Probability and Inferential Statistics

Lecture SS 16

Sample and Estimator

Prof. Dr. Rolf Steyer

Estimate 2
Sample 3
Multivariate Sample 4
Estimator #1 5
Estimator #2 6
Expectation of $\overline{X}$ 7
Expectation of S^2 8
Expectation of S^2 9
Variance of $\overline{X}$ 10
Estimator #3 11
Estimator of σ^2 12
Standard error 13
Estimator #4 14
Estimator #5 15
Estimtor #6 16
Estimator #7 17
Summary of Concepts 18

Estimate

In the last lecture we introduced parameters such as expectation, variances, covariance, correlation, the coefficients of a linear quasi-regression, and conditional expectation value. In empirical applications, hypotheses about such parameters are often of interest.

However, in empirical applications, these parameters are unknown and we are only able to estimate them. For example, the expectation can be estimated by the *mean*

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

in a concrete sample. Similarly, the variance $Var(X)$ can be estimated by the *variance*

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{2}$$

in a concrete sample. These terms are introduced in each course on descriptive statistics.

Sample

Which are the properties of such estimates? Are they unbiased? What can we say about their precision, that is, how strong do they vary between different samples? These and similar questions can be answered if we consider such an estimate to be a value of a random variable, which is also called an *estimator*. A fundamental concept for this kind of questions is the concept of a *simple random sample*.

Definition 1. A *simple random sample of size n* is a sequence $X_1, \ldots, X_n$ of independent and identically distributed random variables on a probability space $(\Omega, \mathscr{A}, P)$. *Sampling* or *drawing a sample* then means conducting the random experiment represented by $(\Omega, \mathscr{A}, P)$. If we draw a sample, then the sequence $x_1, \ldots, x_n$ of values of the random variables $X_1, \ldots, X_n$ are observed. Such a sequence $x_1, \ldots, x_n$ is called the *observed sample, realized sample, concrete sample*, or *data sample*.

Remark 1

According to this definition, the random variables X_i, $i = 1, \ldots, n$, of a sample have one and the same distributions ("identically distributed") $P_{X_i} = P_X$, and therefore the same expectations $E(X_i) := \mu$ (if they exist) and the same variances $Var(X_i) := \sigma^2$ (if they exist). Because the random variables X_i of a sample are assumed to be *independent*, two different such random random variables are uncorrelated, that is

$$Cov(X_i, X_j) \;=\; Corr(X_i, X_j) \;=\; 0, \quad \text{if } i \neq j, \tag{3}$$

provided that these correlations exist.

Sample of a multivariate random variable

Remark 2

Each of the random variables X_i, $i = 1,\ldots,n$, in the definition of a simple random sample can also be a multivariate random variable.

For example, we can assume that

$$(Y_1, Z_1),\ldots,(Y_n, Z_n)$$

is a simple random sample, where

$$(Y_i, Z_i) : (\Omega, \mathscr{A}, P) \to (\overline{\mathbb{R}}^2, \overline{\mathscr{B}}_2), \quad i = 1,\ldots,n\,.$$

Under this assumption, the distributions $P_{(Y_i,Z_i)}$ of the (Y_i, Z_i), $i = 1,\ldots,n$, are identical. If we additionally assume that Y_i and Z_i have finite second moments, then this implies

$$Corr\,(Y_i, Z_i) \;=\; Corr\,(Y_j, Z_j) =: Corr\,(Y, Z), \quad \forall\, i, j = 1,\ldots,n, \tag{4}$$

and

$$Corr\,(Y_i, Y_j) \;=\; Corr\,(Z_i, Z_j) \;=\; Corr\,(Y_i, Z_j) \;=\; 0, \quad \forall\, i \neq j, \quad i, j = 1,\ldots,n. \tag{5}$$

Estimator

Definition 2. If $X_1,\ldots,X_n$ is a simple random sample and $X_i : (\Omega, \mathscr{A}, P) \to (\mathbb{R}, \mathscr{B})$, then the *sample mean*

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6}$$

is a new random variable. A value of $\overline{X}$ is denoted by $\overline{x}$. It is the *concrete sample mean* [see Eq. (1)]. Similarly, the *sample variance*

$$S^2 := \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}\,)^2 \tag{7}$$

is a new random variable. A value of S^2 is denoted by s^2. It is the *concrete sample variance* [see Eq. (2)].

