}} P(B \cap \{X=x\}), \qquad [\,\sigma\text{-additivity of } P\,]$$

we can conclude: $P(B \cap \{X \in A'\}) = 0$ for all $x \in A'$, and with (14.45), also $P(X=x) = 0$ for all $x \in A'$. However, using σ-additivity of P, this implies

$$P_X(A') \;=\; P(X \in A') \;=\; \sum_{\substack{x \in A' \\ P(X=x)>0}} P(X=x) \;=\; 0.$$

[see again Eqs. (5.2) and (5.3)].

▷ **Solution 14-14** (i). This is the definition of P-uniqueness of $E^B(Y|\mathscr{C})$.

(ii), (iii). These propositions have been proved in Theorem 14.44.

(iv). This proposition follows from Theorem 14.44, because $P \underset{\mathscr{C}}{\ll} P^B$ is equivalent to

$$\forall\, C \in \mathscr{C}\colon \quad P^B(C) = 0 \quad \Rightarrow \quad P(C) = 0,$$

and $\mathscr{D} \subset \mathscr{C}$ implies that this implication also holds for all $C \in \mathscr{D}$.

(v). This proposition immediately follows from applying Corollary 5.18 to the $\mathscr{C}$-conditional expectation $E^B(Y|\mathscr{C})$ of Y with respect to P^B.

(vi), (vii). According to Remark 10.11, P^C-uniqueness of $E^B(Y|\mathscr{C})$ implies $V \underset{P^C}{=} V^*$. The-

orem 2.84 then implies $P_V^C = P_{V^*}^C$. If the two distributions are identical, then the corresponding expectations are identical as well (see Cor. 6.17).

(viii). This proposition immediately follows from Remark 2.76 (iii).

(ix). This proposition is an immediate implication of Corollary 14.40, because uniqueness of $E^B(Y|X)$ implies that $E^B(Y|X)$ is P-unique.

(x). This proposition follows from (14.60) and Remark 10.11.

▷ **Solution 14-15** If X is measurable with respect to Z, then

$$\begin{aligned} E^{Z=z}(Y|\mathscr{C},X) &\underset{P^{Z=z}}{=} E^{Z=z}\big[E^{Z=z}(Y|\mathscr{C},Z)\,\big|\,\mathscr{C},X\big] && \text{[Box 10.1 (v)]}\\ &\underset{P^{Z=z}}{=} E^{Z=z}\big[E^{Z=z}(Y|\mathscr{C})\,\big|\,\mathscr{C},X\big] && \text{[(14.68)]}\\ &\underset{P^{Z=z}}{=} E^{Z=z}(Y|\mathscr{C}). && \text{[Box 10.1 (vii)]} \end{aligned}$$

▷ **Solution 14-16** Because $\mathscr{D} \subset \sigma(\mathscr{C},Z)$,

$$\begin{aligned} E^{Z=z}(Y|\mathscr{D}) &\underset{P^{Z=z}}{=} E^{Z=z}[E^{Z=z}(Y|\mathscr{C},Z)\,|\mathscr{D}] && \text{[Box 10.1 (v)]}\\ &\underset{P^{Z=z}}{=} E^{Z=z}[E(Y|\mathscr{C},Z)\,|\mathscr{D}]. && \text{[(14.68)]} \end{aligned}$$

▷ **Solution 14-17** A second version, say V^*, of the conditional expectation $E(Y|X,U)$ is obtained by replacing the value 9 by the value 8 for $\omega_3 = (Joe, yes, -)$ and $\omega_4 = (Joe, yes, +)$. For these two elements, $P(A_1) = 1$, where $A_1 := \{\omega \in \Omega\colon V(\omega) = V^*(\omega)\}$. The probability $P(A_1) = 1$ is obtained from adding the probabilities of all six outcomes ω for which $P(\{\omega\}) > 0$ (see the last column of Table 14.3, p. 387).

The same argument holds for the conditional expectation $E^{X=1}(Y|U)$. Table 14.3 displays a first element, say V_1, that takes on the values 9 for $\omega_1 = (Joe, no, -)$ to $\omega_4 = (Joe, yes, +)$ and .40 for $\omega_5 = (Ann, no, -)$ to $\omega_8 = (Ann, yes, +)$. A second element, say V_1^*, of the conditional expectation $E^{X=1}(Y|U)$ would be obtained by replacing the value 9 by another (arbitrary) value such as 8. For these two elements, $P^{X=1}(C_1) = 1$, where $C_1 := \{\omega \in \Omega\colon V_1(\omega) = V_1^*(\omega)\} = \{(Ann, no, -), (Ann, no, +), (Ann, yes, -), (Ann, yes, +)\}$. This probability $P^{X=1}(C_1) = 1$ is obtained by

$$\begin{aligned} P^{X=1}(C_1) \;=\; P(C_1\,|\,X{=}1) \;&=\; \frac{P(C_1 \cap \{X{=}1\})}{P(X{=}1)} \;=\; \frac{P\big(\{(Ann, yes, -), (Ann, yes, +)\}\big)}{P(X{=}1)}\\ &=\; \frac{P\big(\{(Ann, yes, -)\}\big) + P\big(\{(Ann, yes, +)\}\big)}{P(X{=}1)}\\ &=\; \frac{.228 + .152}{0 + 0 + .228 + .152} \;=\; \frac{.228 + .152}{.38} \;=\; 1. \end{aligned}$$

Chapter 15
Conditional Effect Functions

In chapter 14 we treated $E^{X=x}(Y|Z)$, the Z-conditional expectation of Y with respect to the $(X=x)$-conditional probability measure $P^{X=x}$. There we already noted that, if the values of X represent treatment conditions, then $E^{X=x}(Y|Z)$ refers to the Z-conditional expectation of Y given treatment x. If X is dichotomous with values 0 and 1, then the difference $g_1(Z) := E^{X=1}(Y|Z) - E^{X=0}(Y|Z)$ is the Z-conditional effect function of X. Its values $g_1(z)$ are the effects of X on Y given the value z of Z.

From a methodological point of view, conditioning on a (possibly) multidimensional random variable Z serves at least two purposes. The *first* is to *control for potential confounders*, i. e., to make sure that the effects of X on Y cannot be explained by the dependencies of X and Y on Z. The *second* purpose is to *obtain more specific effects* that are more informative than unconditional effects. Knowing such conditional effects we can choose individualized treatments rather than giving all individuals the same treatment that possibly is suboptimal for all.

In this chapter we introduce the concepts of conditional intercept functions and conditional effect functions and consider these functions for the parametrizations of the conditional expectation $E(Y|X,Z)$ that have been treated in chapters 12 and 13.

15.1 Assumptions and Definitions

In section 14.1 we treated three examples that motivated introducing the conditional expectations $E^{X=x}(Y|Z)$. These examples also motivate this chapter on conditional effect functions. In Examples 14.1 and 14.2, the conditional effect function $g_1(U)$ and each of its values $g_1(u)$ are uniquely defined, whereas in Example 14.3 this is not the case. While in the latter example the value $g_1(Ann)$ is uniquely defined, the value $g_1(Joe)$ is not, and we can choose any real number as the value $g_1(U)(\omega)$ if $U(\omega) = Joe$ and still $g_0(U)$ and $g_1(U)$ satisfy

$$E(Y|X,U) \underset{P}{=} g_0(U) + g_1(U)\cdot X, \tag{15.1}$$

for all versions $E(Y|X,U) \in \mathscr{E}(Y|X,U)$. In other words, the function $g_1(U)$ specified in Example 14.3 is only one out of infinitely many versions of such a conditional effect function satisfying Equation (15.1), and even the expectations $E[g_1(U)]$ of these conditional effect functions are not necessarily identical. There-

fore, we have to introduce an assumption that guarantees that the values of the functions $g_0(U)$ and $g_1(U)$ are not arbitrary and that the expectations $E[g_0(U)]$ and $E[g_1(U)]$ are uniquely defined. Instead of the person variable U used in the examples with Joe and Ann, now we choose Z as a random variable with respect to which we consider conditional intercept and effect functions.

Throughout this chapter we refer to the following assumptions and notation.

Notation and Assumptions 15.1
$X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$, $Y\colon(\Omega,\mathscr{A},P)\to(\overline{\mathbb{R}},\overline{\mathscr{B}})$, and $Z\colon(\Omega,\mathscr{A},P)\to(\Omega'_Z,\mathscr{A}'_Z)$ are random variables, where $E(Y^2)<\infty$ and $Var(Y)>0$. Furthermore, $X\colon(\Omega,\mathscr{A},P)\to(\Omega'_X,\mathscr{A}'_X)$ is discrete with $P(X\in\{x_0,x_1,\ldots,x_n\})=1$ and $P(X=x_i)>0$, for all $i=0,1,\ldots,n$.

According to Remark 6.26, $E(Y^2)<\infty$ implies that $E(Y)$ is finite as well.

Remark 15.2 (Two Additional Assumptions) Two additional assumptions are often used in this chapter. The first is

$$P(X=x_i\,|\,Z)\underset{P}{>}0,\quad\forall\, i=0,1,\ldots,n. \tag{15.2}$$

According to Corollary 14.46, this assumption is equivalent to P-uniqueness of the conditional expectations $E^{X=x_i}(Y|Z)$. Remember, P-uniqueness of the conditional expectations $E^{X=x_i}(Y|Z)$ implies that different versions of $E^{X=x_i}(Y|Z)$ do not only have identical distributions with respect to the measure $P^{X=x_i}$ but also with respect to P. Therefore they also have identical expectations and variances with respect to the measure P. Also note that each of the conditions that are equivalent to (15.2) such as absolute continuity

$$P\underset{\sigma(Z)}{\ll}P^{X=x_i},\quad\forall\, i=0,1,\ldots,n, \tag{15.3}$$

and

$$E(V_i)\;=\;E(V_i^*),\quad\forall\, V_i,V_i^*\in\mathscr{E}^{X=x_i}(Y|Z),\quad\forall\, i=0,1,\ldots,n \tag{15.4}$$

(see again Cor. 14.46) are also equivalent to P-uniqueness of the conditional expectations $E^{X=x_i}(Y|Z)$.

The second additional assumption used in this chapter is finiteness of the second moments of the conditional expectations $E^{X=x_i}(Y|Z)$, i. e.,

$$E\big[E^{X=x_i}(Y|Z)^2\big]<\infty,\quad\forall\, i=0,1,\ldots,n. \tag{15.5}$$

◁

15.2 Conditional Intercept Function and Effect Functions

Theorem 15.3 (Existence of Conditional Intercept and Effect Functions) *Let the assumptions 15.1 hold. Then there are an $E(Y|X,Z) \in \mathscr{E}(Y|X,Z)$ and, for all $i = 0,1,\ldots,n$, real-valued $E^{X=x_i}(Y|Z) \in \mathscr{E}^{X=x_i}(Y|Z)$ such that*

$$E(Y|X,Z) \;=\; g_0(Z) + \sum_{i=1}^{n} g_i(Z) \cdot 1_{X=x_i}, \tag{15.6}$$

with

$$g_0(Z) \;:=\; E^{X=x_0}(Y|Z), \tag{15.7}$$

and

$$g_i(Z) \;:=\; E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z), \quad \forall\, i = 1,\ldots,n. \tag{15.8}$$

(Proof p. 432)

Referring to Equation (15.6), now we can define the conditional intercept function and the conditional effect functions as follows.

