2.4 Estimation of Standard Errors

An important question when estimating parameters: How good are the estimators that we learned in previous sections? Standard errors can serve as measures of their accuracy. To estimate them, we derive an expression for the standard error and estimate all the unknown parameters in it.

Example 2.17. In Example 2.13 (Lecture 5), we estimated the parameter λ of a $Poisson$ distribution by

$$\overline{\lambda} = \overline{X},$$

using the method of moments. Let us estimate its standard error, based on the sample

$$\{7, 7, 11, 6, 5, 6, 7, 4\},$$

for which $\overline{X} = 6.625$ and $s = 2.0659$.

Solution. Recall that for a $Poisson(\lambda)$ distribution, the mean and the variance are

$$\mu = \sigma^2 = \lambda.$$

Also, we know that $V(\overline{X}) = \dfrac{V(X)}{n}$, hence,

$$\mathrm{Std}(\overline{X}) = \sqrt{V(\overline{X})} = \sqrt{\frac{V(X)}{n}} = \frac{\sigma}{\sqrt{n}}.$$

So, there are (at least) two ways to estimate the standard error of $\overline{\lambda}$.

On one hand, $\sigma = \sqrt{\lambda}$ for the $Poisson(\lambda)$ distribution, so we can estimate

$$\sigma_{\overline{\lambda},1} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\lambda}{n}} \approx \sqrt{\frac{\overline{\lambda}}{n}} = \sqrt{\frac{\overline{X}}{n}} = 0.91.$$

On the other hand, we can use the sample standard deviation as an estimate for the population one and get the estimate

$$\sigma_{\overline{\lambda},2} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = 0.7304.$$

Both estimates of the standard error $\sigma_{\overline{\lambda}}$ are rather small, so the estimator $\overline{\lambda}$ seems good.

■

Remark 2.18. Estimation of standard errors can become much harder for just slightly more complex estimators. In some cases, a nice analytic formula for $\sigma_{\overline{\theta}}$ may not exist. Then, other, more modern methods must be employed, such as *bootstrapping*, a method based on computer simulations.

3 The Normal and Student (T) Distributions

3.1 Normal Distribution $N(\mu, \sigma)$

The Normal distribution is, by far, the most important distribution, underlying many of the modern statistical methods used in data analysis. It was first described in the late 1700's by De Moivre, as a limiting case for the Binomial distribution (when n, the number of trials, becomes infinite), but did not get much attention. Half a century later, both Laplace and Gauss (independently of each other) rediscovered it in conjunction with the behavior of errors in astronomical measurements. It is also referred to as the "Gaussian" distribution.

A random variable X has a **Normal** distribution ($\boxed{\text{norm}}$) with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if its pdf is

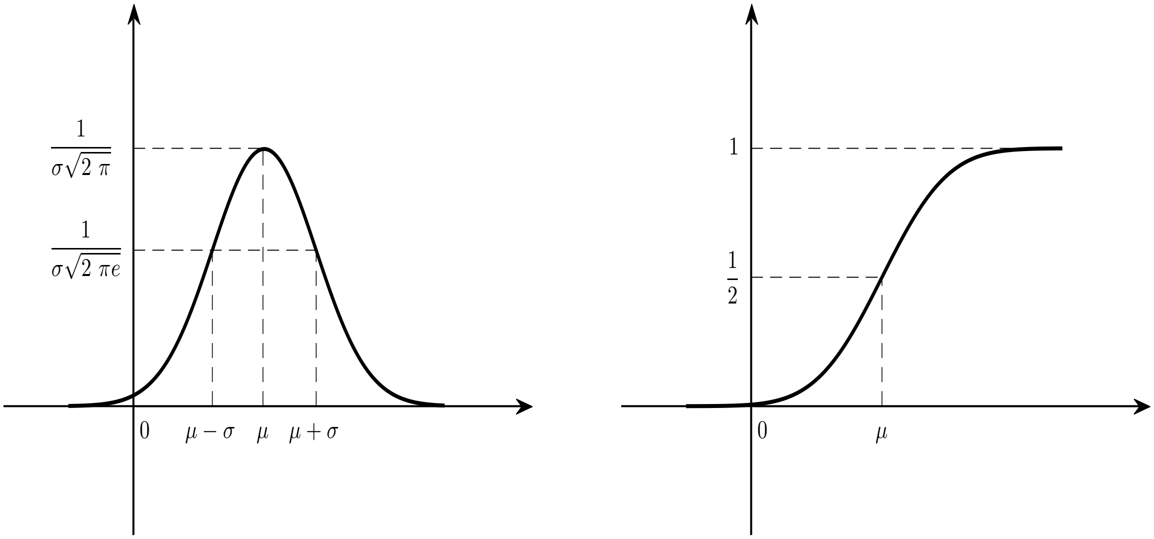
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{3.1}$$

