Probability and Inferential Statistics

Lecture SS 16

Confidence Interval and Central Limit Theorem

Prof. Dr. Rolf Steyer

Distribution of the Sample Mean Assuming a Normal Distribution

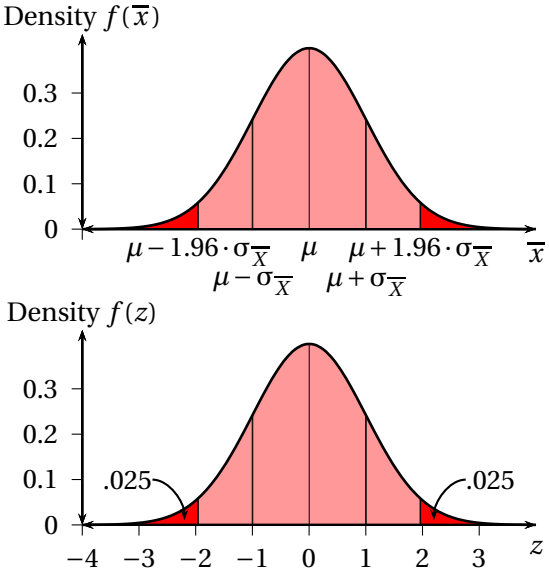
Theorem 1. Let $X_1, \ldots, X_n$ be a *normally distributed sample of size n* with expectation μ and variance σ^2, that is, a sample with normally distributed random variables X_i, $i = 1, \ldots, n$, with expectations $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$. Then the sample mean $\overline{X} := (X_1 + \cdots + X_n)/n$ is normally distributed as well with expectation μ and variance $\sigma^2_{\overline{X}} := \sigma^2/n$.

Under the assumptions of this theorem,

$$Z := \frac{(\overline{X} - \mu)}{\sigma_{\overline{X}}} \tag{1}$$

has a *standard normal distribution*. This implies that the 97.5%-quantile ≈ 1.96 of the standard normal distribution can be transformed to 97.5%-quantile $\approx \mu + 1.96 \cdot \sigma_{\overline{X}}$ of the distribution of $\overline{X}$. Therefore, the probability is $1 - \alpha = .95$ to observe a concrete mean $\overline{x}$ in the interval $\mu \pm 1.96 \cdot \sigma_{\overline{X}}$. Of course, the corresponding implications hold for other quantiles such as the 99.5%-quantile ≈ 2.576 of the standard normal distribution.

Density of a normally distributed sample mean



Confidence interval: Basic idea

Whenever the concrete sample mean $\overline{x}$ is in the interval $\mu \pm 1.96 \cdot \sigma_{\overline{X}}$ (see the light area in the last figure) — and this happens with probability .95 — then the interval

$$\overline{X} \pm 1.96 \cdot \sigma_{\overline{X}} \tag{2}$$

includes μ (see the last figure). This interval is called the 95%-*confidence interval* of the expectation μ.

The lower boundary of this confidence interval is $\overline{X} - 1.96 \cdot \sigma_{\overline{X}}$ and the upper boundary is $\overline{X} + 1.96 \cdot \sigma_{\overline{X}}$. Both boundaries are random variables. There values depend on the values $x_1, \ldots, x_n$ and, therefore, on the concrete mean $\overline{x}$ that realizes if we draw a sample.

Confidence interval if the standard deviation is unknown

If we want to use $\overline{X} \pm 1.96 \cdot \sigma_{\overline{X}}$, then we have to know the true variance $\sigma_{\overline{X}}$ of the sample mean. If this variance is unknown, then we can neither use the normal distribution with expectation μ and standard deviation $\sigma_{\overline{X}}$ nor the standard normal distribution. However, we can estimate the standard error of the sample mean via $\widehat{\sigma}_{\overline{X}}$ and consider the random variable

$$T := \frac{\overline{X} - \mu}{\widehat{\sigma}_{\overline{X}}}.$$

This random variable has a central t-distribution with $df = n - 1$ degrees of freedom. Therefore, we can use the 97.5%-quantile $t_{n-1}(.975)$ of the distribution of T, and specify the interval