Definition 15.4 (Conditional Intercept Function and Effect Functions) *Let the assumptions 15.1 hold as well as* (15.2) *and* (15.5)*. Then the function $g_0\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ is called the Z-conditional intercept function, and, for all $i = 1,\ldots,n$, the function $g_i\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$, the Z-conditional effect function of x_i vs. x_0 on Y pertaining to the version $E(Y|X,Z) \in \mathscr{E}(Y|X,Z)$ in Equation* (15.6)*.*

Remark 15.5 (The Functions g_i vs. the Functions $g_i(Z)$) Note that the functions $g_i(Z)$, $i = 0,1,\ldots,n$, denote the compositions of Z and g_i. Because Z is a random variable on $(\Omega, \mathscr{A}, P)$, the compositions $g_i(Z)$ are random variables on $(\Omega, \mathscr{A}, P)$ as well. In contrast, the functions $g_i\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$, $i = 0,1,\ldots,n$, are random variables on the probability space $(\Omega'_Z, \mathscr{A}'_Z, P_Z)$, where P_Z denotes the distribution of Z. ◁

Remark 15.6 (P-Uniqueness of the Conditional Intercept and Effect Functions) Suppose that the assumptions of Definition 15.4 hold. Then all measurable functions $g_0^*, g_1^*, \ldots, g_n^*\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ with

$$\left(g_0^*(Z) + \sum_{i=1}^{n} g_i^*(Z) \cdot 1_{X=x_i}\right) \in \mathscr{E}(Y|X,Z)$$

satisfy

$$g_i^*(Z) \underset{P}{=} g_i(Z), \quad \forall\, i = 0,1,\ldots,n. \tag{15.9}$$

This property is called *P-uniqueness* of the conditional intercept and effect functions. Among other things, it implies that the expectations, the variances, and the second moments of the functions $g_i(Z)$, $i = 0,1,\ldots,n$, are uniquely defined. ◁

Remark 15.7 (Finite Second Moments of the Effect Functions) Assuming (15.5) implies $E_Z({g_i}^2) < \infty$, for all $i = 0,1,\ldots,n$. According to Equation (6.13),

$$E_Z({g_i}^2) \;=\; E[g_i(Z)^2], \quad \forall\, i = 0,1,\ldots,n\,. \tag{15.10}$$

Hence, assuming (15.5) implies

$$E[g_i(Z)],\, E[g_i(Z)^2] \;<\; \infty, \quad \forall\, i = 0,1,\ldots,n, \tag{15.11}$$

(see Rem. 7.1). ◁

Remark 15.8 (Partial Conditional Expectation $E(Y|X,Z{=}z)$) Let the assumptions of Definition 15.4 hold. Then Equation (15.6) immediately implies

$$E(Y|X,Z{=}z) \;=\; g_0(z) + \sum_{i=1}^{n} g_i(z)\cdot 1_{X=x_i} \tag{15.12}$$

for the partial conditional expectation $E(Y|X,Z{=}z)$ (see Def. 14.26). This equation justifies the terminology introduced in Definition 15.4. ◁

Remark 15.9 ($(Z{=}z)$-Conditional Intercept and Effects) Under the assumptions of Definition 15.4: If $z \in \Omega_Z'$ with $P(Z{=}z) > 0$, then Equation (15.12) and (14.26) imply that there is an $E^{Z=z}(Y|X) \in \mathscr{E}^{Z=z}(Y|X)$ with

$$E^{Z=z}(Y|X) \;=\; E(Y|X,Z{=}z) \;=\; g_0(z) \,+\, \sum_{i=1}^{n} g_i(z)\cdot 1_{X=x_i}. \tag{15.13}$$

If $P(Z{=}z) > 0$, the number $g_0(z)$ is called the $(Z{=}z)$-*conditional intercept* and $g_i(z)$ the $(Z{=}z)$-*conditional effect* of x_i vs. x_0 on Y, where $i = 1,\ldots,n$. Equation (15.13) and Corollary 12.30 imply that $g_0(z)$ is the intercept and $g_i(z)$, $i = 1,\ldots,n$, are the partial regression coefficients pertaining to a linear parametrization of the X-conditional expectation of Y with respect to the measure $P^{Z=z}$, provided that Y has finite second moments with respect to $P^{Z=z}$ and that the matrix of the covariances with respect to $P^{Z=z}$ of the indicators $1_{X=x_1},\ldots,1_{X=x_n}$ is regular. According to Lemma 12.37 and Remark 12.38 this is the case if $P^{Z=z}(X{=}x_i) > 0$ for all $i = 1,\ldots,n$. ◁

Remark 15.10 (Versions of $E^{Z=z}(Y|X)$) Under the assumptions of Definition 15.4, Remark 14.8 and Equation (15.13) immediately imply

$$V_z \underset{P^{Z=z}}{=} g_0(z) \,+\, \sum_{i=1}^{n} g_i(z)\cdot 1_{X=x_i}, \quad \forall\, V_z \in \mathscr{E}^{Z=z}(Y|X)\,. \tag{15.14}$$

◁

Remark 15.11 (Partial Conditional Expectation $E(Y|X{=}x,Z)$) Let assumptions of Definition 15.4 hold. Then Equation (15.12) immediately implies

$$E(Y|X{=}x,Z) \;=\; g_0(Z) \,+\, g_i(Z) \tag{15.15}$$

for the partial conditional expectation $E(Y|X{=}x,Z)$ (see Def. 14.26). ◁

Remark 15.12 (Conditional Expectation $E^{X=x}(Y|Z)$) Under the assumptions of Definition 15.4: If $x \in \Omega_X'$ with $P(X=x) > 0$, then there is an $E^{X=x}(Y|Z) \in \mathscr{E}^{X=x}(Y|Z)$ with

$$E^{X=x}(Y|Z) \;=\; E(Y|X=x,Z) \;=\; g_0(Z) \,+\, g_i(Z). \tag{15.16}$$

◁

Remark 15.13 (Versions of $E^{X=x}(Y|Z)$) Remark 14.8 and Equation (15.16) immediately imply

$$V_x \underset{P^{X=x}}{=} g_0(Z) \,+\, g_i(Z), \quad \forall\, V_x \in \mathscr{E}^{X=x}(Y|Z). \tag{15.17}$$

◁

15.3 Adjusted Conditional Effect Functions

Remark 15.14 (Methodological Background) In Definition 15.4 we considered the Z-conditional intercept function g_0 as well as the Z-conditional effect functions g_i, $i = 1,\ldots,n$, where $Z = (Z_1,\ldots,Z_m)$ can be an m-dimensional random variable consisting of m uni-dimensional random variables, such as *pretest*, *sex*, *educational status*, *body mass index*, and *blood type*. As mentioned before, conditioning on a (possibly multi-dimensional) random variable Z serves at least two purposes. The *first* is to *control for potential confounders*, and the *second* is to obtain more specific effects that are more informative than unconditional effects.

However, the Z-conditional effects described by the effect functions g_i might be relatively fine-grained and one may wish to re-aggregate them. The most radical re-aggregation is to consider the (unconditional) expectation (the 'average') of the Z-conditional effects. However, we might also be interested in the conditional expected values of the Z-conditional effect functions given *sex* = *male* and given *sex* = *female*, or in the conditional expected values of the Z-conditional effects given various scores of the *pretest*.

A particular way of re-aggregation is coarsening the effects. Knowing coarsened effects is important, e. g., if, knowing the Z-conditional effect functions (e. g., from a previous study), we want to conduct an as much as possible individualized treatment, but are not able to assess all components $Z_1,\ldots,Z_m$ of Z, but just Z_1 or just Z_1 and Z_2, for instance.

In Definition 15.15 we consider re-aggregating the Z-conditional intercept function and the Z-conditional effect functions to a W-conditional effect function that is adjusted for Z, where W is another random variable. If $Z = (Z_1,\ldots,Z_m)$ with $m \geq 2$, then $W = Z_1$ and $W = (Z_1, Z_2)$ are examples in case. In these two cases, W would be Z-measurable. However, Z-measurability of W is not absolutely required.

Note that re-aggregating the Z-conditional effect functions is not equivalent to ignoring Z and considering W instead. Conditioning only on W instead of

conditioning on Z we might miss the purpose of controlling and adjusting for important confounders. Re-aggregation as described in Definition 15.15, still controls for Z, and with it we control for potential confounders contained in Z. Re-aggregating the Z-conditional effect functions, only the second purpose of conditioning suffers, because re-aggregation means to obtain less informative and less individualized conditional effects. However, as mentioned above, applying a known probabilistic model that involves Z, sometimes we are only able to assess $W = Z_1$ or $W = (Z_1, Z_2)$, but not all components $Z_1, \ldots, Z_m$ of Z. ◁

Definition 15.15 (Adjusted Conditional Effect Function)
Let the assumptions of Definition 15.4 hold and let $W\colon (\Omega, \mathscr{A}, P) \to (\Omega'_W, \mathscr{A}'_W)$ *be a random variable. Then, for each* $i = 1, \ldots, n$, *the conditional expectation* $E\left[g_i(Z) \middle| W\right]$ *is called a* Z*-adjusted* W*-conditional effect function of* x_i *vs.* x_0 *on* Y. *Furthermore, for each* $i = 1, \ldots, n$, *the expectation* $E\left[g_i(Z)\right]$ *is called the* Z*-adjusted effect of* x_i *vs.* x_0 *on* Y *or the average of the* Z*-conditional effects of* x_i *vs.* x_0 *on* Y *or simply the average effect of* x_i *vs.* x_0, *if there is no ambiguity about* Y *and* Z.

The values of $E\left[g_i(Z) \middle| W\right]$ are the $(W{=}w)$-conditional effects of x_i vs. x_0 on Y that are adjusted for Z.

Remark 15.16 (Z-Adjusted W-Conditional Expectation of Y given $X{=}x_i$) Using the definition of the functions $g_i(Z)$ [see Eq. (15.8)] yields

$$\begin{aligned} & E\left[g_i(Z) \middle| W\right] \\ &= E\left[E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z) \middle| W\right] \\ &= E\left[E^{X=x_i}(Y|Z) \middle| W\right] \; - \; E\left[E^{X=x_0}(Y|Z) \middle| W\right], \quad \forall\, i = 1, \ldots, n. \end{aligned} \tag{15.18}$$

A conditional expectation $E\left[E^{X=x_i}(Y|Z) \middle| W\right]$ is called a *Z-adjusted W-conditional expectation of Y given $X{=}x_i$*, where $i = 0, 1, \ldots, n$. ◁

15.4 Mean Independence of the Conditional Effect Functions

In the following theorem we presume that the assumptions of Definition 15.4 hold, which implies Equation (15.6) for the conditional expectation $E(Y|X,Z)$. Now we consider the implications for the conditional expectation $E(Y|X)$ if we additionally assume

$$E\left[E^{X=x_i}(Y|Z) \middle| X\right] \underset{P}{=} E\left[E^{X=x_i}(Y|Z)\right], \quad \forall\, i = 0, 1, \ldots, n. \tag{15.19}$$

Note that (15.19) follows from independence of X and Z [see Box 10.2 (vi)] and it implies

$$E\left[g_i(Z) \middle| X\right] \underset{P}{=} E\left[g_i(Z)\right], \quad \forall\, i = 0, 1, \ldots, n, \tag{15.20}$$

for the conditional intercept and conditional effect functions.

Theorem 15.17 (Mean Independence and Average Effects)
Let the assumptions of Definition 15.4 hold. Then there is a version $E(Y|X) \in \mathscr{E}(Y|X)$ *with*

$$E(Y|X) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot 1_{X=x_i} \tag{15.21}$$

with

$$\beta_0 = E(Y|X=x_0) \tag{15.22}$$

and

$$\beta_i = E(Y|X=x_i) - E(Y|X=x_0), \tag{15.23}$$

If additionally Equations (15.19) *hold, then*

$$\beta_i = E[g_i(Z)], \quad \forall\, i = 0,1,\ldots,n. \tag{15.24}$$

(Proof p. 433)

Remark 15.18 (Uniqueness of Regression Coefficients) According to Equation (15.22) the regression coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are uniquely defined. The crucial assumption is $P(X \in \{x_0, x_1, \ldots, x_n\}) = 1$ and $P(X=x_i) > 0$, for all $i = 0,1,\ldots,n$. ◁

Remark 15.19 (Independence of X and Z) If we assume that the Equations (15.19) hold, then the regression coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are the expectations of the functions $g_i(Z)$, $i = 0,1,\ldots,n$ [see Eq. (15.24)]. As mentioned before, Equations (15.19) follow from $X \perp\!\!\!\perp_P Z$ [see Box 10.2 (vi)]. ◁

Remark 15.20 (The Role of Randomization) From a methodological point of view it should be noted that independence of a treatment variable X and a variable Z can be created by randomly assigning the observational unit (e. g., a person) to one of the treatment conditions, provided that Z represents a pretreatment variable. Random assignment creates independence of X and *all* pretreatment variables Z. Examples for such pretreatment variables are the person variable U (see Table 14.1 for a concrete example) as well as any function of U such as sex, race, and any other attribute of persons prior to treatment. ◁

According to Corollary 15.21, Equation (15.2) — and therefore P-uniqueness of the functions $g_i(Z)$ — follows from independence of X and Z, if we presume $P(X=x_i) > 0$ for all $i = 0,1,\ldots,n$.

Corollary 15.21 (Independence of X and Z and P-Uniqueness)
Let the assumptions of Definition 15.4 hold. If $X \perp\!\!\!\perp_P Z$, *then* $E^{X=x_i}(Y|Z)$ *and the functions* $g_i(Z)$ *in Equations* (15.7) *and* (15.8) *are P-unique, for all* $i = 0,1,\ldots,n$.

(Proof p. 433)

15.5 Conditional Logit Effect Functions

In the previous sections of this chapter we studied the conditional intercept and conditional effect functions $g_0, g_1, \ldots, g_n$. Now we consider the special case, in which Y is dichotomous with values 0 and 1. In this case $E(Y|X,Z)$ is also called a conditional probability and is also denoted by $P(Y=1|X,Z)$. If Y is dichotomous there are also Z-conditional logit intercept and logit effect functions, denoted f_i, and we can consider the functions g_i and f_i at the same time.