The cdf of a Normal variable is then given by

$$F(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\dfrac{(t-\mu)^2}{2\sigma^2}}\, dt = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\dfrac{t^2}{2}}\, dt. \tag{3.2}$$

The graph of the Normal density is a symmetric, bell-shaped curve (known as "Gauss's bell" or "Gauss's bell curve") centered at the value of the first parameter μ, as can be seen in Figure 1(a). The graph of the cdf of a Normally distributed random variable is given in Figure 1(b) and this is approximately what the graph of the cdf of *any* continuous random variable looks like.

Remark 3.1.
1. There is an important particular case of a Normal distribution, namely $N(0, 1)$, called the **Standard (or Reduced) Normal Distribution**. A variable having a Standard Normal distribution is



(a) Density Function (pdf)

(b) Cumulative Distribution Function (cdf)

Fig. 1: Normal Distribution

usually denoted by Z. The density and cdf of Z are given by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \;\; x \in \mathbb{R} \quad \text{and} \quad F_Z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \tag{3.3}$$

The function F_Z given in (3.3) is known as *Laplace's function* (or *the error function*) and its values can be found in tables or can be computed by most mathematical software.

3. As noticed from (3.2) and (3.3), there is a relationship between the cdf of any Normal $N(\mu, \sigma)$ variable X and that of a Standard Normal variable Z, namely

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right) \; .$$

Asymptotic Normality

By the Central Limit Theorem, the sum of observations, and therefore, the sample mean have approximately Normal distribution if they are computed from a large sample. That is, the distribution of $\overline{X}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and that of

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(which is the *reduced* variable of $\overline{X}$) is approximately Standard Normal ($N(0,1)$) as $n \to \infty$. This property is called *asymptotic normality*. The same is true for other statistics, e.g. the difference of means:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \longrightarrow N(0,1), \text{ as } n_1, n_2 \to \infty.$$

3.2 Student Distribution

The **Student (T)** distribution appeared as a necessity, when the sample size was small and asymptotic normality could not be used. It was developed in the early 1900's by W. S. Gosset under the pseudonym "Student". It has one parameter, denoted by n or ν or simply, df and it stands for "number of degrees of freedom". The T-distribution is symmetric and bell-shaped, like the Normal one, only it is narrower (see Figure 2).

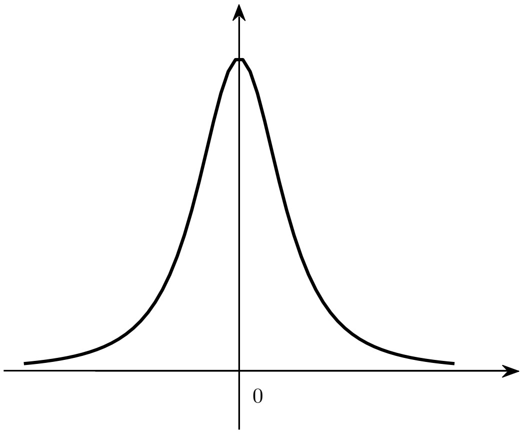


Fig. 2: Student (T) Distribution pdf

Quantiles

Recall *quantiles*. A quantile of a given order $\alpha \in (0,1)$ for a random variable X with cdf F, is a value q_α with the property that

$$F(q_\alpha) \;=\; P(X \le q_\alpha) \;=\; \alpha,\; q_\alpha \;=\; F^{-1}(\alpha),$$

i.e., that the area under the graph of the pdf, to the *left* of q_α is α.