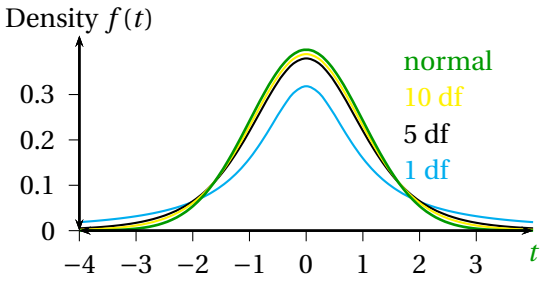
$$\overline{X} \pm t_{df}(.975) \cdot \widehat{\sigma}_{\overline{X}}\,, \tag{3}$$

which includes the expectation μ with probability .95. Note that the distribution of $\overline{X}$ is the normal distribution. Because we do not know $\sigma_{\overline{X}}$, we have to use the t-distribution, which yields a larger confidence interval.

Densities of some t-distribution with $df = 1, 5, 10$ and the standard normal distribution

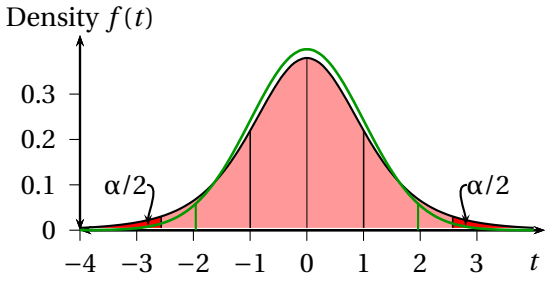
The t-distribution depends on the degrees of freedom $df = n-1$, which in turn depends on the sample size. Considering the random variable $T := (\overline{X} - \mu)/\hat{\sigma}_{\overline{X}}$ the degrees of freedom is $n-1$, because $n-1$ values of $X_i - \overline{X}$ can realize, if we consider the estimated standard error of the sample mean $\hat{\sigma}_{\overline{X}} = \hat{\sigma}/\sqrt{n}$, where

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}\,.$$



t-Distribution with five degrees of freedom

As mentioned before, the random variable $T := (\overline{X} - \mu)/\hat{\sigma}_{\overline{X}}$ has a t-distribution with $n-1$ degrees of freedom. Therefore, we observe with probability $1-\alpha = .95$ a concrete mean $\overline{x}$ in the interval $\mu \pm t_{df}(.975) \cdot \hat{\sigma}_{\overline{X}}$, where $t_{df}(.975) \approx 2.57$ is the 97.5%-quantile of the t-distribution.
If we have a normally distributed sampe of size $n = 6$, then we observe a concrete sample mean $\overline{x}$ in the interval $\mu \pm 2.57 \cdot \hat{\sigma}_{\overline{X}}$ (see the light area in the figure below) with probability .95. Therefore, the interval $\overline{X} \pm 2.57 \cdot \hat{\sigma}_{\overline{X}}$ includes the expectation μ with probability .95 as well. If we have a normally distributed sample with unknown variance of size $n = 6$, then this interval is called the 95%-*confidence interval of the expectation* μ. For sample size n, the confidence interval is: $\overline{X} \pm t_{df}(.975) \cdot \hat{\sigma}_{\overline{X}}$.



Confidence interval: Difference variable

Equation (3) for the confidence interval can also be applied to a difference variable $Y_1 - Y_2$. For example, $Y_1 - Y_2$ can represent the difference between a pretest and a post-test. We might then be interested in the true mean change, for example, after a psychological intervention. If we assume a normally distributed sample $D_1 := Y_{11} - Y_{12}, \ldots, D_n := Y_{n1} - Y_{n2}$, then the interval

$$\overline{D} \pm t_{df}(1-\alpha/2) \cdot \widehat{\sigma}_{\overline{D}} \tag{4}$$

includes the expectation $\mu_D := E(D_i) = E(Y_{i1} - Y_{i2})$ with probability $1-\alpha$. Under the assumptions (normally distributed sample), the interval (4) is the $100 \cdot (1-\alpha)\%$-*confidence interval of the expectation* μ_D.