Theorem 15.22 (Existence of the Conditional Logit Effect Functions) *Let the assumptions 15.1 hold, let Y be dichotomous with values 0 and 1, and suppose there is a $P(Y=1|X,Z) \in \mathscr{P}(Y=1|X,Z)$ with $0 < P(Y=1|X,Z) < 1$. Then there are a version $P(Y=1|X,Z) \in \mathscr{P}(Y=1|X,Z)$, measurable functions $g_0, g_1, \ldots, g_n \colon (\Omega_Z', \mathscr{A}_Z') \to (\mathbb{R}, \mathscr{B})$ with finite second moments, measurable functions $f_0, f_1, \ldots, f_n \colon (\Omega_Z', \mathscr{A}_Z') \to (\mathbb{R}, \mathscr{B})$, and, for all $i = 0, 1, \ldots, n$, a real-valued $P^{X=x_i}(Y=1|Z) \in \mathscr{P}^{X=x_i}(Y=1|Z)$ such that*

$$P(Y=1|X,Z) = g_0(Z) + \sum_{i=1}^{n} g_i(Z) \cdot 1_{X=x_i} \tag{15.25}$$

$$= \frac{\exp\left[f_0(Z) + \sum_{i=1}^{n} f_i(Z) \cdot 1_{X=x_i}\right]}{1 + \exp\left[f_0(Z) + \sum_{i=1}^{n} f_i(Z) \cdot 1_{X=x_i}\right]} \tag{15.26}$$

with

$$g_0(Z) := P^{X=x_0}(Y=1|Z) \tag{15.27}$$

$$= \frac{\exp\left[f_0(Z)\right]}{1 + \exp\left[f_0(Z)\right]} \tag{15.28}$$

and

$$g_i(Z) := P^{X=x_i}(Y=1|Z) - P^{X=x_0}(Y=1|Z) \tag{15.29}$$

$$= \frac{\exp\left[f_0(Z) + f_i(Z)\right]}{1 + \exp\left[f_0(Z) + f_i(Z)\right]} - \frac{\exp\left[f_0(Z)\right]}{1 + \exp\left[f_0(Z)\right]}. \tag{15.30}$$

(Proof p. 433)

Remark 15.23 (P-Uniqueness of the Functions $f_i(Z)$) Note that Remark 15.5 still applies to the functions $g_i(Z)$, and it also applies to the functions $f_i(Z)$, $i = 0, 1, \ldots, n$. If we additionally assume (15.2), then, for all $i = 0, 1, \ldots, n$, the conditional probabilities $P^{X=x_i}(Y=1|Z)$ and the functions $g_i(Z)$ are P-unique. If we assume (15.2), then the functions $f_i(Z)$ are P-unique as well, i. e., all measurable functions $f_0^*, f_1^*, \ldots, f_n^* : (\Omega_Z', \mathscr{A}_Z') \to (\mathbb{R}, \mathscr{B})$ with

$$\frac{\exp\left(f_0^*(Z) + \sum_{i=1}^{n} f_i^*(Z) \cdot 1_{X=x_i}\right)}{1 + \exp\left(f_0^*(Z) + \sum_{i=1}^{n} f_i^*(Z) \cdot 1_{X=x_i}\right)} \in \mathscr{P}(Y=1|X,Z)$$

satisfy

$$f_i^*(Z) \underset{P}{=} f_i(Z), \quad \forall\, i = 0,1,\ldots,n, \tag{15.31}$$

(see Exercise 15-1). ◁

Definition 15.24 (Conditional Logit Intercept and Effect Functions)
Let the assumptions of Theorem 15.22 as well as (15.2) *hold and suppose that the second moments of the functions f_i, $i = 0,1,\ldots,n$, are finite. Then the function $f_0\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ is called the Z-conditional logit intercept function, and, for all $i = 1,\ldots,n$, the function $f_i\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$, the Z-conditional logit effect function of x_i vs. x_0 on Y pertaining to the version $P(Y{=}1|X,Z) \in \mathscr{P}(Y{=}1|X,Z)$ in Equation* (15.26).

Remark 15.25 (Partial Conditional Probability $P(Y{=}1|X,Z{=}z)$) Let the assumptions of Theorem 15.22 hold. Then Equation (15.26) immediately implies

$$P(Y{=}1|X,Z{=}z) \;=\; \frac{\exp\left(f_0(z) + \sum_{i=1}^n f_i(z)\cdot 1_{X=x_i}\right)}{1+\exp\left(f_0(z) + \sum_{i=1}^n f_i(z)\cdot 1_{X=x_i}\right)} \tag{15.32}$$

for the partial conditional probability $P(Y{=}1|X,Z{=}z)$ (see Rem. 14.27). This equation justifies the terminology introduced in Definition 15.24. ◁

Remark 15.26 ($(Z{=}z)$-Conditional Logit Intercept and Effects) If the assumptions of Theorem 15.22 hold and $z \in \Omega'_Z$ with $P(Z{=}z) > 0$, then Equation (15.32) and (14.26) imply that there is a $P^{Z=z}(Y{=}1|X) \in \mathscr{P}^{Z=z}(Y{=}1|X)$ with

$$P^{Z=z}(Y{=}1|X) \;=\; P(Y{=}1|X,Z{=}z) \;=\; \frac{\exp\left(f_0(z) + \sum_{i=1}^n f_i(z)\cdot 1_{X=x_i}\right)}{1+\exp\left(f_0(z) + \sum_{i=1}^n f_i(z)\cdot 1_{X=x_i}\right)}. \tag{15.33}$$

Equation (15.33) and Corollary 13.18 imply that $f_0(z)$ is the intercept and $f_i(z)$, $i = 1,\ldots,n$, are the coefficients pertaining to a linear logistic parametrization of $P^{Z=z}(Y{=}1|X)$, provided that the matrix of the covariances of the random variables $X_1,\ldots,X_n$ with respect to the measure $P^{Z=z}$ is regular (see Th. 13.18). According to Lemma 12.37 and Remark 12.38 this is the case if $P^{Z=z}(X{=}x_i) > 0$ for all $i = 1,\ldots,n$.

Under the assumptions of Theorem 15.22, Remark 14.8 and Equation (15.33) immediately imply

$$V_z \underset{P^{Z=z}}{=} \frac{\exp\left(f_0(z) + \sum_{i=1}^n f_i(z)\cdot 1_{X=x_i}\right)}{1+\exp\left(f_0(z) + \sum_{i=1}^n f_i(z)\cdot 1_{X=x_i}\right)}, \quad \forall\, V_z \in \mathscr{P}^{Z=z}(Y{=}1|X). \tag{15.34}$$

◁

Remark 15.27 (Conditional Probability $P^{X=x_i}(Y{=}1|Z)$) Remark 14.8 and Equation (15.25) immediately imply

$$V_0 \underset{P^{X=x}}{=} \frac{\exp\big(f_0(Z)\big)}{1+\exp\big(f_0(Z)\big)}, \quad \forall\, V_0 \in \mathscr{P}^{X=0}(Y{=}1|Z), \tag{15.35}$$

and, for all $i = 1,\dots,n$,

$$V_i \underset{P^{X=x_i}}{=} \frac{\exp\big(f_0(Z)+f_i(Z)\big)}{1+\exp\big(f_0(Z)+f_i(Z)\big)}, \quad \forall\, V_i \in \mathscr{P}^{X=x_i}(Y{=}1|Z)\,. \tag{15.36}$$

◁

Remark 15.28 (Second Moments of the Functions $f_i(Z)$) Note that the second moments of the functions $f_i(Z)$ are not necessarily finite unless assumptions are introduced that are additional to those mentioned in Theorem 15.22. ◁

Remark 15.29 (Conditional Log Odds Functions) In terms of conditional probabilities, the Z-conditional logit intercept function can be written

$$f_0(Z) \;=\; \ln\left[\frac{P^{X=x_0}(Y{=}1|Z)}{1-P^{X=x_0}(Y{=}1|Z)}\right]. \tag{15.37}$$

Hence, f_0 may also be called the *Z-conditional log odds function* of x_0. Similarly,

$$f_0(Z)+f_i(Z) \;=\; \ln\left[\frac{P^{X=x_i}(Y{=}1|Z)}{1-P^{X=x_i}(Y{=}1|Z)}\right], \quad \forall\, i = 1,\dots,n \tag{15.38}$$

(see Exercise 15-2). The function $f_0+f_i\colon (\Omega'_Z,\mathscr{A}'_Z)\to(\mathbb{R},\mathscr{B})$ satisfying $(f_0+f_i)(Z) = f_0(Z)+f_i(Z)$ is called the *Z-conditional log odds function* of x_i. ◁

Remark 15.30 (Conditional Log Odds Ratio Functions) Equations (15.37) and (15.38) immediately imply

$$f_i(Z) \;=\; \ln\left[\frac{P^{X=x_i}(Y{=}1|Z)}{1-P^{X=x_i}(Y{=}1|Z)}\right] - \ln\left[\frac{P^{X=x_0}(Y{=}1|Z)}{1-P^{X=x_0}(Y{=}1|Z)}\right] \tag{15.39}$$

$$= \ln\left[\frac{\dfrac{P^{X=x_i}(Y{=}1|Z)}{1-P^{X=x_i}(Y{=}1|Z)}}{\dfrac{P^{X=x_0}(Y{=}1|Z)}{1-P^{X=x_0}(Y{=}1|Z)}}\right], \quad \forall\, i = 1,\dots,n, \tag{15.40}$$

for the Z-conditional logit effect functions. Hence, $f_i(Z)$, $i = 1,\dots,n$, is the difference between the Z-conditional log odds functions of x_i and x_0, respectively [see Eq. (15.39)]. Referring to Equation (15.40), f_i is also called the *Z-conditional log odds ratio function of x_i and x_0*. ◁

Remark 15.31 (Conditional Odds Ratio Functions) The exponential function of $f_i(Z)$ is

$$\exp[f_i(Z)] \;=\; \frac{\dfrac{P^{X=x_i}(Y{=}1|Z)}{1-P^{X=x_i}(Y{=}1|Z)}}{\dfrac{P^{X=x_0}(Y{=}1|Z)}{1-P^{X=x_0}(Y{=}1|Z)}}, \quad \forall\, i = 1,\dots,n\,. \tag{15.41}$$

The composite function $\exp(f_i)\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ is called the *Z-conditional odds ratio function of x_i and x_0*. ◁

Remark 15.32 (Conditional Risk Ratio Functions) Another closely related function is

$$k_i(Z) := \frac{P^{X=x_i}(Y=1|Z)}{P^{X=x_0}(Y=1|Z)}, \quad \forall\, i = 1,\dots,n. \tag{15.42}$$

The function $k_i\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ satisfying (15.42) is called the *Z-conditional risk ratio function of x_i and x_0*. ◁

Remark 15.33 (Four Conditional Effect Functions) Hence, under the assumptions of Theorem 15.22, we considered four different Z-conditional effect functions: g_i, f_i, $\exp(f_i)$, and k_i. They all describe Z-conditional effect functions of x_i compared to x_0 on Y on different scales. ◁

Remark 15.34 (Independence, Regression Coefficients, and Adjusted Effects) A property of the conditional effect functions g_i not shared by the other effect functions is the following one: If

$$E\big[g_i(Z)\,\big|\,X\big] \;=\; E\big[g_i(Z)\big], \quad \forall\, i = 0,1,\dots,n,$$

[see Eq. (15.20)] — and, according to Remark 15.19 this follows from independence of X and Z — then according to Theorem 15.17,

$$E[g_i(Z)] \;=\; \beta_i, \quad i = 0,1,\dots,n, \tag{15.43}$$

holds for the expectations of the functions $g_i(Z)$, where β_i are the parameters in the equation

$$P(Y=1|X) \;=\; \beta_0 + \sum_{i=1}^{n} \beta_i \cdot 1_{X=x_i}. \tag{15.44}$$

Note that, under the assumptions of Theorem 15.22, there are always coefficients $\beta_0, \beta_1, \dots, \beta_n \in \mathbb{R}$ such that Equation (15.44) holds. From a methodological point of view, this means that we can ignore pretreatment variables Z in a randomized experiment with treatment variable X and dichotomous response variable Y and still interpret the coefficient β_i as the average effect $E[g_i(Z)]$ of x_i compared to x_0 on Y, where $i = 1,\dots,n$ (see Rem. 15.20). ◁

Remark 15.35 (Mean Independence and Logit Effect Functions) Note that the analog to Equation (15.43) *does not hold* for the expectation of the logit effect functions. That is, although, under the assumptions of Theorem 15.22, there are always coefficients $\lambda_0, \lambda_1, \dots, \lambda_n \in \mathbb{R}$ such that

$$P(Y=1|X) \;=\; \frac{\exp(\lambda_0 + \sum_{i=1}^{n} \lambda_i \cdot 1_{X=x_i})}{1 + \exp(\lambda_0 + \sum_{i=1}^{n} \lambda_i \cdot 1_{X=x_i})} \tag{15.45}$$

holds, neither Equation (15.20) nor independence of X and Z imply that λ_i is identical to $E[f_i(Z)]$. In fact it is even possible that λ_i is negative and $E[f_i(Z)]$ is positive (see Example 15.37). ◁

Example 15.36 (Joe and Ann With Random Assignment–continued) In Example 14.1, we considered the conditional expectation

$$\begin{aligned} P(Y{=}1|X,U) &= .2 + .5\cdot 1_{U=Joe} + .2\cdot X - .1\cdot 1_{U=Joe}\cdot X \\ &= (.2+.5\cdot 1_{U=Joe}) + (.2-.1\cdot 1_{U=Joe})\cdot X \\ &= g_0(U)+g_1(U)\cdot X. \end{aligned} \tag{15.46}$$

with intercept function g_0 satisfying

$$g_0(U) = P^{X=0}(Y{=}1|U) = .2 + .5\cdot 1_{U=Joe}$$

and effect function g_1 satisfying

$$g_1(U) = P^{X=1}(Y{=}1|U) - P^{X=0}(Y{=}1|U) = .2 - .1\cdot 1_{U=Joe}$$

[see Eq. (14.1)]. The average of the U-conditional effects of X on Y is

$$E[g_1(U)] = E\big(.2 - .1\cdot 1_{U=Joe}\big) = .2 - .1\cdot E(1_{U=Joe}) = .2 - .1\cdot .5 = .15.$$

In this example this average effect is equal to β_1 in the equation

$$P(Y{=}1|X) = \beta_0 + \beta_1\cdot X = .45 + .15\cdot X, \tag{15.47}$$

which illustrates Equations (15.24), (15.43), and (15.44).