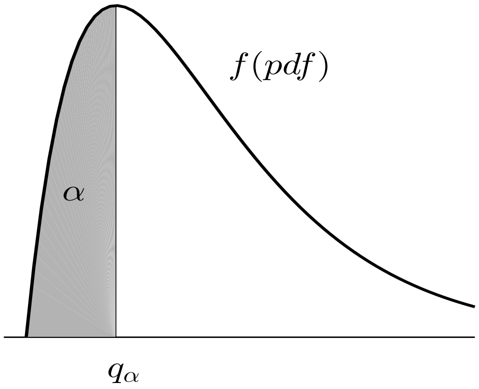


Fig. 3: Quantile of order $\alpha \in (0,1)$

4 Estimation by Confidence Intervals

4.1 Basic Concepts; General Framework

So far, point estimators provided one single value, $\overline{\theta}$, to estimate the value of an unknown parameter θ, but little measure of the accuracy of the estimate. In contrast, an **interval estimator** specifies a *range* of values, within which the parameter is estimated to lie. More specifically, the sample will be used to produce *two* sample functions, $\overline{\theta}_L(X_1,\ldots,X_n) < \overline{\theta}_U(X_1,\ldots,X_n)$, with values $\overline{\theta}_L = \overline{\theta}_L(x_1,\ldots,x_n), \overline{\theta}_U = \overline{\theta}_U(x_1,\ldots,x_n)$, respectively, such that for a given $\alpha \in (0,1)$,

$$P(\overline{\theta}_L \le \theta \le \overline{\theta}_U) \;=\; 1-\alpha. \tag{4.1}$$

Then
$-$ the range $(\overline{\theta}_L, \overline{\theta}_U)$ is called a **confidence interval (CI)**, more specifically, a $100(1-\alpha)\%$ confidence interval,

$-$ the values $\overline{\theta}_L, \overline{\theta}_U$ are called (lower and upper) **confidence limits**,
$-$ the quantity $1-\alpha$ is called **confidence level** or **confidence coefficient** and
$-$ the value α is called **significance level**.

Remark 4.1.
1. It may seem a little peculiar that we use $1-\alpha$ instead of simply α in (4.1), since both values are in $(0,1)$, but the reasons are in close connection with *hypothesis testing* and will be revealed in the next sections.
2. The condition (4.1) *does not* uniquely determine a $100(1-\alpha)\%$ CI.
3. Evidently, the smaller α and the length of the interval $\overline{\theta}_U - \overline{\theta}_L$ are, the better the estimate for θ. Unfortunately, as the confidence level increases, so does the length of the CI, thus, reducing accuracy.

To produce a CI estimate for θ, we need a *pivotal quantity*, i.e. a statistic S that satisfies two conditions:
$-$ $S = S(X_1, \dots, X_n; \theta)$ is a function of the sample measurements and the unknown parameter θ, this being the *only* unknown,
$-$ the distribution of S is known and does not depend on θ.

We will use the pivotal method to find $100(1-\alpha)\%$ CI's. Depending on which population parameter we wish to estimate, the expression and the pdf of the pivot will change, but the principles will stay the same. So, we start with the case where the pivot has a (possibly asymptotically) $N(0,1)$ distribution, so we can better understand the ideas.

Let θ be a target parameter and let $\overline{\theta}$ be an unbiased estimator for θ ($E(\overline{\theta}) = \theta$), with standard error $\sigma_{\overline{\theta}}$, such that, under certain conditions, it is know that

$$Z = \frac{\overline{\theta} - \theta}{\sigma_{\overline{\theta}}} \left(= \frac{\overline{\theta} - E(\overline{\theta})}{\sigma(\overline{\theta})} \right) \tag{4.2}$$

has an approximately Standard Normal $N(0,1)$ distribution. We can use Z as a pivotal quantity to construct a $100(1-\alpha)\%$ CI for estimating θ. Since the pdf of Z is known, we can choose two values, Z_L, Z_U such that for a given $\alpha \in (0,1)$,

$$P(Z_L \le Z \le Z_U) = 1 - \alpha. \tag{4.3}$$

How to choose them? Of course, there are infinitely many possibilities. Recall that for continuous random variables, the probability in (4.3) represents an *area*, namely the area under the graph of the pdf and above the x-axis, between the values Z_L and Z_U. Basically, the values Z_L and Z_U should be chosen so that that area is $1 - \alpha$. We will take advantage of the symmetry of the Standard Normal pdf and choose the two values so that the area $1 - \alpha$ is in "the middle". That means (since the total area under the graph is 1) the two portions left on the two sides, both should have an area of $\dfrac{\alpha}{2}$, as seen in Figure 4.

