2.2 Absolutely Correct and Consistent Estimators

Recall that we seek unbiased (or nearly unbiased) estimators that have *low* standard error or at least, "decreasing" standard error. There are several ways to interpret that.

One way is for the variance to decrease as the sample size increases.

Definition 2.1. *An estimator $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$ is called an **absolutely correct** estimator for θ, if it satisfies the conditions*

(i) $E(\overline{\theta}) = \theta$,

(ii) $\lim\limits_{n\to\infty} V(\overline{\theta}) = 0$.

Remark 2.2. The sample mean $\overline{X}$ is an absolutely correct estimator for the theoretical mean $\mu = E(X)$. More generally, the sample moment of order k, $\overline{\nu}_k$, is an absolutely correct estimator for the population moment of order k, $\nu_k = E\left(X^k\right)$. In fact, *all* the unbiased estimators that we discussed (in the Table from last time) are absolutely correct.

Also, we would expect that as the sample size n increases, $\overline{\theta}$ gets "closer" to θ, at least in a probabilistic sense. That is the idea behind consistent estimators.

Definition 2.3. *An estimator $\overline{\theta} = \overline{\theta}_n$, found from a sample of size n, is said to be a **consistent** estimator for θ, if $\overline{\theta}_n \xrightarrow{p} \theta$ ($\overline{\theta}_n$ converges in probability to θ), i.e. if for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\Big(|\overline{\theta}_n - \theta| < \varepsilon\Big) = 1.$$

The property of consistency of a point estimator ensures the fact that the larger the sample size, the better the estimate. The estimate "improves consistently" with increasing the sample size. The following assertion is a direct consequence of Chebyshev's inequality.

Proposition 2.4. *An absolutely correct estimator is consistent.*

Remark 2.5. The sample moment of order k, $\overline{\nu}_k$, is a consistent estimator for the population moment of order k, $\nu_k = E\left(X^k\right)$, since it is absolutely correct. In particular, the sample mean $\overline{X}$ is a consistent estimator for the theoretical mean $\mu = E(X)$.

The notions of *unbiasedness* and *consistency* seem to be very close, however they are not equivalent: Unbiasedness is a statement about the expected value of the sampling distribution of the estimator. Consistency is a statement about "where the sampling distribution of the estimator is going" as the sample size increases. Let us consider a few examples.

Example 2.6. Let $X_1, \ldots, X_n$ be a random sample drawn from a $N(\mu, \sigma)$ population, with both parameters $\mu \in \mathbb{R}, \sigma > 0$ unknown.

For estimating the mean μ, consider the estimator $\overline{\mu} = X_1$. Obviously it is an unbiased estimator for μ, since

$$E(X_1) \;=\; E(X) \;= \mu.$$

But, $\overline{\mu}$ is *not* consistent, since its distribution does *not* become more concentrated around μ as the sample size increases, it stays $N(\mu, \sigma)$, no matter how large the sample size gets.

Example 2.7. Let $X_1, \ldots, X_n$ be a random sample drawn from a population with pdf

$$X \begin{pmatrix} -a & a \\ 0.5 & 0.5 \end{pmatrix},$$

with $a > 0$ unknown.
Consider the estimator $\hat{\theta} = \max\{X_1, \ldots, X_n\}$ for the estimation of a.

First, we compute the population mean and variance

$$\begin{aligned} E(X) &= -a \cdot 0.5 + a \cdot 0.5 = 0, \\ V(X) &= E\left(X^2\right) - (E(X))^2 = a^2, \end{aligned}$$

the last assertion following from the fact that $X^2 \equiv a^2$ (X^2 takes a single value, namely a^2, with probability 1).

Let us find the pdf of $\hat{\theta}$. Obviously, $\hat{\theta}$ can only take the values a or $-a$. The only way that the maximum of the X_i's is $-a$ is if *all* variables X_i take the value $-a$. That means that

$$\begin{aligned} P(\hat{\theta} = -a) &= P(X_1 = -a) \ldots P(X_n = -a) = \frac{1}{2^n} \text{ and, consequently,} \\ P(\hat{\theta} = a) &= 1 - \frac{1}{2^n}. \end{aligned}$$

Thus, the pdf of $\hat{\theta}$ is

$$\hat{\theta} \begin{pmatrix} -a & a \\ \dfrac{1}{2^n} & 1 - \dfrac{1}{2^n} \end{pmatrix},$$

and its mean is

$$E(\hat{\theta}) \;= -\frac{a}{2^n} + a\left(1 - \frac{1}{2^n}\right) \;=\; a\left(1 - \frac{1}{2^{n-1}}\right) \;<\; a.$$

So $\hat{\theta}$ is *biased*. However, it is a consistent estimator of a because the error probability $\frac{1}{2^n}$converges to 0 as the sample size increases, so the limit of the pdf of $\hat{\theta}$ as $n \to \infty$ is the constant random variable $\begin{pmatrix} a \\ 1 \end{pmatrix}$.