The same random experiment is also considered in Example 13.22. There, we computed

$$P(Y{=}1|X) = \frac{\exp(\lambda_0+\lambda_1\cdot X)}{1+\exp(\lambda_0+\lambda_1\cdot X)} \approx \frac{\exp(-.201+.606\cdot X)}{1+\exp(-.201+.606\cdot X)} \tag{15.48}$$

and

$$\begin{aligned} P(Y{=}1|X,U) &\underset{P}{=} \frac{\exp\big((\gamma_0+\gamma_2\cdot 1_{U=Ann})+(\gamma_1+\gamma_3\cdot 1_{U=Ann})\cdot X\big)}{1+\exp\big((\gamma_0+\gamma_2\cdot 1_{U=Ann})+(\gamma_1+\gamma_3\cdot 1_{U=Ann})\cdot X\big)} \\ &\approx \frac{\exp\big((.847-2.234\cdot 1_{U=Ann})+(.539+.442\cdot 1_{U=Ann})\cdot X\big)}{1+\exp\big((.847-2.234\cdot 1_{U=Ann})+(.539+.442\cdot 1_{U=Ann})\cdot X\big)}, \end{aligned} \tag{15.49}$$

showing that the logit of $P(Y{=}1|X)$ is $f_0(U)+f_1(U)\cdot X$ with logit intercept function

$$f_0(U) = \gamma_0+\gamma_2\cdot 1_{U=Ann} \approx .847 - 2.234\cdot 1_{U=Ann}$$

and logit effect function

$$f_1(U) = \gamma_1+\gamma_3\cdot 1_{U=Ann} = .539 + .442\cdot 1_{U=Ann}.$$

Note that

$$\lambda_1 \approx .606$$

Table 15.1. Joe and Ann: Reversed Average Logit Effect

15-3

Outcomes ω				Observables			Conditional expectations						
Unit	Treatment	Success	$P(\{\omega\})$	Person variable U	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$P(X=1\|U)$	$E^{X=0}(Y\|U)$	$E^{X=1}(Y\|U)$	Effect function $g_1(U)$	Logit effect function $f_1(U)$
(*Joe*,	*no*,	−)	.067	*Joe*	0	0	.732	.732	.5	.732	.984	.252	3.114
(*Joe*,	*no*,	+)	.183	*Joe*	0	1	.732	.732	.5	.732	.984	.252	3.114
(*Joe*,	*yes*,	−)	.004	*Joe*	1	0	.984	.626	.5	.732	.984	.252	3.114
(*Joe*,	*yes*,	+)	.246	*Joe*	1	1	.984	.626	.5	.732	.984	.252	3.114
(*Ann*,	*no*,	−)	.067	*Ann*	0	0	.732	.732	.5	.732	.268	−.464	−2.010
(*Ann*,	*no*,	+)	.183	*Ann*	0	1	.732	.732	.5	.732	.268	−.464	−2.010
(*Ann*,	*yes*,	−)	.183	*Ann*	1	0	.268	.626	.5	.732	.268	−.464	−2.010
(*Ann*,	*yes*,	+)	.067	*Ann*	1	1	.268	.626	.5	.732	.268	−.464	−2.010

$$\neq E[f_1(U)] \approx .539 + .442 \cdot E(1_{U=Ann}) \approx .760.$$

Hence, although X and U are independent, the logit effect λ_1 of X in the logistic parametrization of $P(Y{=}1|X)$ is *not* equal to the expectation of the logit effect function $f_1(U)$ of the logit in the logistic parametrization of $P(Y{=}1|X,U)$.

From a methodological point of view this means that random assignment of a unit to one of two treatment conditions—which creates independence of a treatment variable X and the person variable U—does not imply that the slope λ_1 of the logit in the logistic parametrization of $P(Y{=}1|X)$ can be interpreted as the expectation of the U-conditional logit effect function of treatment variable X on Y. As mentioned before, it is even possible that λ_1 is negative and $E[f_i(Z)]$ is positive (see Example 15.37).

In examples in which $f_1(U) = \gamma_1$ is a constant, this implies that $\lambda_1 = \gamma_1$ does not follow from independence of X and U. In contrast, compare the corresponding invariance property formulated in Theorem 12.42 for a linear parametrization of a conditional expectation. ◁

Example 15.37 (Joe and Ann: Reversed Average Logit Effect) Table 15.1 displays an example in which the coefficient λ_1 in the equation

$$P(Y{=}1|X) = \frac{\exp(\lambda_0 + \lambda_1 \cdot X)}{1 + \exp(\lambda_0 + \lambda_1 \cdot X)} \tag{15.50}$$

is negative, whereas the expectation $E[f_1(U)]$ of the function $f_1(U)$ in the equation

$$P(Y{=}1|X,U) \underset{P}{=} \frac{\exp\big(f_0(U)+f_1(U)\cdot X\big)}{1+\exp\big(f_0(U)+f_1(U)\cdot X\big)} \tag{15.51}$$

is positive, although X and U are independent. The coefficients of Equation (15.50) are $\lambda_0 \approx 1.005$ and $\lambda_1 \approx -.490$, whereas the expectation of the conditional logit effect function is $E[f_1(U)] \approx .552$ (see Exercises 15-3 and 15-4). ◁

Example 15.38 (Joe and Ann With Self-Selection – continued) In this example, the results are the same as in 15.36. However, whereas in Example 15.36 the average effect $E[g_1(U)] = .15$ is equal to β_1 in Equation (15.47), this is not the case in Example 15.38. Here, $\beta_1 = -.18$ [see Eq. (14.6)]. ◁

Example 15.39 (No Treatment for Joe – continued) In Example 14.3, the function $g_1(U)$ is not P-unique although the intercept function $g_0(U)$ and the conditional effect $g_1(Ann)$ are uniquely defined. In this example, there are (infinitely) many functions $g_1(U)$ satisfying Equation (15.1). ◁

15.6 Proofs

Proof of Theorem 15.3

For all $i = 0, 1, \ldots, n$,

$$\begin{aligned}
E(Y^2) < \infty \;&\Rightarrow\; E(Y) < \infty && \text{[Rem. 6.26]}\\
&\Rightarrow\; E^{X=x_i}(Y) < \infty && \text{[Rem. 14.6]}\\
&\Rightarrow\; \exists \text{ a real-valued } E^{X=x_i}(Y|Z) \in \mathscr{E}^{X=x_i}(Y|Z). && \text{[Box 10.2 (x)]}
\end{aligned}$$

According to Theorem 10.9 and Corollary 10.23, finiteness of $E(Y)$ also implies that there is a $g(X,Z) \in \mathscr{E}(Y|X,Z)$ such that, for real-valued versions $E^{X=x_i}(Y|Z)$

$$\begin{aligned}
g(X,Z) &\underset{P}{=} \sum_{i=0}^{n} E^{X=x_i}(Y|Z)\cdot \mathbf{1}_{X=x_i} && \text{[(14.28)]}\\
&\underset{P}{=} E^{X=x_0}(Y|Z)\cdot \mathbf{1}_{X=x_0} + \sum_{i=1}^{n} E^{X=x_i}(Y|Z)\cdot \mathbf{1}_{X=x_i}\\
&\underset{P}{=} E^{X=x_0}(Y|Z) - \sum_{i=1}^{n} E^{X=x_0}(Y|Z)\cdot \mathbf{1}_{X=x_i} + \sum_{i=1}^{n} E^{X=x_i}(Y|Z)\cdot \mathbf{1}_{X=x_i} && \text{[(5.32)]}\\
&\underset{P}{=} E^{X=x_0}(Y|Z) + \sum_{i=1}^{n} \big[E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z)\big]\cdot \mathbf{1}_{X=x_i}.
\end{aligned}$$

Because the function on the right-hand side of the last equation is (X,Z)-measurable and P-equivalent to $g(X,Z) \in \mathscr{E}(Y|X,Z)$ (see Th. 2.57), it is an element of $\mathscr{E}(Y|X,Z)$. Defining the specific version

$$E(Y|X,Z) := E^{X=x_0}(Y|Z) + \sum_{i=1}^{n} \big[E^{X=x_i}(Y|Z) - E^{X=x_0}(Y|Z)\big]\cdot \mathbf{1}_{X=x_i}$$

completes the proof.

Proof of Theorem 15.17

The existence of a version $E(Y|X) \in \mathscr{E}(Y|X)$ for which Equations (15.21) to (15.23) hold has already been proved in Theorem 12.36. Hence, we only have to prove Equation (15.24). For all versions $E(Y|X) \in \mathscr{E}(Y|X)$,

$$\begin{aligned}
E(Y|X) &\underset{P}{=} E[E(Y|X,Z)|X] && \text{[Box 10.2 (v)]}\\
&\underset{P}{=} E\left[g_0(Z) + \sum_{i=1}^{n} g_i(Z)\cdot \mathbf{1}_{X=x_i} \,\middle|\, X\right] && \text{[(15.6)]}\\
&\underset{P}{=} E[g_0(Z)|X] + \sum_{i=1}^{n} E\left[g_i(Z)\cdot \mathbf{1}_{X=x_i} \,\middle|\, X\right] && \text{[Box 10.2 (xv)]}\\
&\underset{P}{=} E[g_0(Z)|X] + \sum_{i=1}^{n} E[g_i(Z)|X]\cdot \mathbf{1}_{X=x_i} && \text{[Box 10.2 (xiv), } \sigma(\mathbf{1}_{X=x_i}) \subset \sigma(X)\text{]}\\
&\underset{P}{=} E[g_0(Z)] + \sum_{i=1}^{n} E[g_i(Z)]\cdot \mathbf{1}_{X=x_i}. && \text{[(15.20)]}
\end{aligned}$$

According to Lemma 12.37 (i), the second moments of the indicators $\mathbf{1}_{X=x_1},\ldots,\mathbf{1}_{X=x_n}$ are finite. Because in Definition 15.4 we also assume (15.5), which implies that the second moments of the functions g_i, $i=0,1,\ldots,n$, are finite, we can apply Rule (xiv) of Box 10.2. Furthermore, Remark 7.1 and (15.11) imply $E[g_i(Z)\cdot X_i] < \infty$. Hence, we can apply Rule of (xv) Box 10.2.

According to Remark 2.18, the right-hand side of the last equation above is X-measurable. Because it is P-equivalent to all versions $E(Y|X) \in \mathscr{E}(Y|X)$, the existence of a specific version $E(Y|X)$ satisfying Equations (15.21) and (15.24) follows from (10.8).

Proof of Corollary 15.21

According to Box 10.2 (vi) and Equations (6.5) and (10.4), independence of X and Z implies $P(X{=}x_i|Z) \underset{P}{=} P(X{=}x_i)$. Because $P(X{=}x_i) > 0$, we can conclude $P(X{=}x_i|Z) \underset{P}{>} 0$. Furthermore, $P(X{=}x_i|Z) \underset{P}{>} 0$ is equivalent to P-uniqueness of $E^{X=x_i}(Y|Z)$ [see Cor. 14.46 (a) and (c)]. According to Box 14.1 (viii), this implies P-uniqueness of the functions $g_i(Z)$, $i=0,1,\ldots,n$.

Proof of Theorem 15.22

By definition, $P(Y{=}1|X,Z) = E(\mathbf{1}_{Y=1}|X,Z)$. Hence, the existence of measurable functions $g_0, g_1, \ldots, g_n\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ satisfying Equations (15.25), (15.27), and (15.29) has already been proved in Theorem 15.3. Assuming $0 < P^{X=x_i}(Y{=}1|Z) < 1$ implies finiteness of the second moments of the $P^{X=x_i}(Y{=}1|Z)$, which in turn implies finiteness of the second moments of the $g_i = P^{X=x_i}(Y{=}1|Z) - P^{X=x_0}(Y{=}1|Z)$. In order to show that there are measurable functions $f_0, f_1, \ldots, f_n\colon (\Omega'_Z, \mathscr{A}'_Z) \to (\mathbb{R}, \mathscr{B})$ satisfying Equation (15.26) we define

$$l_i(Z) := \operatorname{logit}\left[P^{X=x_i}(Y{=}1|Z)\right], \quad \forall\, i = 0,1,\ldots,n, \tag{15.52}$$

using the logit of $P^{X=x_i}(Y{=}1|Z)$ defined by Equation (13.4). Furthermore, we define

$$f_0(Z) := l_0(Z), \tag{15.53}$$

and

$$f_i(Z) := l_i(Z) - l_0(Z), \quad \forall\, i = 1, \ldots, n. \tag{15.54}$$

These definitions and Equations (13.5) then yield

$$P^{X=x_i}(Y=1|Z) = \frac{\exp\big(\operatorname{logit}[P^{X=x_i}(Y=1|Z)]\big)}{1+\exp\big(\operatorname{logit}[P^{X=x_i}(Y=1|Z)]\big)} = \frac{\exp\big(l_i(Z)\big)}{1+\exp\big(l_i(Z)\big)}, \quad \forall\, i = 0, 1, \ldots, n.$$

Hence, for $i = 0$, Equation (15.27) implies

$$g_0 = P^{X=x_0}(Y=1|Z) = \frac{\exp\big(f_0(Z)\big)}{1+\exp\big(f_0(Z)\big)},$$

and Equation (15.29) yields

$$\begin{aligned} g_i(Z) &= P^{X=x_i}(Y=1|Z) - P^{X=x_0}(Y=1|Z) \\ &= \frac{\exp\big(f_0(Z)+f_i(Z)\big)}{1+\exp\big(f_0(Z)+f_i(Z)\big)} - \frac{\exp\big(f_0(Z)\big)}{1+\exp\big(f_0(Z)\big)}, \quad \forall\, i = 1, \ldots, n. \end{aligned}$$

15.7 Exercises

▷ **Exercise 15-1** Prove that, under the assumptions of Theorem 15.22, assuming (15.2) implies that the conditional probabilities $P^{X=x_i}(Y=1|Z)$, their logit transformations $l_i(Z) := \operatorname{logit}\big[P^{X=x_i}(Y=1|Z)\big]$, $i = 0, 1, \ldots, n$, and the differences of these logit transformations are P-unique.