Expectation and variance of an estimator

The sample mean $\overline{X}$ is a numerical random variable. Therefore, we can also consider its expectation $E(\overline{X})$ and its variance $Var(\overline{X})$.

Hence, we can now raise the following questions:

- Is $\overline{X}$ an *unbiased estimator* of the parameter $E(X_i) := \mu$,
  that is, does $E(\overline{X}) = \mu$ hold?
- Is S^2 an *unbiased estimator* of the parameter $Var(X_i) := \sigma^2$, that is, does $E(S^2) = \sigma^2$ hold?
- What is the variance $Var(\overline{X})$ or the standard deviation $SD(\overline{X})$ of the sample mean?

Definition 3. The standard deviation $SD(\overline{X})$ of the sample mean is also called the *standard error of the sample mean*. It is also denoted by $\sigma_{\overline{X}}$.
Correspondingly, $SD(S^2)$ is called the *standard error of the sample variance*.

Expectation of the sample mean

Theorem 1. Let $X_1, \ldots, X_n$ be a simple random sample with $E(X_1) = \ldots = E(X_n) := \mu$.
The the sample mean $\overline{X}$ has the expectation $E(\overline{X}) = \mu$.
Hence, $\overline{X}$ is an unbiased estimator of μ.

Proof:

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n} \cdot E\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} \mu \\ &= \frac{1}{n} \cdot n \cdot \mu \\ &= \mu. \end{aligned}$$

Expectation of the sample variance

Theorem 2. Let $X_1, \ldots, X_n$ be a simple random sample with $E(X_1) = \ldots = E(X_n) := \mu$ and $Var(X_1) = \ldots = Var(X_n) := \sigma^2$. Then the expectation of the sample variance is $E(S^2) = \sigma^2 - Var(\overline{X})$. Hence, the sample variance S^2 is *not an unbiased estimator* of the variance σ^2.

Proof:

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2\right) \\
&= E\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)\right) \\
&= E\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 - 2\overline{X}\,\frac{1}{n} \cdot \sum_{i=1}^{n} X_i + \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{X}^2\right) \\
&= E\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) \\
&= \frac{1}{n} \cdot \sum_{i=1}^{n} E\left(X_i^2\right) - E\left(\overline{X}^2\right) \\
&= \frac{1}{n} \cdot n \cdot E\left(X_i^2\right) - E\left(\overline{X}^2\right) \\
&= E\left(X_i^2\right) - E\left(\overline{X}^2\right).
\end{aligned}$$

Expectation of the sample variance (continued)

Proof (continued): Because

$$\begin{aligned}
\sigma^2 := Var(X_i) &= E\left(X_i^2\right) - E(X_i)^2 \\
&= E\left(X_i^2\right) - \mu^2
\end{aligned}$$

$E\left(X_i^2\right) = \sigma^2 + \mu^2$ holds, and therefore

$$E(S^2) = E\left(X_i^2\right) - E\left(\overline{X}^2\right) = \sigma^2 + \mu^2 - E\left(\overline{X}^2\right).$$

Furthermore,

$$\begin{aligned}
Var(\overline{X}) &= E\left(\overline{X}^2\right) - E(\overline{X})^2 \\
&= E\left(\overline{X}^2\right) - \mu^2
\end{aligned}$$

and therefore $E\left(\overline{X}^2\right) = Var(\overline{X}) + \mu^2$. However, this implies

$$\begin{aligned}
E(S^2) &= \sigma^2 + \mu^2 - E\left(\overline{X}^2\right) \\
&= \sigma^2 + \mu^2 - \left(Var(\overline{X}) + \mu^2\right) \\
&= \sigma^2 - Var(\overline{X}).
\end{aligned}$$

Variance and standard error of the sample mean

Theorem 3. Let $X_1, \ldots, X_n$ be a simple random sample with $Var(X_1) = \ldots = Var(X_n) := \sigma^2$.
The sample mean has the variance $Var(\overline{X}) = \sigma^2/n$.

Proof:

$$\begin{aligned} Var(\overline{X}) &= Var\left(\frac{1}{n}\cdot\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\cdot\left(\sum_{i=1}^{n} Var(X_i) + 2\cdot\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j)\right) \\ &= \frac{1}{n^2}\cdot\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\cdot\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

The standard deviation of the sample means, that is, the standard error of the sample mean, is

$$\sigma_{\overline{X}} := SD(\overline{X}) = \sigma/\sqrt{n}. \tag{8}$$

Standard error of the sample mean

The estimate of the sample mean *varies* from sample to sample. The *standard error of the sample mean* is

$$SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}. \tag{9}$$

According to this equation, the standard deviation of a sample mean is smaller for larger sample size n and the precision of the estimate also depends on the standard deviation $SD(X_i) = \sigma$ of the random variables X_i of the sample $X_1, \ldots, X_n$.