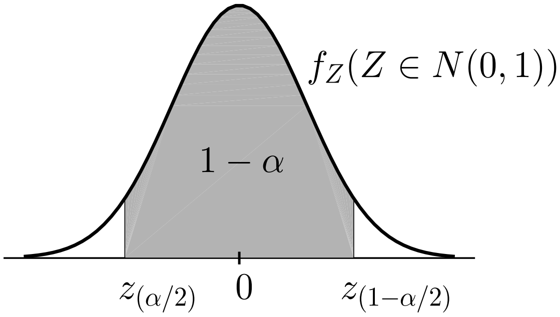


Fig. 4: Confidence interval for the $N(0,1)$ distribution

Since for Z_L we want the area to its left to be $\alpha/2$, we choose it to be the quantile of order $\alpha/2$ for Z,

$$Z_L = z_{\alpha/2}.$$

For the value Z_U, the area to its *right* should be $\alpha/2$, which means the area to the left is $1 - \alpha/2$. Thus, we choose

$$Z_U = z_{1-\alpha/2}.$$

Indeed, now we have

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha,$$

as in (4.3).

From here, we proceed to rewrite the inequality inside, until we get the limits of the CI for θ. We have

$$\begin{aligned} 1 - \alpha &= P\left(z_{\frac{\alpha}{2}} \leq \frac{\overline{\theta} - \theta}{\sigma_{\overline{\theta}}} \leq z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}} \leq \overline{\theta} - \theta \leq \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(-\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} \leq \theta - \overline{\theta} \leq -\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right) \\ &= P\left(\overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} \leq \theta \leq \overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right), \end{aligned}$$

so the $100(1-\alpha)\%$ CI for θ is given by

$$\left[\overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}},\, \overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right]. \tag{4.4}$$

Remark 4.2.
1. Since the Standard Normal distribution is symmetric about the origin, we have $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ and the CI can be written as

$$\left[\overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}},\, \overline{\theta} + \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}\right] \quad \text{or} \quad \left[\overline{\theta} + \sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}},\, \overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right].$$

2. As mentioned earlier, for estimating various population parameters, the pivot will be different, but the procedure of finding the CI will be the same, even when the distribution of the pivot *is not* symmetric.

One-sided confidence intervals

The CI we determined is a **two-sided CI**, because it gives bounds on both sides. A two-sided CI is not always the most appropriate for the estimation of a parameter θ. It may be more relevant to make a statement simply about how *large* or how *small* the parameter might be, i.e. to find confidence intervals of the form $(-\infty, \overline{\theta}_U]$ and $[\overline{\theta}_L, \infty)$, respectively, such that the probability that θ is in the CI is $1-\alpha$. These are called **one-sided confidence intervals** and they can be found the same way, using quantiles of an appropriate order.

• *Lower confidence interval* for θ
We want to find θ_U such that $P(\theta \leq \theta_U) = 1-\alpha$. We have, successively.

$$\begin{aligned} 1-\alpha &= P(\theta \leq \theta_U) \;=\; P(-\theta \geq -\theta_U) \\ &= P\Big(\frac{\overline{\theta}-\theta}{\sigma_{\overline{\theta}}} \geq \frac{\overline{\theta}-\theta_U}{\sigma_{\overline{\theta}}}\Big) \\ &= P\Big(Z \geq \frac{\overline{\theta}-\theta_U}{\sigma_{\overline{\theta}}}\Big). \end{aligned}$$

But we know that $P(Z \geq z_\alpha) = 1-\alpha$, so, by equating $\dfrac{\overline{\theta}-\theta_U}{\sigma_{\overline{\theta}}} = z_\alpha$, we get $\theta_U = \overline{\theta} - \sigma_{\overline{\theta}} \cdot z_\alpha$ and the lower CI

$$\left(-\infty, \overline{\theta} - \sigma_{\overline{\theta}} \cdot z_\alpha\right] \;=\; \left(-\infty, \overline{\theta} + \sigma_{\overline{\theta}} \cdot z_{1-\alpha}\right],$$

the last equality coming from the symmetry of the quantiles $z_{1-\alpha} = -z_\alpha$.