▷ **Exercise 15-2** Prove Equations (15.37) and (15.38).

▷ **Exercise 15-3** Compute the values of the conditional expectations displayed in Table 15.1 from the probabilities of the elementary events in the second column of this table.

▷ **Exercise 15-4** Using the results displayed in Table 15.1, compute the coefficients λ_0 and λ_1 of Equation (15.50) as well as the expectation of the conditional logit effect function f_1.

Solutions

▷ **Solution 15-1** According to Theorem 14.44, assuming (15.2) implies that the conditional probabilities $P^{X=x_i}(Y=1|Z)$ are P-unique, which in turn implies that the functions g_i are P-unique [see (2.36)]. If, for all $i = 0, 1, \ldots, n$, the $P^{X=x_i}(Y=1|Z)$ are P-unique, then the functions $l_i(Z)$ and their differences [see Eqs. (15.53) and (15.54)] are P-unique as well [see (2.34) and Rem. 2.76].

▷ **Solution 15-2** The conditional logit intercept function can be written

$$\begin{aligned} f_0(Z) &= \operatorname{logit}\big(P^{X=x_0}(Y=1|Z)\big) && [(15.28), (13.5)] \\ &= \ln\left[\frac{P^{X=x_0}(Y=1|Z)}{1-P^{X=x_0}(Y=1|Z)}\right]. && [(13.4)] \end{aligned}$$

Furthermore,

$$\begin{aligned} f_0(Z) + f_i(Z) &= \text{logit}\big(P^{X=x_i}(Y{=}1|Z)\big) && [(15.30), (13.5)] \\ &= \ln\left[\frac{P^{X=x_i}(Y{=}1|Z)}{1-P^{X=x_i}(Y{=}1|Z)}\right]. && [(13.4)] \end{aligned}$$

▷ **Solution 15-3** All these values are displayed in Table 15.1.

▷ **Solution 15-4** Using Equations (13.5), (13.7), as well as the conditional probabilities $P(Y{=}1|X{=}x)$ displayed in Table 15.1 yields

$$\begin{aligned} \lambda_0 &\approx \ln\big(.732/(1-.732)\big) \approx 1.005, \\ \lambda_0 + \lambda_1 &\approx \ln\big(.984/(1-.984)\big) \approx .515, \end{aligned}$$

and $\lambda_1 = .515 - 1.005 = -.490$. The expectation of the logit effect function f_1 is

$$\begin{aligned} E_U(f_1) &= E[f_1(U)] && [(6.15)] \\ &\approx 3.114 \cdot .5 - 2.010 \cdot .5 = .552. && [\text{Table } 15.1] \end{aligned}$$

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Bauer, H. (1996). *Probability theory*. Berlin, Germany and New York, NY: de Gruyter.

Bauer, H. (2001). *Measure and integration theory*. Berlin, Germany and New York, NY: de Gruyter.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York, NY: Wiley-Interscience.

Ellis, R., & Gulick, D. (2006). *Calculus* (6^{th} ed.). Mason, OH: Thomson.

Elstrodt, J. (2007). *Maß- und Integrationstheorie [Measure and integration theory]* (4th ed.). Berlin, Germany: Springer.

Fisz, M. (1963). *Probability theory and mathematical statistics*. New York, NY: Wiley.

Fristedt, B., & Gray, L. (1997). *A modern approach to probability theory*. Boston, MA: Birkhäuser.

Georgii, H.-O. (2008). *Stochastics – Introduction to probability and statistics*. Berlin, Germany: de Gruyter.

Hoffmann-Jørgensen, J. (1994). *Probability with a view toward statistics* (Vol. 1). New York, NY: Chapman & Hall.

Horn, R. A., & Johnson, C. R. (1991). *Matrix analysis*. Cambridge, UK: Cambridge University Press.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). New York, NY: Wiley.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York, NY: Wiley.

Kheyfits, A. (2010). *A primer in combinatorics*. Berlin/New York: De Gruyter.

Klenke, A. (2008). *Probability theory – A comprehensive course*. London, England: Springer.

Kolmogoroff, A. N. (1933/1977). *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Reprinted ed.). Berlin, Germany: Springer.

Kolmogorov, A. N. (1933/1977). *Grundbegriffe der Wahrscheinlichkeitsrechnung [Foundations of the Theory of Probability]* (Reprinted ed.). Berlin, Germany: Springer.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd ed.; N. Morrison, Trans.). New York, NY: Chelsea.

McCullagh, P., & Nelder, J. A. (1989). *Monographs on statistics and applied probability: Vol. 37. Generalized linear models* (2nd ed.; D. R. Cox, D. V. Hinkley, N. Reid, D. B. Rubin, & D. V. Silverman, Eds.). Chapman & Hall.

Michel, H. (1978). *Maß- und Integrationstheorie I. [Measure and integration theory I].* Berlin: VEB Deutscher Verlag der Wissenschaften.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.

Rosen, K. (2012). *Discrete mathematics and its applications* (7^{th} ed.). New York, NY: McGraw-Hill.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression [Probability and regression].* Berlin, Germany: Springer.

Tong, Y. L. (1990). *The multivariate normal distribution.* New York, NY: Springer.

Author Index

Agresti, A., 336, 515

Balakrishnan, N., 257, 515
Bauer, H., VI, 16, 24, 91–93, 98, 101, 102, 104, 106, 107, 109, 171, 205, 247, 291, 312, 409, 474, 497, 515
Billingsley, P., VI, 515

Ellis, R., VIII, 104, 185, 379, 489, 515
Elstrodt, J., 106, 515

Fisz, M., 256, 257, 259, 261, 490, 515
Fristedt, B., VI, 515

Georgii, H.-O., 4, 7, 12, 143, 252–254, 256, 258, 261, 262, 515
Gray, L., VI, 515
Gulick, D., VIII, 104, 185, 379, 489, 515

Hoffmann-Jørgensen, J., VI, 515
Horn, R. A., 362, 515

Johnson, C. R., 362, 515
Johnson, N. L., 250, 257, 259, 515

Kemp, A. W., 250, 515
Kheyfits, A., 8, 515
Klenke, A., VI, 12, 13, 15, 22, 25, 27, 50, 51, 60, 62, 64, 80, 86, 90, 104, 105, 109, 110, 213, 248, 252, 300, 301, 317, 367, 474, 515
Kolmogoroff, A. N., V, VII, 515
Kolmogorov, A. N., 127, 289, 515
Kotz, S., 250, 257, 515

McCullagh, P., 370, 376, 516
Michel, H., 11, 516

Nelder, J. A., 370, 376, 516

Rao, C. R., 360, 516
Rosen, K., VIII, 516

Steyer, R., VII, 516

Tong, Y. L., 262, 516

Subject Index

absolute continuity, 109
  and P-equivalence, 161
  and independence, 169
  of a conditional-probability measure, 139
  of marginal distributions, 177
absolute value function, 63
additivity of a measure
  σ-additivity, 17
  finite additivity, 17
adjusted conditional effect function, 424
adjusted effect, 424
adjusted logit effect, 427
almost all, 158
almost everywhere, 100
average effect, 424

Bayes' theorem
  for densities, 491
  for events, 137
Bernoulli distribution, 242
Bernoulli variable, 242
binomial coefficient, 243
binomial distribution, 244
  approximation by Poisson distribution, 246
bivariate normal distribution, 261
Borel σ-algebra on $\mathbb{R}$, 12
Borel σ-algebra on $\mathbb{R}^n$, 13, 15
Borel set, 12

cardinality, 8, 21
Cartesian product, 3
Cauchy distribution, 257
Cauchy-Schwarz inequality, 218
central F-distribution, 258
central t-distribution, 256
central limit theorem, 253
central moment, 205, 206
χ^2-distribution, 255
closed interval, 6
CNS-uniqueness, 475
codomain of a mapping, 41
coefficient of determination
  of a conditional expectation, 325
  of a linear quasi-regression, 227
common null set uniqueness, 475
complement of a set, 5
composition of two mappings, 58
conditional correlation
  given a value of a random variable, 337
conditional covariance
  given a σ-algebra, 329
  given a random variable, 330
  given a value of a random variable, 331
  rules of computation, 334
conditional density, 491, 492
conditional distribution, 468
  and P-equivalence, 477
  and conditional independence, 483
  and conditional independence given a random variable, 482
  and independence, 473
  and mean independence, 484
  existence, 474
  of a random variable given a σ-algebra, 468
  of a random variable given a random variable, 468
  of a random variable given a value of a random variable, 470

conditional distribution function, 490
conditional effect function, 421
conditional expectation
w.r.t. a conditional-probability measure, 389–391
and conditional densities, 492, 493
and joint distribution, 306
and mean-squared error, 301
coefficient of determination, 325
convergence, 299, 300
discrete, 278
given a σ-algebra, 290
given a random variable, 290
linear parametrization, 350
marginalization, 487
monotonicity, 299
rules of computation, 297, 298
uniqueness w.r.t. a probability measure, 292
conditional expectation value
and conditional distribution, 486
given a value of a random variable, 272, 303
given an event, 272
of a composition, 274
rules of computation, 276, 277
conditional independence
of random variables given an event, 171
conditional independence given a σ-algebra
and conditional mean independence, 450
characterizations, 452
family of ∩-stable set systems, 455
family of events, 454
family of random variables, 455
family of set systems, 454
notation, 445
of two events, 439
of two random variables, 442
of two set systems, 441
properties, 448
conditional independence given a random variable
and conditional distributions, 482
family of events, 454
family of random variables, 455
family of set systems, 454
of two events, 440
of two random variables, 442
of two set systems, 442
properties, 449
conditional independence given a value of a random variable
of two events, 441
conditional independence given an event
family of ∩-stable set systems, 146
family of set systems, 146
of two events, 144
conditional intercept function, 421
conditional logit effect function, 427
conditional logit intercept function, 427
conditional mean independence, 306, 307
and conditional independence, 450
and independence, 308
from a σ-algebra , 306
from random variable, 307
conditional probability
w.r.t. a conditional-probability measure, 389–391
given a σ-algebra , 290
given a random variable, 290
given a value of a random variable, 303
given an event, 132
of an event given a value of a random variable, 273
of an event given an event, 273
conditional probability density, *see* conditional density
conditional standard deviation
given a σ-algebra, 330
given a random variable, 330
given a value of a random variable, 331
conditional variance
given a σ-algebra, 330
given a random variable, 330
given a value of a random variable, 331
rules of computation, 335
conditional-probability measure, 138, 478
constant mapping, 48
continuity of a measure from above, 25
continuity of a measure from below, 25
continuous random variable, 183
without expectation, 200
continuous uniform distribution, 129, 250

convergence of conditional expectations, 299, 300
convergence of integrals, 103
correlation, 220
  and slope of a linear quasi-regression, 221
  invariance under linear transformations, 221
correlational independence
  mean independence, 311
countable intersection, 6
countable union, 5
countably generated σ-algebra , 11
counting measure, 21
covariance, 216
  rules of computation, 218
covariance matrix, 224
  rules of computation, 226

density, *see* aso probability density183
  and probability function, 177
  of the χ^2-distribution, 255
  of the F-distribution, 258
  of the t-distribution, 256
  of the bivariate normal distribution, 261
  of the Cauchy distribution, 257
  of the multivariate normal distribution, 260
  of the standard normal distribution, 108
  of the univariate normal distribution, 252
density of a measure w.r.t. another measure, 106
dichotomous function, 48
dichotomous random variable, 156
Dirac measure, 21
discrete conditional expectation given a random variable, 278
discrete conditional probability given a random variable, 278
discrete distribution, 172
discrete random variable, 172
  without expectation, 199
discrete regression, 279
discrete uniform distribution, 241
disjoint sets, 4
distribution, 154
  and distribution function, 179, 181
  Bernoulli, 242
  binomial, 244
  χ^2, 255
  Cauchy, 257
  discrete uniform, 241
  F, 258
  geometric, 248
  multivariate normal, 260
  of a composition, 155
  of an indicator, 155
  Poisson, 246
  t, 256
  univariate normal, 252
  univariate standard normal, 252
distribution function, 179
  and distribution, 179, 181
  and independence, 182
  joint and marginal distribution function, 182
  of a binomial distribution, 244
  of the geometric distribution, 250
  of the Poisson distribution, 246
  of the standard normal distribution, 252
  of the univariate normal distribution, 252
domain of a mapping, 41
dominated convergence of integrals, 104
Dynkin system, 16

effect function, 421
elementary event, 128
elementary function, 83
equivalence w.r.t. a measure, 65
  necessary and sufficient conditions, 107
equivalence w.r.t. a probability measure
  of random variables, 158
  of two factorizations of a conditional expectation, 302
equivalence class w.r.t. a measure, 66
equivalence of probability densities w.r.t. a measure, 176
equivalence relation, 66
event, 127
existence
  of a conditional distribution, 474

of a conditional expectation given a σ-algebra , 291
of an expectation, 197
expectation
w.r.t. P, 197
w.r.t. a conditional-probability measure, 198, 273
of a distribution, 202
of a random matrix, 224
of a random variable, 197
of a random variable with a countable number of real values, 199
of a random variable with a finite number of real values, 198
of a random variable with density, 200
of a random vector, 223
of a sample mean, 204
of an indicator, 198
of the product of random variables under independence, 205
of the product of two random variables, 217
rules of computation, 204
expectation of a random matrix
rules of computation, 225

factorial of an integer, 244
factorization
equivalence of two versions w.r.t. a probability measure, 302
of a conditional expectation given a random variable, 301
uniqueness, 302
factorization of a composition, 60
F-distribution, 258
final σ-algebra, 54
finite additivity of a measure, 17
finite intersection, 6
finite measure, 23
finite union, 5
Fubini's theorem, 110