Unbiased estimator of the variance σ^2

Theorem 4. Let $X_1, \ldots, X_n$ be a simple random sample with $Var(X_1) = \ldots = Var(X_n) := \sigma^2$. Then

$$\widehat{\sigma}^2 := S^2 \cdot \frac{n}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{10}$$

is an unbiased estimator of the variance σ^2, that is $E\left(\widehat{\sigma}^2\right) = \sigma^2$.

Proof: According to Theorem 2, $E(S^2) = \sigma^2 - Var(\overline{X})$ is the expectation of the sample variance and according to Theorem 3, $Var(\overline{X}) = \sigma^2/n$. Hence,

$$E(S^2) = \sigma^2 - \frac{\sigma^2}{n}.$$

Therefore

$$\begin{aligned} E\left(S^2 \cdot \frac{n}{n-1}\right) &= \frac{n}{n-1} \cdot E\left(S^2\right) \\ &= \frac{n}{n-1} \cdot \left(\sigma^2 - \frac{\sigma^2}{n}\right) \\ &= \frac{n\,\sigma^2}{n-1} - \frac{\sigma^2}{n-1} = \frac{\sigma^2 \cdot (n-1)}{n-1} = \sigma^2. \end{aligned}$$

Estimation of the standard error of the sample mean

In many empirical applications, the variance σ^2 is unknown. In this case, Equation (9) for the standard error of the sample mean cannot be applied and we have to estimate the variance σ^2. In this case, we cannot compute but only *estimate* the standard error of $\overline{X}$,

According to Theorem 4

$$\widehat{\sigma}^2 := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{11}$$

is an unbiased estimator of the variance σ^2, and this also applies to the estimator

$$\widehat{\sigma}_{\overline{X}} := \frac{\widehat{\sigma}}{\sqrt{n}} \tag{12}$$

of the standard error $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Standard error of the sample mean — Example

Example

Consider again our parapsychological experiment, guessing the outcome of a coin flip. If we conduct it with $n = 10$ trials, then the sample size is 10 and the relative frequency of hits is

$$\overline{1}_H := \frac{X}{n} = \frac{\sum_{i=1}^{n} 1_{H_i}}{n}, \tag{13}$$

where H_i denotes the event "hit at trial i" and $X = \sum_{i=1}^{n} 1_{H_i}$ the *number of hits*. The random variable $\overline{1}_H$ is an estimator of the probability $p := P(H_i)$ of a hit in the ith trial, $i = 1, \ldots, n$. The relative frequency $\overline{1}_H$ varies between samples. Obviously, $\overline{1}_H$ is also a sample mean.

If we increase the sample size n, that is, the number of trials, then the deviation between the relative frequencies in these samples is smaller. More precisely, the standard error of $\overline{1}_H$ decreases and the estimate of the probability of hits is more precise. Now we can even tell exactly, *how* the precision of this estimate of the probability $p := P(H_i)$ of a hit at trial i improves.

Standard error of the sample mean — Example continued

Which is the standard error of $\overline{1}_H$ if the sample sizes are $n = 10$ and $n = 100$, respectively?

In this example:

$$\begin{aligned}
\sigma^2 := Var(1_{H_i}) &= E(1_{H_i}^2) - E(1_{H_i})^2 \\
&= E(1_{H_i}) - E(1_{H_i})^2 = E(1_{H_i}) \cdot \big(1 - E(1_{H_i})\big) \\
&= P(H_i = 1) \cdot \big(1 - P(H_i = 1)\big) \\
&= p \cdot (1 - p),
\end{aligned}$$

and the standard error of $\overline{1}_H$ is

$$\sigma_{\overline{1}_H} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p \cdot (1-p)}}{\sqrt{n}}.$$

If we assume $p = .5$ and $n = 10$, this yields

$$\sigma_{\overline{1}_H} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{.5 \cdot .5}}{\sqrt{10}} = \frac{.5}{\sqrt{10}} = .1581139.$$