2.3 Method of Moments

So far, we have discussed desirable properties of point estimators, how to distinguish "good" from "bad" or "better" estimators, based on how reliable they are in approximating the value of a population parameter. In all the procedures we analyzed and all the examples we discussed, the value of a point estimator $\overline{\theta}$ was given for a target parameter θ, based on sample variables $X_1, X_2, \ldots, X_n$, i.e. $\overline{\theta} = \overline{\theta}(X_1, X_2, \ldots, X_n)$. But *how* to *actually* find an estimator, an approximating value? Sometimes, such a value may be "guessed" from past experience or from observing many samples over time. But, most of the time, we need mathematical ways of producing a point estimator, which can then be analyzed from the various points of view discussed in the previous section.

There are several popular methods for pointwise estimation. In what follows, we present one of the oldest and easiest methods for obtaining point estimators, first formalized by K. Pearson in the late 1800's, the *method of moments*.

Let us recall, for a population characteristic X, we define the *moments of order* k as

$$\nu_k \;=\; E\left(X^k\right) \;=\; \begin{cases} \displaystyle\sum_{i\in I} x_i^k p_i, & \text{if } X \text{ is discrete with pdf } X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I} \\ \displaystyle\int_{\mathbb{R}} x^k f(x)\, dx, & \text{if } X \text{ is continuous with pdf } f : \mathbb{R} \to \mathbb{R}. \end{cases} \tag{2.1}$$

For a sample drawn from the distribution of X, i.e. sample variables $X_1, \ldots, X_n$ (iid), the *sample moments of order* k are defined by

$$\overline{\nu}_k \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^k. \tag{2.2}$$

Also, let us recall that

$$\begin{aligned} E(\overline{\nu}_k) &= \nu_k, \\ V(\overline{\nu}_k) &= \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right) \;\to\; 0, \text{ as } n \to \infty, \end{aligned} \tag{2.3}$$

so the sample moment of order k is an *absolutely correct* (and, hence, a *consistent*) estimator for the population moment of the same order.

That is precisely the idea of this method. Since our sample comes from a family of distributions $\{f(\theta)\}$, we choose such a member of this family whose properties are close to properties of our data. Namely, we shall match the moments. As the theoretical (population) moments in (2.1) contain the target parameters that are to be estimated, while the sample moments in (2.2) are all known, computable from the sample data, simply set the two to be equal and solve the resulting system. To estimate k parameters, equate the first k population and sample moments:

$$\left\{ \begin{array}{rcl} \nu_1 & = & \overline{\nu}_1 \\ \ldots & \ldots & \ldots \\ \nu_k & = & \overline{\nu}_k \end{array} \right. \tag{2.4}$$

The left-hand sides of these equations depend on the distribution parameters. The right-hand sides can be computed from data. The **method of moments estimator** is the solution of this $k \times k$ system of equations.

Remark 2.8. We state, without proof, the fact that an estimator $\overline{\theta}_n$ obtained by the method of moments is a *consistent* estimator.

Example 2.9. Consider a $Poisson$ distribution of parameter $\lambda > 0$, unknown. Its pdf is

$$X \left(\begin{array}{c} k \\ \frac{\lambda^k}{k!} e^{-\lambda} \end{array} \right)_{k=0,1,\ldots}.$$

Let us estimate the parameter λ by the method of moments, based on a sample $\{X_1, \ldots, X_n\}$.
Solution. There is only one unknown parameter, hence we write one equation:

$$\nu_1 \ = \ \overline{\nu}_1,$$

where $\nu_1 = \mu = \lambda$ is the mean of the Poisson distribution and $\overline{\nu}_1 = \overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$ is the sample mean. So, we are solving the simple equation

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

"Solving" it for λ, we obtain

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

the method of moments estimator of λ. So, for instance, if we have the sample

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

based on that, we find the estimator

$$\overline{\lambda} \;=\; \frac{53}{8} \;=\; 6.625.$$

■

Example 2.10. The following sample

$$\{-1, 1, 1, 2, -1, 2, 1, 1, 1, 2\}$$

was drawn from a distribution with pdf

$$X \begin{pmatrix} -1 & 1 & 2 \\ \frac{1}{4}\theta & 1 - \frac{1}{2}\theta & \frac{1}{4}\theta \end{pmatrix},$$

with $0 < \theta < 2$, unknown. What is the method of moments estimator of θ?