Γ-function, 255
generating system
of a σ-algebra, 9
of a product σ-algebra, 16
geometric distribution, 248
distribution function, 250

identically distributed random variables, 155
identification
linear logit regression, 376
linear regression, 358
identity
of mappings, 46
of random variables, 158
identity mapping, 48
i. i. d., 168
image measure, 68
under a step function, 70
image of a set under a mapping, 41, 43
increasing sequence of nonnegative step functions, 91
independence
and P-equivalence, 169
and absolute continuity, 169
and conditional mean independence, 308
and distribution function, 182
and probability densities, 184
and product measure, 169
conditional, *see* conditional independence
family of σ-algebras, 143
family of events, 142
family of random variables, 167
family of set systems, 142
measurable mappings, 170
of ∩-stable set systems, 143
of n random variables, 168
of a constant and a set of events, 170
of a random variable and a set system, 168
of an event and a set system, 143
of three events, 142
of two events, 142
of two random variables, 166
independent and identically distributed, 168
indicator, 20
infinite measure, 23
integrable, 94
integral
w.r.t. a Dirac measure, 92
w.r.t. a finite weighted sum of measures, 101

w.r.t. a measure with density, 107
w.r.t. a weighted sum of Dirac measures, 101
w.r.t. a weighted sum of measures, 101
w.r.t. an image measure, 102
w.r.t. the Lebesgue measure, 94
of μ-equivalent functions, 99
of a constant, 86
of a function with a finite number of values, 103
of a measurable function, 94
of a nonnegative measurable function, 92
of a nonnegative step function, 86
of a positive measurable function, 98
over a null set, 98
over a subset, 86, 96
over the union of two sets, 97
intercept
of a simple linear quasi-regression, 215
of a simple linear regression, 352
intercept function, 421
intersection
of countably many sets $A_1, A_2, \ldots$, 6
of finitely many sets $A_1, \ldots, A_n$, 6
invariance of regression coefficients, 364
inverse image of a set under a mapping, 41

joint distribution, 164
joint distribution function, 181

Lebesgue integral and Riemann integral, 104
Lebesgue measure, 22
linear combination of two functions, 63
linear logistic regression, 376
linear logit parametrization, 372
identification, 376
linear logit regression, 376
linear parametrization
differences between means as coefficients, 361
identification, 358
means as coefficients, 360
linear parametrization of a conditional expectation, 350
linear quasi-regression, 213
and linear regression, 356
and regression, 301
equivalent characterizations, 219
linear regression, 359
and linear quasi-regression, 356
and normal distribution, 360
identification, 358
linearity
of the integral, 96
link function, 370
log odds, 374
log odds ratio, 374
log-odds, 370
logit effect function, 427
logit intercept function, 427
logit of a conditional probability, 370
logit transformation, 369

marginal density, 178
marginal distribution, 165
marginal distribution function, 491
marginal probability density, 491
marginalization of a conditional expectation, 487
mean centered random variable, 206
mean independence, 307
and conditional distributions, 484
and correlational independence, 311
mean squared error, 213, 227
measurability w.r.t. a mapping, 57
measurable mapping, 46
measurable set, 5
measurable space, 6
measure, 17
measure space, 17
measure with density, 106
mixture of probability measures, 131
moment, 206
moment of a numerical random variable, 205
monotonicity
of a conditional expectation, 299
of a measure, 19
of an integral, 93, 100
MSE, mean squared error function, 213
μ-almost everywhere, 100
multiple correlation, 327
multiple linear logistic regression, 376

multiple linear quasi-regression, 227
multiple linear regression, 359
multiplication rule for probabilities, 133
multivariate Bernoulli distribution, 243
multivariate Bernoulli variable, 242
multivariate mapping, 55
multivariate normal distribution, 260
multivariate random variable, 164

negative part of a function, 63
nonnegative step function, 83
    and counting measure, 88
    and Dirac measure, 87
    normal representation, 84
normal distribution, 252
    and linear regression, 360
    bivariate, 261
    density, 252
    multivariate, 260
    standard, 252
    univariate, 252
normal representation of a nonnegative step function, 84
null set, 27
    integral over a null set, 98
null-set equivalence of two measures, 109

odds ratio, 375
outcome of a random experiment, 128

pairwise independence, 142
P-almost all, 158
partial $(X, Z{=}z)$-conditional expectation, 393
partial correlation, 336
partition of a set, 10
P-equivalence, 158
    and absolute continuity, 161
    and conditional distributions, 477
    and distributions, 160
    and independence, 169
    of compositions, 160
    of random variables, 158
P^B-expectation, 198
pointwise convergence, 91
Poisson distribution, 246
    approximation of the binomial distribution, 246
    distribution function, 246
positive part of a function, 63
power set, 6
probability density, *see* aso density176
    of a continuous real-valued random variable, 183
    of a probability measure, 175
    of a random variable, 176
probability function, 172
    and density, 177
    of a marginal distribution, 174
probability measure, 127
    with density, 175
probability of an event, 127
probability space, 127
product σ-algebra, 14, 56
product measure, 24
    and independence, 169
product set, 3
projection mapping, 56
P-uniqueness, 292
    of a conditional expectation, 292
    of a conditional expectation w.r.t. conditional-probability measure, 403
P^B-uniqueness, 392

quantile, 180
quantile function, 180
    and inverse distribution function, 180
quasi-integrable, 94
quasi-regression
    linear, 213

Radon-Nikodym derivative, 109
Radon-Nikodym theorem, 109
    and probability density, 176
random sample, 168
random variable, 153
    continuous, 183
    discrete, 172
    discrete real-valued, 173
    numerical, 154
    real-valued, 154
regressand, 279, 301
regression, 301
    and linear quasi-regression, 301
    discrete, 279
regression coefficients, 359
    invariance, 364

regressor, 279, 301
residual
  w.r.t. a conditional expectation, 322
  w.r.t. a linear quasi-regression, 219, 228
residual w.r.t. a conditional expectation
  rules of computation, 323
restriction of a measure, 23
Riemann integral and Lebesgue integral, 104
risk ratio, 375
rules of computation
  for a residual w.r.t. a conditional expectation, 323
  for conditional covariances, 334
  for conditional expectation values given a value of a random variable, 277
  for conditional expectation values given an event, 276
  for conditional expectations given a σ-algebra , 297
  for conditional expectations given a random variable, 298
  for conditional variances, 335
  for covariance matrices, 226
  for covariances, 218
  for expectations of random matrices, 225
  for measures, 19
  for probabilities, 129
  for the expectation of a random variable, 204
  for variances, 207

set system, 4
σ-additivity of a measure, 17
σ-algebra, 5
  Borel, 12
  countably generated, 11
  final, 53
  generated by a composition, 59
  generated by a family of mappings, 55
  generated by a mapping, 52
  generated by a multivariate mapping, 55
  generated by a set system, 9
  generated by an indicator, 52
  trivial, 28
σ-field, *see* σ-algebra
σ-finite measure, 23
σ-subadditivity, 19
sign function, 64
simple function, 83
simple linear logistic regression, 376
simple linear regression, 352, 359
  identification of the intercept, 359
  identification of the slope, 360
  intercept, 352
  slope, 352
singleton, 6
skewness, 209
slope
  of a linear quasi-regression and correlation, 221
  of a simple linear quasi-regression, 215
  of a simple linear regression, 352
stability of a set system w.r.t. intersections, ∩-stability, 15
standard deviation, 206
  of the sample mean, 208
standard error
  of the sample mean, 208
standard normal distribution, 252
step function, 49
  normal representation, 84
strictly diagonally dominant matrix, 362

t-distribution, 256
theorem
  B. Levi, 103
  Bayes' theorem for densities, 491
  Bayes' theorem for events, 137
  central limit, 253
  Fubini, 110
  of total probability, 136
  Radon-Nikodym, 109
  transformation theorem for a conditional expected value, 274
  transformation theorem for an integral, 102
  transformation theorem of an expectation, 201
trace σ-algebra, 8
trace of a set system, 8
transformation theorem

for a conditional expected value, 274
for an expectation, 201
for an integral, 102
for conditional distributions, 486
triple-wise independence, 142
trivial σ-algebra, 28
trivial σ-algebra w.r.t. a measure, 28

uncorrelated random variables, 217
uncountable union, 6
uniform distribution, 250
union
of countably many sets, 5
of finitely many sets, 5
uniqueness
of a conditional expectation w.r.t. a probability measure, 292
of a conditional expectation given a σ-algebra , 291
of a factorization, 302
univariate normal distribution, 252
density, 252
distribution function, 252

variance, 206
of an indicator, 206
of the sample mean, 208
rules of computation, 207

Z-transformation, 208

List of Symbols

$\neg$	Aa not, IX
$\vee$	Ab or, IX
$\wedge$	Ab and, IX
$\Rightarrow$	Aca implies, IX
$\Leftrightarrow$	Acb equivalent to, IX
$\exists$	Ada there is (synonymously, there exists), IX
$\forall$	Adb for all, IX
$\in$	Sacaaa element of, IX
1_A	SadaaaVmqaaa indicator (function) of the set A, 20
$\cup$	Saeaaa union of sets, IX
$\cap$	Safaaa intersection of sets, IX
$\setminus$	Sahaaa set difference, IX
$A^c := \Omega \setminus A$	Sahcaa complement of a set A with respect to a set Ω, IX
$\subset$	Sajaaa subset or equal, IX
$\times$	Sajbaa Cartesian product or product set, IX
$\otimes$	Sajbba product σ-algebra, 14
$\otimes$	Sajbba product measure, 24
$\circ$	Sajbbb composition of two mappings, 58
$\odot$	SajbbbUfoaaa measure with density, 106
$\underset{\mu}{=}$	SajbbbVmaaaa μ-equivalence of mappings, 65
$\underset{\mu}{<}$	SajbbbVmhaaa smaller than except for a μ-null set, 67
$\underset{\mu}{>}$	SajbbbVmkaaa greater than except for a μ-null set, 67
$\underset{\mu}{\leq}$	SajbbbVmmaaa smaller than or equal except for a μ-null set, 67
$\underset{\mu}{\geq}$	SajbbbVmoaaa greater than or equal except for a μ-null set, 67
$\underset{P}{=}$	SajbbbVmogaa P-equivalence of random variables, 158
$\underset{\mu\text{-a.a.}}{=}$	SajbbbVmsaaa equal for μ-almost all $\omega \in \Omega$, 65
$\underset{\mathscr{A}}{\ll}$	SajbbbVpuaaa absolute continuity of a measure with respect to another measure, 109
$\underset{\mathscr{A}}{\approx}$	SajbbbVpvaaa null-set equivalence of two measures, 109