If $n = 100$, then

$$\sigma_{\overline{1}_H} = \frac{.5}{\sqrt{100}} = .05.$$

Standard error of the sample mean — Example continued

In the computation of the standard error $\sigma_{\overline{1}_H}$ we presume $P(1_{H_i} = 1) = .5$. If this assumption cannot be made, then we can only estimate the standard error using Equation (12). Suppose, $n = 10$ and we had $X(\omega) = \sum_{i=1}^{n} 1_{H_i}(\omega) = 6$ hits. Then

$$\begin{aligned}\widehat{\sigma}^2(\omega) &= \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(1_{H_i}(\omega) - \overline{1}_H(\omega)\right)^2 = \frac{1}{10-1} \cdot \sum_{i=1}^{10} \left(1_{H_i}(\omega) - .6\right)^2 \\ &= \frac{1}{10-1} \cdot \sum_{i=1}^{10} \left(1_{H_i}(\omega) - .6\right)^2 = \frac{1}{9} \cdot \left(4 \cdot (0-.6)^2 + 6 \cdot (1-.6)^2\right) \\ &= \frac{1}{9} \cdot (4 \cdot .36 + 6 \cdot .16) \approx 0.267.\end{aligned}$$

Hence, the estimate of the standard error of the relative frequency of hits is:

$$\widehat{\sigma}_{\overline{1}_H}(\omega) = \widehat{\sigma}(\omega)/\sqrt{10} \approx \sqrt{0.267}/\sqrt{10} \approx 0.163.$$

If we would have had $X(\omega) = 7$ hits, then $\widehat{\sigma}^2(\omega) = 1/9 \cdot (3 \cdot .49 + 7 \cdot .09) \approx 0.233$ and

$$\widehat{\sigma}_{\overline{1}_H}(\omega) = \widehat{\sigma}(\omega)/\sqrt{10} = \sqrt{0.233}/\sqrt{10} \approx 0.153.$$

Hence, the estimate of the standard error does not only depend on the sample size, but also on the number $X(\omega)$ of hits obtained. The estimator $\widehat{\sigma}_{\overline{1}_H}$ of the standard error is again a random variable. Its value $\widehat{\sigma}_{\overline{1}_H}(\omega)$ depends on the concrete outcome ω of the random experiment "guessing the outcome of a coin flip 10 times".

Standard error of the sample mean — Example continue

Table 1: Probability distribution of the number of hits, estimated variances and estimated standard errors of the relative frequency of hits for guessing 10 times with probability $P(1_{H_i} = 1) = p = 1/2$.

Number of hits $X := \sum_{i=1}^{n} 1_{H_i}$	$p_X = P(X=x)$	$\overline{X} := X/10$	$\widehat{\sigma}^2 := \widehat{Var}(1_{H_i})$	$\widehat{\sigma}_{\overline{1}_H} := \widehat{\sigma}/\sqrt{10}$
0	.00098	0	.000	.000
1	.00977	1/10	.100	.100
2	.04395	2/10	.178	.133
3	.11719	3/10	.233	.153
4	.20508	4/10	.267	.163
5	.24609	5/10	.278	.167
6	.20508	6/10	.267	.163
7	.11719	7/10	.233	.153
8	.04395	8/10	.178	.133
9	.00977	9/10	.100	.100
10	.00098	1	.000	.000
	Expectations	.500	.250	$\sqrt{.025} \approx .1581$

Note: $1_{H_i}(\omega) = 1$, if the ith guess is correct, $1_{H_i}(\omega) = 0$, otherwise.

Summary of new concepts

A *simple random sample of size n* is a sequence $X_1, \ldots, X_n$ of independent and identically distributed random variables on a probability space $(\Omega, \mathscr{A}, P)$.

Drawing a sample means conducting the random experiment that is represent by $(\Omega, \mathscr{A}, P)$.

A *concrete sample of size n* is a sequence $x_1, \ldots, x_n$ of values of $X_1, \ldots, X_n$.

Sample mean: $\overline{X} := (1/n) \cdot \sum_{i=1}^{n} X_i$.

Sample variance: $S^2 := (1/n) \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Realized (or concrete) sample mean: $\overline{x} := (1/n) \cdot \sum_{i=1}^{n} x_i$.

Realized (or concrete) sample variance: $s^2 := (1/n) \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$.

Standard error of the sample mean: $\sigma_{\overline{X}} := \sigma / \sqrt{n}$.

Estimator of the standard error of the sample mean: $\widehat{\sigma}_{\overline{X}} := \widehat{\sigma} / \sqrt{n}$.

Unbiased estimator of the variance: $\widehat{\sigma}^2 := \dfrac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$.