Solution. Again, we have one unknown, so one equation

$$\nu_1 \;=\; \overline{\nu}_1,$$

where

$$\nu_1 \;=\; \mu \;=\; -1 \cdot \frac{1}{4}\theta + 1 \cdot \left(1 - \frac{1}{2}\theta\right) + 2 \cdot \frac{1}{4}\theta \;=\; 1 - \frac{\theta}{4}$$

is the population mean and

$$\overline{\nu}_1 \;=\; \overline{X} \;=\; \frac{X_1 + \ldots + X_{10}}{10} \;=\; \frac{9}{10}$$

is the sample mean. So, we have

$$1 - \frac{\theta}{4} \;=\; \overline{X},$$

which yields the estimator

$$\hat{\theta} \;=\; 4(1 - \overline{X}) \;=\; \frac{2}{5} \;=\; 0.4.$$

■

Example 2.11. Let us recall the example we used before (Example 4.4, Lecture 3), where to evaluate the effectiveness of a processor, a sample of CPU times for $n = 30$ randomly chosen jobs (in seconds) was considered:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

The histogram we did suggested that the CPU times have a *Gamma* distribution with some unknown parameters $\alpha > 0$ and $\lambda > 0$. Let us use this sample to estimate them by the method of moments.

Solution. For the Gamma distribution with parameters $\alpha, \lambda > 0$, it is known that the population mean and variance are given by

$$\begin{aligned} \mu &= \alpha\lambda, \\ \sigma^2 &= \alpha\lambda^2. \end{aligned}$$

There are two unknown parameters, so we will need two equations to estimate them. We have $\nu_1 = \mu$. Since we also have the variance (the population *central* moment of order 2), let us use that for our second equation. In other words, we consider the (population and sample) moments of order 1 (for the first equation) and the (population and sample) *central* moments of order 2 (for the second equation). Recall that for this sample, we found (in Lectures 3 and 4) that the sample mean and variance are

$$\begin{aligned} \overline{\nu}_1 &= \overline{X} = 48.2333 \text{ and} \\ s^2 &= 703.1506 \end{aligned}$$

We cam also compute the sample central moment of order 2, as

$$\overline{\mu}_2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) = \frac{70^2 + \ldots + 19^2 - 30 \cdot 48.2333^2}{30} = \frac{90185 - 69794}{30} = 679.7.$$

So, we solve the system

$$\begin{cases} \alpha\lambda &= \overline{X} = 48.2333 \\ \alpha\lambda^2 &= \overline{\mu}_2 = 679.7, \end{cases}$$

which has the solution

$$\begin{cases} \overline{\alpha} &= \dfrac{\overline{X}^2}{\overline{\mu}_2} = 3.4228, \\ \overline{\lambda} &= \dfrac{\overline{\mu}_2}{\overline{X}} = 14.0919. \end{cases}$$

■

Remark 2.12. Method of moments estimates are typically easy to compute. However, on rare occasions, when k equations are not enough to estimate k parameters, higher moments (i.e. more equations) can be considered. Also, as we have seen in Example 2.11, *central* (population and sample) moments can be used, to make computations easier.

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 2.9 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,

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

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 $N(\mu, \sigma)$

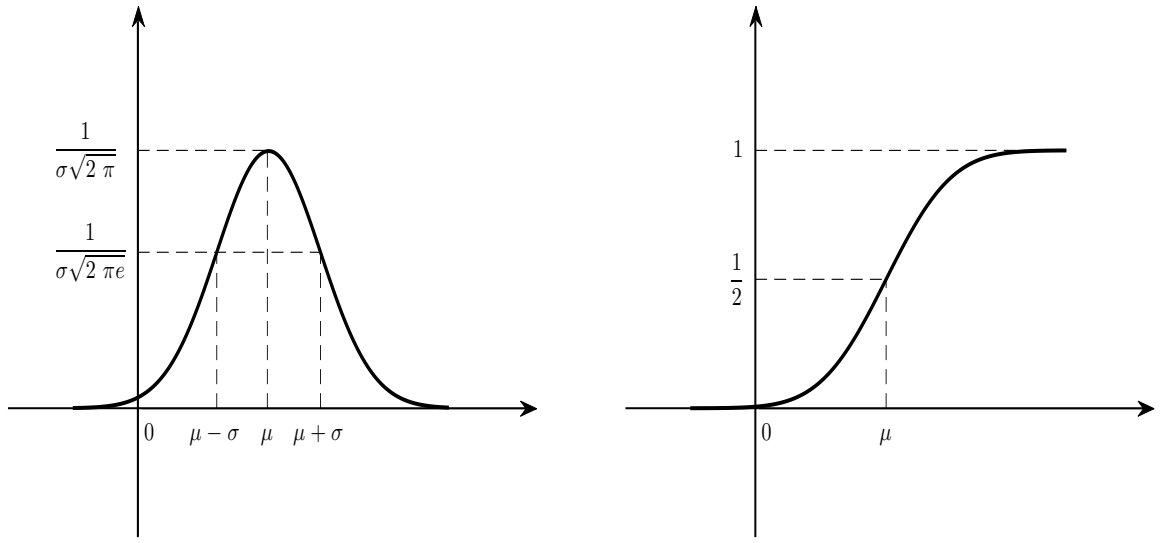
The Normal distribution is, by far, the most important distribution, underlying many of the modern statistical methods used in data analysis. It was first described in the late 1700's by De Moivre, as a limiting case for the Binomial distribution (when n, the number of trials, becomes infinite), but 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 (norm) with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \;\; x \in \mathbb{R}. \tag{3.5}$$