$\underset{P}{\perp\!\!\!\perp}$	SajbbbVtvaaa independence with respect to the probability measure P, 142
$[a,b]$	Sajbca closed interval between real numbers a and b, 6
$]a,b]$	Sajbcc half-open interval including b but not a, 12
$\{x\}$	Sajbcg singleton, i. e., the set that contains x as the only element, 6
$\emptyset$	Sajbcj empty set, IX
$(A_i, i \in I)$	Sajhcg family of sets A_i, $i \in I$, where the index set I may be finite, countable, or uncountable, IX
$\bigcup_{i\in I} A_i$	Sakaaa union of the sets A_i, $i \in I$, IX
$\bigcap_{i\in I} A_i$	Samaaa intersection of the sets A_i, $i \in I$, IX
$\bigcup_{i=1}^{n} A_i$	Samcaa union of finitely many sets $A_1,\dots,A_n$, 5
$\bigcup_{i=1}^{\infty} A_i$	Sameaa union of countably many sets $A_1, A_2,\dots$, 5
$\bigcap_{i=1}^{n} A_i$	Samfba intersection of finitely many sets $A_1,\dots,A_n$, 6
$\bigcap_{i=1}^{\infty} A_i$	Samfca intersection of countably many sets $A_1, A_2,\dots$, 6
$\bigtimes_{i=1}^{n} A_i$	Samfga Cartesian product or product set of n sets A_i, IX
$\lim_{n\to\infty} a_n$	Sbaaaa limit of a sequence $a_1, a_2,\dots$ of real numbers, IX
$\sum_{i=1}^{n} a_i$	Scaaaa sum of the numbers $a_1,\dots,a_n$, IX
$\sum_{i=1}^{\infty} a_i$	Sccaaa $\lim_{n\to\infty} \sum_{i=1}^{n} a_i$, IX
$\prod_{i=1}^{n} a_i$	Sdaaaa product of the real numbers $a_1,\dots,a_n$, IX
$\mathscr{A}_1 \otimes \dots \otimes \mathscr{A}_n$	Sdchaa product σ-algebra of the $\mathscr{A}_1,\dots,\mathscr{A}_n$, 14
$\bigotimes_{i=1}^{n} \mathscr{A}_i$	Seaaaa product σ-algebra of the $\mathscr{A}_1,\dots,\mathscr{A}_n$, 14
$\int f\,d\mu$	Siaaaa integral of a measurable function $f\colon (\Omega,\mathscr{A},\mu) \to (\overline{\mathbb{R}},\overline{\mathscr{B}})$ with respect to the measure μ, 94
$\int_A f\,d\mu$	Sigaaa integral of a measurable function $f\,(\Omega,\mathscr{A},\mu) \to (\mathbb{R},\mathscr{B})$ with respect to the measure μ over a subset A of Ω, 96
$\int_a^b f(x)\,dx$	Simaaa Riemann integral of the function f from a to b, 105
$\mathscr{A}\big\vert_{\Omega_0}$	Svcaaa trace of the set system $\mathscr{A}$ in the set Ω_0, 8
$\mathscr{C}'_f$	Svgaaa final σ-algebra of $\mathscr{C}$ under f, 54
$(\Omega,\mathscr{A})$	Tcaaaa measurable space, 6
$(\Omega,\mathscr{A},\mu)$	Tcfaaa measure space, 17
$(\Omega,\mathscr{A},P)$	Tcjaaa probability space, 127
$f\colon A \to B$	Uaaaaa mapping f assigning to each $a \in A$ one and only one element $b \in B$, IX
$f(A)$	Udaaaa image of the set A under f, 41
$f^{-1}(A')$	Ugaaaa inverse image of the set A' under f, 41
$\{f \in A'\}$	Uhaaaa $:= f^{-1}(A')$, 41
$\{f = \omega'\}$	Ujaaaa $:= f^{-1}(\{\omega'\})$, 42
$(f_i, i \in I)$	Umaaaa family of mappings, 55
$f^{-1}(\mathscr{E}')$	Umaaaa set of the inverse images of all sets $A' \in \mathscr{E}'$. If $\mathscr{E}'$ is a σ-algebra, then $f^{-1}(\mathscr{E}')$ is the σ-algebra generated by f, 49
$g \circ f,\ g(f)$	Umgaaa composition of f and g, 58

$\lvert f \rvert$	Umiaaa absolute value function of f, 63
f^+	Umkaaa positive part of the function f, 63
f^-	Umlaaa negative part of the function f, 63
$f_n \uparrow f$	Umoaaa the sequence $f_1, f_2, \ldots$ of functions converges pointwise and monotonically from below to f, 91
$f\colon (\Omega, \mathscr{A}) \to (\Omega', \mathscr{A}')$	Vaaaaa $(\Omega, \mathscr{A})$, $(\Omega', \mathscr{A}')$ are measurable spaces and the mapping $f\colon \Omega \to \Omega'$ is $(\mathscr{A}, \mathscr{A}')$-measurable, 47
$f\colon (\Omega, \mathscr{A}, \mu) \to \Omega'$	Veaaaa $(\Omega, \mathscr{A}, \mu)$ is a measure space and $f\colon \Omega \to \Omega'$ is a mapping, 65
$f\colon (\Omega, \mathscr{A}, \mu) \to (\Omega', \mathscr{A}')$	Vhaaaa $f\colon \Omega \to \Omega'$ is an $(\mathscr{A}, \mathscr{A}')$-measurable mapping and μ is a measure on $(\Omega, \mathscr{A})$, 65
$f \underset{\mu}{=} g$	Vmaaaa the mappings f and g are μ-equivalent, 65
$f \underset{\mu}{<} g$	Vmhaaa f is smaller than g except for a set A of arguments with $\mu(A) = 0$, 67
$f \underset{\mu}{>} g$	Vmkaaa f is greater than g except for a set A of arguments with $\mu(A) = 0$, 67
$f \underset{\mu}{\leqq} g$	Vmmaaa f is smaller than or equal to g except for a set A of arguments with $\mu(A) = 0$, 67
$f \underset{\mu}{\geqq} g$	Vmoaaa f is greater than or equal to g except for a set A of arguments with $\mu(A) = 0$, 67
$f(\omega) \underset{\mu\text{-a.a.}}{=} g(\omega)$	Vmsaaa $f(\omega) = g(\omega)$, for μ-almost all $\omega \in \Omega$, 65
μ_f	Vptaaa image measure of μ under f, 68
$\nu \underset{\mathscr{A}}{\ll} \mu$	Vpuaaa the measure ν is absolutely continuous with respect to the measure μ, 109
$\nu \underset{\mathscr{A}}{\approx} \mu$	Vpvaaa the measures μ and ν are null-set equivalent, i. e., they are absolutely continuous with respect to each other, 109
$A \underset{P}{\perp\!\!\!\perp} B$	Vtvaaa the events A and B are independent with respect to the probability measure P, 142
$\underset{P}{\perp\!\!\!\perp} (A_i, i \in I)$	Vtvea a family of events A_i that are independent with respect to the probability measure P, 142
$A \underset{P}{\perp\!\!\!\perp} C \mid B$	Vtveda the events A and C are B-conditionally independent with respect to the probability measure P, 144
$\mathscr{E}_1 \underset{P}{\perp\!\!\!\perp} \mathscr{E}_2$	Vtvgaa the set systems $\mathscr{E}_1$ and $\mathscr{E}_2$ are independent with respect to the probability measure P, 143
$\underset{P}{\perp\!\!\!\perp} (\mathscr{E}_i, i \in I)$	Vtvhaa a family of set systems $\mathscr{E}_i$ that are independent with respect to the probability measure P, 143
$A \underset{P}{\perp\!\!\!\perp} \mathscr{E}$	Vtvoaa the event A and the set system $\mathscr{E}_1$ are independent with respect to the probability measure P, 143
$\underset{P}{\perp\!\!\!\perp} (X_i, i \in I)$	Vtvtha a family of random variables X_i that are independent with respect to the probability measure P, 167

$\underset{P}{\perp\!\!\!\perp} X_1,\ldots,X_n$ — Vtvtla the random variables $X_1,\ldots,X_n$ are independent with respect to the probability measure P, 168

$\mathscr{E} \underset{P}{\perp\!\!\!\perp} X$ — Vtvtoa the set system $\mathscr{E}$ and the random variable X are independent with respect to the probability measure P, 168

$X \underset{P}{\perp\!\!\!\perp} (Y_i, i \in I)$ — Vtvtua the random variable X and the (σ-algebra generated by the) family of random variables are independent with respect to the probability measure P, 168

$\underset{P}{\perp\!\!\!\perp} (\mathscr{E}_i, i \in I) \,|\, B$ — Vtvwaa a family of set systems $\mathscr{E}_i$ that are B-conditionally independent with respect to the probability measure P, 146

$\overline{Y}$ — Xtyaaa arithmetic mean (sample mean) of the random variables $Y_1,\ldots,Y_n$, 204

$\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D}, \mathscr{E}$ — Zababe set systems, sometimes σ-algebras , 4

$\mathscr{A}_1 \otimes \ldots \otimes \mathscr{A}_n$ — Zaeabe product σ-algebra of the $\mathscr{A}_1,\ldots,\mathscr{A}_n$, 14

$\mathscr{A}\big|_{\Omega_0}$ — Zahabe trace of the set system $\mathscr{A}$ in the set Ω_0, 8

$\mathscr{B}$ — Zbcaaa Borel σ-algebra on $\mathbb{R}$, 12

$\mathscr{B}_n$ — Zbccaa Borel σ-algebra on $\mathbb{R}^n$, 13

$\overline{\mathscr{B}}$ — Zbcfaa Borel σ-algebra on $\overline{\mathbb{R}}$, 13

$\overline{\mathscr{B}}_n$ — Zbgaaa Borel σ-algebra on $\overline{\mathbb{R}}^n$, 13

$Cov(X,Y)$ — Zccaaa covariance of the random variables X and Y, 216

$Cov(Y_1, Y_2 \,|\, X{=}x)$ — Zccmaa ($X{=}x$)-conditional covariance of Y_1 and Y_2, 331

$Cov^{X=x}(Y_1, Y_2)$ — Zccoaa covariance of Y_1 and Y_2 with respect to the conditional-probability measure $P^{X=x}$, 332

$Cov(Y_1, Y_2 \,|\, \mathscr{C})$ — Zcgaaa a version of the $\mathscr{C}$-conditional covariance of Y_1 and Y_2, 329

$Corr\,(X,Y)$ — Zcgaaa correlation of the random variables X and Y, 220

$Cov(Y_1, Y_2 \,|\, X)$ — Zcgcaa a version of the X-conditional covariance of Y_1 and Y_2, 330

$\mathscr{C}'_f$ — Zchaaa final σ-algebra of $\mathscr{C}$ under f, 54

$Corr\,(Y_1, Y_2 \,|\, X{=}x)$ — Zcllaa ($X{=}x$)-conditional correlation of Y_1 and Y_2, 337

$Corr\,(Y_1, Y_2\,;\mathscr{C})$ — Zclmaa partial correlation of Y_1 and Y_2 given $\mathscr{C}$, 336

$Corr\,(Y_1, Y_2\,; X)$ — Zclnaa partial correlation of Y_1 and Y_2 given X, 336

$\dfrac{d\nu}{d\mu}$ — Zdgaaa Radon-Nikodym density (also Radon-Nikodym derivative) of ν with respect to μ, 109

δ_ω — Zdmaaa Dirac measure at (point) ω, 21

$E(Y)$ — Zeaaaa expectation of the random variable Y, 197

$E(\boldsymbol{x})$ — Zefaaa column vector of expectations, 223

$E^B(Y)$ — Zehaaa expectation of the random variable Y with respect to the probability measure P^B, 198

$E(Y|B)$ — Zeiaaa conditional expectation value of Y given the event B, 272

$E(Y|X{=}x)$ — Zeifaa conditional expectation value of Y given the event $\{X{=}x\}$, also denoted by $E(Y|\{X{=}x\})$, 272

$E(Y|X{=}x)$ — Zeifga ($X{=}x$)-conditional expectation value of Y, 303

$E(\boldsymbol{X})$	Zejaaa matrix of expectations, 224
$E_Y(g)$	Zejaaa expectation of the random variable g with respect to the distribution of the random variable Y, 201
$E_Y^{X=x}(g)$	Zelaaa expectation of g with respect to the distribution $P_Y^{X=x}$, 274
$E^{X=x}(Y)$	Zemaaa expectation of Y with respect to the conditional-probability measure $P^{X=x}$, 273
$E^{\{X=x\}}(Y)$	Zemaaa expectation of Y with respect to the conditional-probability measure $P^{X=x}$, 273
ε	Zepaaa residual of a random variable Y with respect to its $\mathscr{C}$-conditional expectation, 322
ϵ	Zepbaa residual with respect to a (multiple) linear quasi-regression, 228
$E(Y\|X)$	Zeqgaa a version of the X-conditional expectation of Y, 290
$\mathscr{E}(Y\|X)$	Zeqgca set of all versions of the X-conditional expectation of Y, 292
$E(Y\|X_1,\ldots,X_n)$	Zeqgce a version of the conditional expectation of Y given the multivariate regressor $X_1,\ldots,X_n$, 291
$E(Y\|\mathscr{C})$	Zeqgea a version of the $\mathscr{C}$-conditional expectation of Y, 290
$\mathscr{E}(Y\|\mathscr{C})$	Zeqgha set of all versions of the $\mathscr{C}$-conditional expectation of Y, 292
$E^B(Y\|\mathscr{C})$	ZEYCCc a version of the $\mathscr{C}$-conditional expectation of Y with respect to the measure P^B, 389
$\mathscr{E}^B(Y\|\mathscr{C})$	ZEYCCf the set of all versions of the $\mathscr{C}$-conditional expectation of Y with respect to the measure P^B, 389
$E^B(Y\|X)$	ZEYXXc a version of the X-conditional expectation of Y with respect to the measure P^B, 390
$E^B(Y\|X{=}x)$	ZEYXXe an $(X{=}x)$-conditional expectation value of Y with respect to the measure P^B, 396
$\mathscr{E}^B(Y\|X)$	ZEYXXf the set of all versions of the X-conditional expectation of Y with respect to the measure P^B, 390
$E^{Z=z}(Y\|\mathscr{C})$	ZEZCCc a version of the $\mathscr{C}$-conditional expectation of Y with respect to the measure $P^{Z=z}$, 390
$\mathscr{E}^{Z=z}(Y\|\mathscr{C})$	ZEZCCf the set of all versions of the $\mathscr{C}$-conditional expectations of Y with respect to the measure $P^{Z=z}$, 390
$E^{Z=z}(Y\|X{=}x)$	ZEZXXb an $(X{=}x)$-conditional expectation value of Y with respect to the measure P^B, 397
$E(Y\|X,Z{=}z)$	ZEZXXbb a version of the partial $(X,Z{=}z)$-conditional expectation of Y (with respect to the measure P), 393
$E^{Z=z}(Y\|X)$	ZEZXXc a version of the X-conditional expectation of Y with respect to the measure $P^{Z=z}$, 391
$\mathscr{E}^{Z=z}(Y\|X)$	ZEZXXf the family of all versions of the X-conditional expectation of Y with respect to the measure $P^{Z=z}$, 391
$E^{Z=z}(Y\|X{=}x)$	ZEZXXj an $(X{=}x)$-conditional expectation value of Y with respect to the measure $P^{Z=z}$, 398
$E^{Z=z}(Y\|X,W)$	ZEZXXk a version of the (X,W)-conditional expectation of Y with respect to the measure $P^{Z=z}$, 408