Confidence interval: Mean difference of two independent samples

Theorem 2. Let $X_{11}, \ldots, X_{n_1 1}$ and $X_{12}, \ldots, X_{n_2 2}$ be two *independent* normally distributed samples of size n_1 and n_2 with expectations μ_1 and μ_2, respectively, and variance $\sigma_1^2 = \sigma_2^2 := \sigma^2$. According to this assumption, *all* $n_1 + n_2$ random variables $X_{11}, \ldots, X_{n_1 1}$, $X_{12}, \ldots, X_{n_2 2}$ are independent. Then

$$T := \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \tag{5}$$

has a t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom. In this equation

$$\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2} := \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}\,, \tag{6}$$

where

$$S_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_{i1} - \overline{X}_1)^2 \qquad \text{and} \qquad S_2^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (X_{j2} - \overline{X}_2)^2 \tag{7}$$

are the two sample variances. Therefore,

$$(\overline{X}_1 - \overline{X}_2) \pm t_{df}(1-\alpha/2) \cdot \widehat{\sigma}_{\overline{X}_1 - \overline{X}_2} \tag{8}$$

is the $100 \cdot (1-\alpha)\%$-confidence interval of the difference $\mu_1 - \mu_2$ of expectations, where $df = n_1 + n_2 - 2$.

General Definition of a Confidence Interval

Definition 1. Let $X_1, \ldots, X_n$ be sample and θ a parameter of the distributions of the random variables X_i. Furthermore, let $\boldsymbol{X} := (X_1, \ldots, X_n)$ be the vector of the random variables $X_1, \ldots, X_n$ and $f_1(\boldsymbol{X})$, $f_2(\boldsymbol{X})$ numerical functions of $\boldsymbol{X}$. If

$$P\big(f_1(\boldsymbol{X}) \leq \theta \leq f_2(\boldsymbol{X})\big) \;=\; 1-\alpha, \qquad 0<\alpha<1,$$

the interval $\big[f_1(\boldsymbol{X}), f_2(\boldsymbol{X})\big]$ is called *the* $100\cdot(1-\alpha)\%$*-confidence interval of* θ.

Under the standard normal distribution and the t-distribution the confidence intervals are always symmetric. Under other distributions, this is not necessarily the case. An example is the binomial distribution. There we may ask for the interval including (e.g., with probability .95) the unknown probability p. In this case, the confidence intervals are not symmetric any more around the relative frequency.

Central Limit Theorem for the Sample Mean

Theorem 3. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed (i. i. d.) random variables with finite expectation $E(X_i) := \mu$ and finite variance $Var(X_i) := \sigma^2$. Furthermore, let $\overline{X}_n := (X_1 + \ldots + X_n)/n$ denote the mean of the first n random variables X_i. Then the distribution of the standardized sample mean $\big(\overline{X}_n - \mu\big)/\sigma_{\overline{X}_n}$ converges to the standard normal distribution for $n \to \infty$. Hence, if $\Phi(z)$ denotes the distribution function of standard normal distribution, then

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \leq z\right) \;=\; \Phi(z).$$

In Theorem 3 we do not assume a distribution of a specific kind. Hence, the random variables X_i, $i = 1, \ldots, n$ can have a binomial distribution, a uniform distribution, a normal distribution, or any other kind of distribution. The theorem also applies even if the sample has a bimodal distribution. The only assumption is that $X_1, \ldots, X_n$ is a sample, that is, the random variables X_i are i. i. d. However, for a bimodal distribution, the convergence of the distribution of the standardized sample mean is slower than its convergence if the X_i have a normal distribution.