The cdf of a Normal variable is then given by

$$F(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\dfrac{(t-\mu)^2}{2\sigma^2}}\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\dfrac{t^2}{2}}\, dt. \tag{3.6}$$



(a) Density Function (pdf) (b) Cumulative Distribution Function (cdf)

Fig. 1: Normal Distribution

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

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R} \quad \text{and} \quad F_Z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \tag{3.7}$$

The function F_Z given in (3.7) 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 (3.6) and (3.7), 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 $N\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big)$ 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:

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

Quantiles

In many inferential statistical procedures, we will need to use *quantiles*. Recall that 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 (see Figure 2).

For *symmetric* distributions, the symmetry is reflected in the computation of quantiles. By symmetry, we have

$$\begin{aligned} P(X \le -q_\alpha) \ &\overset{\text{sym}}{=} \ P(X \ge q_\alpha) \ = \ 1 - P(X \le q_\alpha) \\ &= \ 1-\alpha \ = \ P(X \le q_{1-\alpha}), \end{aligned}$$

therefore,

$$q_{1-\alpha} \ = \ -q_\alpha, \ \ \forall \alpha \in (0,1). \tag{3.8}$$

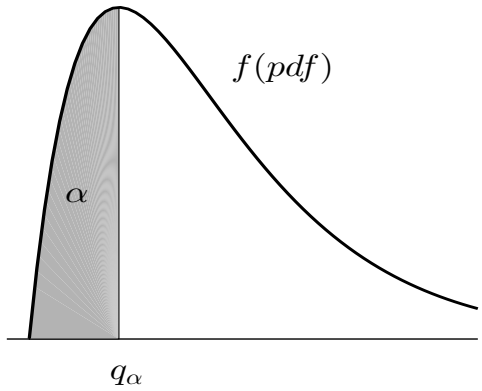


Fig. 2: Quantile of order $\alpha \in (0, 1)$

This is certainly the case for the Standard Normal distribution (see Figure 3):

$$z_{1-\alpha} \;=\; -z_\alpha, \;\; \forall \alpha \in (0, 1).$$

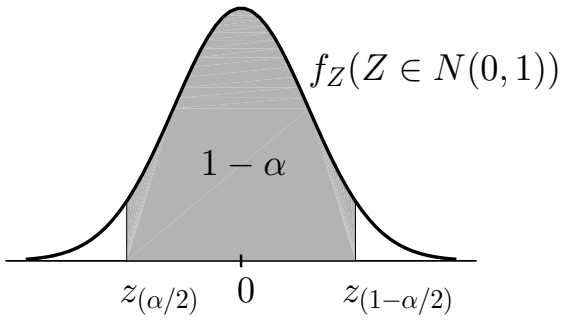


Fig. 3: Quantiles for the $N(0, 1)$ distribution

3.2 Student (T) 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. Since it is symmetric, its quantiles also satisfy relation (3.8) (see Figure 4):

$$t_{1-\alpha} \;=\; -t_\alpha, \;\; \forall \alpha \in (0,1).$$

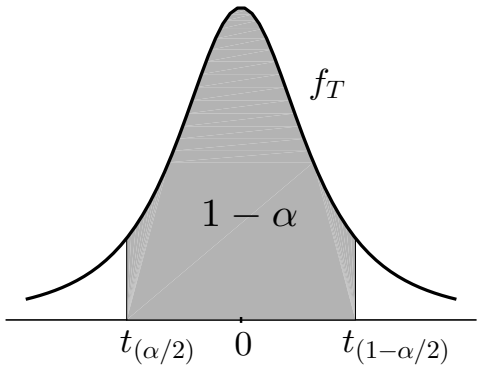


Fig. 4: Student T Distribution pdf and quantiles

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:
$-$ $S = S(X_1, \ldots, X_n; \theta)$ 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 3.

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

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

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

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

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

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