• *Upper confidence interval* for θ

Similarly, to find θ_L such that $P(\theta \geq \theta_L) = 1 - \alpha$, we use

$$\begin{aligned} 1-\alpha &= P(\theta \geq \theta_L) \;=\; P(-\theta \leq -\theta_L) \\ &= P\Big(\frac{\overline{\theta}-\theta}{\sigma_{\overline{\theta}}} \leq \frac{\overline{\theta}-\theta_L}{\sigma_{\overline{\theta}}}\Big) \\ &= P\Big(Z \leq \frac{\overline{\theta}-\theta_L}{\sigma_{\overline{\theta}}}\Big) \;=\; P(Z \leq z_{1-\alpha}), \end{aligned}$$

so the upper CI is

$$\big[\overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{1-\alpha}, \infty\big) \;=\; \big[\overline{\theta} + \sigma_{\overline{\theta}} \cdot z_{\alpha}, \infty\big)\,.$$

4.2 Confidence Intervals for One Population Mean

Let X be a population characteristic, with mean $\mu = E(X)$ and variance $V(X) = \sigma^2$, whose pdf depends on a parameter θ, $f(x;\theta)$. Let $X_1, X_2, \ldots, X_n$ be a sample drawn from the pdf of X.
The formulas for finding confidence intervals for the mean μ are based on the following results.

Proposition 4.3. *Assume that $X \in N(\mu, \sigma)$ or that the sample size is large enough ($n > 30$). Then*

$$\text{a) } Z = \frac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} \in N(0,1) \quad \textit{and} \quad \text{b) } T = \frac{\overline{X}-\mu}{\dfrac{s}{\sqrt{n}}} \in T(n-1).$$

<u>CI for the mean, known variance</u>

If either $X \in N(\mu, \sigma)$ or the sample is large enough ($n > 30$) and σ is known, then by Proposition 4.3, we can use the pivot

$$Z = \frac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} \in N(0,1).$$

The procedure will go *exactly* as described in the previous section, with $\theta = \mu, \overline{\theta} = \overline{X}, \sigma_{\overline{\theta}} = \dfrac{\sigma}{\sqrt{n}}$.

The $100(1-\alpha)\%$ CI for the mean is given by

$$\mu \;\in\; \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\; \overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]. \tag{4.5}$$

Since $N(0,1)$ is symmetric (and one quantile is the negative of the other), we can write it in short as

$$\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \overline{X} \mp z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}. \tag{4.6}$$

CI for the mean, unknown variance

In practice, it is somewhat unreasonable to expect to know the value of σ, if the value of μ is unknown. We can find CI's for the mean, without knowing the variance. If either $X \in N(\mu, \sigma)$ or the sample is large enough ($n > 30$), then by Proposition 4.3, we can use the pivot

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1).$$

The same computations as before will lead to the $100(1-\alpha)\%$ CI for the mean:

$$\mu \ \in \ \left[\overline{X} - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}},\ \overline{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]. \tag{4.7}$$

Notice that we change the notations for the quantiles, according to the pdf of the pivot (z for $N(0,1)$, t for $T(n-1)$, etc.). The Student $T(n-1)$ is also symmetric (see Figure 5), so again, we can write the CI in short as

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad \text{or} \quad \overline{X} \mp t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}. \tag{4.8}$$

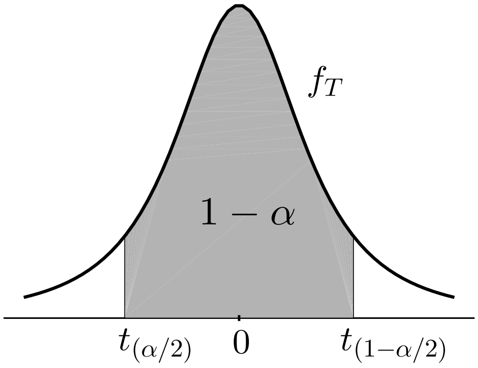


Fig. 5: Confidence interval for the T distribution

Remark 4.4. The parameter of a Student T distribution, ν, is generally called *number of degrees of freedom*. One might wonder why in estimating the mean, this parameter is $\nu = n-1$ and not $\nu = n$, the sample size. The sample variables $X_1, \dots, X_n$ are independent, so it would seem that there are $\nu = n$ degrees of freedom. But its meaning is the dimension of the vector used to estimate the sample variance

$$s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2,$$

where we use the vector $X_1 - \overline{X}, \dots, X_n - \overline{X}$. Notice that by subtracting the sample mean $\overline{X}$ from each observation, there exists a linear relation among the elements, namely

$$\sum_{k=1}^{n} (X_k - \overline{X}) = 0,$$

so we lose 1 degree of freedom due to this constraint.