$E^{Z=z}(Y|\mathscr{C},\mathscr{D})$ ZEZXXl a version of the $\sigma(\mathscr{C}\cup\mathscr{D})$-conditional expectation of Y with respect to the measure $P^{Z=z}$, 407

$E^{Z=z}(Y|\mathscr{C},Z)$ ZEZXXn a version of the $\sigma[\mathscr{C}\cup\sigma(Z)]$-conditional expectation of Y with respect to the measure $P^{Z=z}$, 407

F_X Zfcaaa distribution function of a real-valued random variable X, 179

$F_{X_1,\ldots,X_n}$ Zfcgaa joint distribution function of $X_1,\ldots,X_n$, 181

$F(x)\Big|_a^b$ Zfgaaa $F(b)-F(a)$, 105

f_X Zfgbaa probability density of a continuous real-valued random variable X, 183

$|f|$ Zfgjaa absolute value function of f, 63

f^+ Zfhaaa positive part of the function f, 63

f^- Zfiaaa negative part of the function f, 63

$(f_i, i\in I)$ Zflaaa family of mappings, 55

$f\odot\mu$ Zfoaaa measure with density f with respect to μ, 106

$g\circ f,\ g(f)$ Zgfaaa composition of f and g, 58

1_A Ziaaaa indicator (function) of the set A, 20

$1_{X\in A'}$ Ziahaa indicator of the event $\{X\in A'\}$, i. e., $1_{X\in A'}=1_{X^{-1}(A')}$, 156

$1_{X=x}, 1_{\{X=x\}}$ Zialaa indicator variable of the event $\{X=x\}$, 273

$\mathscr{I}_1$ Zicaaa set system of all half-open intervals in $\mathbb{R}$, 12

id Zidaaa identity mapping, 48

$\mathscr{I}_n$ Zigaaa set system of all half-open cuboids in $\mathbb{R}^n$, 13

logit Zloaaa logit transformation, 369

logit $[P(Y=1|\mathscr{C})]$ Zlobaa logit of $P(Y=1|\mathscr{C})$, 370

λ, λ_n Zlraaa Lebesgue measure on $(\mathbb{R}^n,\mathscr{B}_n)$, where $\lambda:=\lambda_1$, 22

$\mu,\ \nu$ Zmaaaa general symbols for measures, 17

$\mu_\#$ Zmaaaa counting measure, 21

μ_f Zmfaaa image measure of μ under f, 68

$\mu_1\otimes\ldots\otimes\mu_n$ Zmgaaa product measure of $\mu_1,\ldots,\mu_n$, 24

$\nu \underset{\mathscr{A}}{\ll} \mu$ Zncabe the measure ν is absolutely continuous with respect to the measure μ, 109

$\nu \underset{\mathscr{A}}{\approx} \mu$ Zngabe the measures μ and ν are null-set equivalent, i. e., they are absolutely continuous with respect to each other, 109

$\mathbb{N}$ Znhabe set of all positive integers without zero, i. e., $\mathbb{N}=\{1,2,\ldots,\}$, 5

$\mathbb{N}_0$ Znhcbe set of all nonnegative integers including zero, i. e., $\mathbb{N}_0=\{0,1,2,\ldots,\}$, 4

$(\Omega,\mathscr{A},P)$ Zohaaa probability space, 127

$\mathscr{P}(\Omega)$ Zpbhaa power set of Ω, 6

P Zpcaaa probability measure, 127

$P(A)$ Zpceaa probability of the event A, 127

$P(X=x)$ Zpcgaa probability of the event $\{X=x\}=X^{-1}(\{x\})$, 154

$P(X \in A')$ Zpcjaa probability of the event $\{X \in A'\} = X^{-1}(A')$, 154

$P(A|B)$ Zpcjaa conditional probability of A given B (with respect to the probability measure P), 132

$P(X_1 \in A', X_2 \in B')$ Zpcjca probability of the event $\{X_1 \in A'\} \cap \{X_2 \in B'\}$, 166

$P(A|X{=}x)$ Zpcnca conditional probability of the event A given the event $\{X{=}x\}$ with $P(\{X{=}x\}) > 0$, also denoted by $P(A|\{X{=}x\})$, 273

$P(A|X{=}x)$ Zpcncc conditional probability of the event A given the event $\{X{=}x\}$, also called $(X{=}x)$-conditional probability of A, 303

$P(Y{=}y|\{X{=}x\})$ Zpcncfa conditional probability of the event $\{Y{=}y\}$ given the event $\{X{=}x\}$ with $P(\{X{=}x\}) > 0$, also denoted by $P(Y{=}y|X{=}x)$, 274

$P(Y{=}y|X{=}x)$ Zpcnda conditional probability of the event $\{Y{=}y\}$ given the event $\{X{=}x\}$, also called $(X{=}x)$-conditional probability of $\{Y{=}y\}$ and also denoted by $P(\{Y{=}y\}|X{=}x)$, 303

$P(A|\mathscr{C})$ Zpcnga a version of the $\mathscr{C}$-conditional probability of the event A, 290

$\mathscr{P}(A|\mathscr{C})$ Zpcnha set of all versions of the $\mathscr{C}$-conditional expectation of the event A, 292

$P(A|X)$ Zpcnia conditional probability of the event A (with respect to P) given the discrete random variable X, 278

$P(A|X)$ Zpcniaa a version of the X-conditional probability of the event A, 290

$\mathscr{P}(A|X)$ Zpcnja set of all versions of the X-conditional expectation of the event, 292

$P(Y{=}y|X)$ Zpcnjb a version of the X-conditional probability of the event $\{Y{=}y\}$, 290

$P^{X=x}$ Zpcpaa the $(X{=}x)$-conditional probability measure on $(\Omega, \mathscr{A})$, 161

P_X Zpcpca distribution of the random variable X (with respect to P), 154

$P_{X_1,\ldots,X_n}$ Zpcpha joint distribution of the random variables $X_1, \ldots, X_n$, the distribution of the multivariate random variable $X = (X_1, \ldots, X_n)$, 164

$(P_X)_g$ Zpcqaa image measure of P_X under g, 155

P^B Zpcnaa B-conditional-probability measure, 138

P_X^B Zpfgaa distribution of X with respect to the conditional-probability measure P^B, 161

p_X Zpiaaa probability function of a discrete random variable X, 172

p_{X_1,X_2} Zpifaa probability function of the bivariate random variable $X = (X_1, X_2)$, 174

π_j Zpkaaa jth projection mapping, 56

$P^B(A|\mathscr{C})$ ZPYCCc a version of the $\mathscr{C}$-conditional probability of the event A with respect to the measure P^B, 389

$\mathscr{P}^B(A|\mathscr{C})$ ZPYCCf the set of all versions of the $\mathscr{C}$-conditional probability of the event A with respect to the measure P^B, 389

$P^B(A|X)$ ZPYXXc a version of the X-conditional probability of the event A with respect to the measure P^B, 390

$\mathscr{P}^B(A|X)$ ZPYXXe the set of all versions of the X-conditional probability of the event A with respect to the measure P^B, 390

$P^B(A|X{=}x)$ ZPYXXe an $(X{=}x)$-conditional probability of A with respect to the measure P^B, 396

$P^{Z=z}(A|X{=}x)$ ZPYXXe an $(X{=}x)$-conditional probability of A with respect to the measure $P^{Z=z}$, 397

$P^{Z=z}(A|\mathscr{C})$ ZPZCCc a version of the $\mathscr{C}$-conditional probability of the event A with respect to the measure P^B, 390

$\mathscr{P}^{Z=z}(A|\mathscr{C})$ ZPZCCf the family of all versions of the $\mathscr{C}$-conditional probability of the event A with respect to the measure $P^{Z=z}$, 390

$P^{Z=z}(A|X)$ ZPZXXc a version of the X-conditional probability of the event A with respect to the measure $P^{Z=z}$, 391

$\mathscr{P}^{Z=z}(A|X)$ ZPZXXe the family of all versions of the X-conditional probability of the event A with respect to the measure $P^{Z=z}$, 391

Q_X Zqfaaa quantile function of a real-valued random variable X, 180

$\mathbb{Q}$ Zqhaaa set of all rational numbers, 13

$Q_{lin}(Y|X)$ Zqlaaa the composition of X and the linear quasi-regression (or linear least-squares regression of Y on X), 214

$Q_{lin}(Y|X_1,\ldots,X_n)$ Zqmaaa linear quasi-regression of Y on $X_1,\ldots,X_n$, 227

$R_{Y|X}$ Zrgaaa multiple correlation of Y and X, 327

$R_{Y|X_1,\ldots,X_n}$ Zrgcaa multiple correlation of Y and $(X_1,\ldots,X_n)$, 327

$R^2_{Y|\mathscr{C}}$ Zrhjaa coefficient of determination of $E(Y|\mathscr{C})$, 325

$R^2_{Y|X}$ Zrhjca coefficient of determination of $E(Y|X)$, 325

$R^2_{Y|X_1,\ldots,X_n}$ Zrhjda coefficient of determination of $E(Y|X_1,\ldots,X_n)$, 326

$\mathbb{R}$ Zrjcbe set of all real numbers, 3

$\mathbb{R}^2$ Zrjfaa Cartesian product $\mathbb{R}\times\mathbb{R}$, 3

$\mathbb{R}^n$ Zrjfaaa n-fold Cartesian product $\mathbb{R}\times\ldots\times\mathbb{R}$, 3

$\overline{\mathbb{R}}$ Zrjjaa extended set of all real numbers, i. e., $\overline{\mathbb{R}}=\mathbb{R}\cup\{\infty,-\infty\}$, 13

$SD(Y)$ Zscaaa standard deviation of the random variable Y, 206

$SE(\overline{Y})$ Zscdaa standard error of the sample mean of the random variables $Y_1,\ldots,Y_n$, 208

$\operatorname{sgn}(f)$ Zscdha sign function of f, 64

$SD(Y|X{=}x)$ Zsckaa $(X{=}x)$-conditional standard deviation of Y, 331

$SD(Y|\mathscr{C})$ Zscmaa a version of the $\mathscr{C}$-conditional standard deviation of Y, 330

$SD(Y|X)$ Zscmaa a version of the X-conditional standard deviation of Y, 330

$\sigma(\mathscr{E})$ Zsdhaa σ-algebra generated by the set system $\mathscr{E}$, 9

$\sigma(f)$ Zsdhda σ-algebra generated by the mapping f, 52

$\sigma(f_1,\ldots,f_n)$ Zsdhfa σ-algebra generated by the mappings $f_1,\ldots,f_n$, 55

$\sigma(X)$	Zsdhja σ-algebra generated by the random variable X, 166
$\boldsymbol{\Sigma_{xy}}$	Zsiaaa covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$, 224
$\boldsymbol{\Sigma_{xx}}$	Zsmaaa variance-covariance matrix of $\boldsymbol{x}$, 226
$Var(Y)$	Zvaaaa variance of the random variable Y, 206
$Var(Y\|X{=}x)$	Zvcmaa ($X{=}x$)-conditional variance of Y, 331
$Var(Y\|X)$	Zvcmaa a version of the X-conditional variance of Y, 330
$Var(Y\|\mathscr{C})$	Zvgaaa a version of the $\mathscr{C}$-conditional variance of Y, 330
$\boldsymbol{x}$	Zxcaaa column vector of numerical random variables, 223
$\boldsymbol{X}$	Zxdaaa matrix of numerical random variables, 223
$X^{-1}(\mathscr{A}'_X)$	Zxdhjb σ-algebra generated by the random variable X, 166
$X \underset{P}{=} Y$	Zxhaaa X and Y are P-equivalent, 158
$X(\omega) \underset{P\text{-}a.a.}{=} Y(\omega)$	Zxhjaa X and Y are identical for P-almost all $\omega \in \Omega$, 158
$\overline{X}, \overline{Y}$	Zyjaaa arithmetic mean (sample mean) of the random variables $X_1, \ldots, X_n$ or $Y_1, \ldots, Y_n$, respectively, 204
Z_Y	Zzaaaa Z-transformation of the random variable Y, 208