In general, the number of degrees of freedom can be computed as

$$\nu \quad = \quad \text{sample size} \quad - \quad \text{number of estimated parameters.}$$

However, it should be noted that this issue is important only when the sample size is *small* ($n < 30$), when there is significant difference in the values of the quantiles. When n is large, we may use the quantiles for $T(\nu)$ with $\nu = n$ or $\nu = n-1$, since for both distributions, we have

$$T(n),\ T(n-1) \overset{n\to\infty}{\longrightarrow} N(0,1),$$

so both quantiles are approximately equal to the z quantiles.

Selecting the sample size

Notice that in the case of a Normal distribution of the pivot, the CI we find is symmetric and its length is

$$2\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}.$$

We can revert the problem and ask a very practical question: How large a sample should be collected to provide a certain desired precision of our estimator? In other words, what sample size n guarantees that the margin of a $(1-\alpha)100\%$ CI does not exceed a specified limit Δ? To answer this question, we only need to solve the inequality

$$2\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} \;\;\leq\;\; \Delta \tag{4.9}$$

in terms of n. Typically, parameters are estimated more accurately based on larger samples, so that the standard error $\sigma_{\overline{\theta}}$ and the margin are decreasing functions of the sample size n. Then, (4.9) will be satisfied for sufficiently large n.

For example, when estimating the mean in the case of known variance, inequality (4.9) comes down to

$$2\frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \;\;\leq\;\; \Delta,$$

so we require

$$n \;\;\geq\;\; \left(\frac{2\sigma}{\Delta} z_{1-\frac{\alpha}{2}}\right)^2 \tag{4.10}$$

Example 4.5. Consider a sample of measurements

$$2.5,\ 7.4,\ 8.0,\ 4.5,\ 7.4,\ 9.2,$$

drawn from an approximately Normal distribution.
a) Find a 95% confidence interval for the population mean, if the measurement device guarantees a standard deviation of $\sigma = 2.2$.
b) How many measurements should be taken in order for the length of the 95% confidence interval for the mean to not exceed 1?
c) Without any information on the population variance, construct 95% two- and one-sided CI's for the mean of the population.

Solution. This sample has size $n = 6$ and sample mean $\overline{X} = 6.5$. To attain a confidence level of $1 - \alpha = 0.95$, we need $\alpha = 0.05$ and $\alpha/2 = 0.025$.
a) Since $\sigma = 2.2$ is known, we use formula (4.5). Hence, we need quantiles

$$z_{0.025} \;=\; -1.96,\;\; z_{0.975} \;=\; 1.96.$$

We find the 95% CI for the mean

$$\left[\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] \;=\; [4.74,\ 8.26].$$

That means that the mean μ of the population from which the sample was drawn is between 4.74 and 8.26 with probability 0.95.

b) Notice that the length of the CI found in part a) is $\approx\ 3.52$, quite large (not much precision). If we want to improve the accuracy of our estimate (shorten the length of the interval), we need to enlarge the sample, take more measurements.
With $\sigma = 2.2,\ z_{0.975} = 1.96$ and $\Delta = 1$, we find from (4.10),

$$n \;\geq\; \left(\frac{2\sigma}{\Delta} z_{1-\frac{\alpha}{2}}\right)^2 \;=\; 74.37,$$

so, a sample of size at least 75 will ensure the fact that the length of the 95% CI for the mean does not exceed 1.

c) If σ is not known, we use s instead. We have $s\ =\ 2.497$ and the quantiles for the $T(5)$ distribution

$$\begin{aligned} t_{1-\alpha/2} &\;=\; t_{0.975} \;=\; 2.57 \\ t_{1-\alpha} &\;=\; t_{0.95} \;=\; 2.02. \end{aligned}$$

We find the two-sided CI to be

$$[\overline{X} \mp \frac{s}{\sqrt{n}} t_{1-\alpha/2}] \;=\; [3.88, 9.12],$$

the lower CI

$$(-\infty, \overline{X} + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}] \;=\; (-\infty, 8.55]$$

and the upper CI

$$[\overline{X} - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}, \infty) \;=\; [4.45, \infty).$$

■

4.3 Confidence Intervals for One Population Proportion

Recall (from Lecture 5) that a *population proportion* is

$$p = P(i \in A),$$

where A is a subpopulation.

Based on a random sample $X_1, \dots, X_n$, we define the *sample proportion* as

$$\overline{p} = \frac{\text{number of sampled items from } A}{n}.$$

Then

$$\begin{aligned} E\left(\overline{p}\right) &= p, \\ V\left(\overline{p}\right) &= \frac{p(1-p)}{n} = \frac{pq}{n}. \end{aligned} \qquad (4.11)$$

So $\overline{p}$ is an absolutely correct estimator for p and by a CLT,

$$Z = \frac{\overline{p} - p}{\sqrt{\dfrac{pq}{n}}} \qquad (4.12)$$

converges in distribution to a Standard Normal $N(0,1)$ variable, as $n \to \infty$.

Now, as p is unknown, we estimate the standard error $\sigma_{\overline{p}} = \sqrt{V\left(\overline{p}\right)} = \sqrt{\dfrac{p(1-p)}{n}}$ by

$$s_{\overline{p}} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}.$$

So, again, for large samples ($n > 30$), we can use

$$Z = \frac{\overline{p} - p}{\sqrt{\dfrac{\overline{p}(1-\overline{p})}{n}}} \in N(0,1)$$

as a pivot to construct a confidence interval for p.

For a given confidence level $1-\alpha$, with the same computations as before, we obtain a $100(1-\alpha)\%$

CI for the population proportion p as

$$\left[\overline{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right]. \tag{4.13}$$

Remark 4.6. With the same procedure as before, one-sided CI's for a population proportion can be also derived:

$$\Big(-\infty, \overline{p} - z_{\alpha}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\Big] \text{ and } \Big[\overline{p} - z_{1-\alpha}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \infty\Big).$$

Selecting the sample size

Just as we did for the population mean (in the case of known variance), we can derive a formula for the sample size that will provide a certain precision of our interval estimator. The length of the CI in (4.13) is

$$2\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} z_{1-\frac{\alpha}{2}}.$$

Notice that for any $\overline{p} \in (0, 1)$, we have

$$\overline{p}(1-\overline{p}) \;\leq\; \frac{1}{4}.$$

Then to get a desired precision

$$2\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} z_{1-\frac{\alpha}{2}} \;\leq\; \Delta,$$

we solve

$$2\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} z_{1-\frac{\alpha}{2}} \;\leq\; 2 \cdot \frac{1}{2}\frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \;\leq\; \Delta,$$

for n. We get

$$n \;\geq\; \left(\frac{z_{1-\frac{\alpha}{2}}}{\Delta}\right)^2. \tag{4.14}$$

Example 4.7. A company has to accept or reject a large shipment of items. For quality control purposes, they collect a sample of 200 items and find 12 defective items in it.
a) Find a 99% confidence interval for the proportion of defective items in the whole shipment.

b) How many items should be tested to ensure a 99% confidence interval of length at most 0.05?
c) Find a 99% *upper* confidence interval for the proportion of defective items in the whole shipment.

Solution. The sample is large enough and we have

$$\overline{p} \;=\; \frac{12}{200} \;=\; 0.06.$$

For $1-\alpha = 0.99$, $\alpha = 0.01$, $\alpha/2 = 0.005$, the quantile is

$$z_{0.005} \;=\; -2.576.$$

Then the 99% confidence interval for the proportion of defective items is

$$\left[\overline{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] \;=\; \left[0.06 \pm 2.576\sqrt{\frac{0.06\cdot 0.94}{200}}\right] \;=\; [0.017,\; 0.103]\,.$$

So, with 99% confidence, the percentage of defective items is between 1.7% and 10.3%.

b) The length of the 99% CI we found is 0.086. For a margin of $\Delta \le 0.05$ of the 99% CI, we need a sample size of

$$n \;\ge\; \left(\frac{z_{0.995}}{\Delta}\right)^2 \;=\; \left(\frac{2.576}{0.05}\right)^2 \;=\; 2653.898 \;\approx\; 2654.$$

c) We now use the quantile $z_{0.99} = 2.326$. The upper CI for p is

$$\left[\overline{p} - z_{0.99}\sqrt{\frac{\overline{p}(1-\overline{p})}{200}}, \infty\right) \;= [0.021, \infty).$$

That means that with 99% confidence, the percentage of defective items is at least 2.1%.

■