2. Parameter Estimation

In statistical inference we are interested in making conclusions about the population of interest. Often this means making a statement about an unknown parameter describing the population. In this chapter we discuss methods using data that can infer the value of the unknown parameter.

2.1 Point estimation

Suppose we are given a random sample from a population $f(x|\theta)$ depending on an unknown parameter θ with values in the parameter space Θ, i.e., $\theta \in \Theta$. We wish to use the sample to infer the value of θ within Θ. Any function of the sample which can be used for this purpose is an estimator for θ.

Definition 2.1 (Estimator)**.**
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ be a random sample from a population which depends on a parameter $\theta \in \Theta$. Any statistic $T = T(X_1, \ldots, X_n)$ taking values in a subset of Θ, i.e., $T \in \Theta$, is called an **estimator** for the parameter θ. Suppose we observe $X_1 = x_1, \ldots, X_n = x_n$ and evaluate $t = T(x_1, \ldots, x_n)$. The value t corresponding to the observed values $x_1, \ldots, x_n$ is called an **estimate** of θ.

Note that an estimator, being a function of the random sample, is itself a random variable. We can therefore talk about the distribution of this estimator. We can potentially come up with several estimators so using their distribution we can evaluate their performance. Two of the most commonly used criteria for evaluating estimators are the **bias** and the **mean squared error** which we define below.

Definition 2.2 (Bias)**.**
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and let $T = T(X_1, \ldots, X_n)$ be an estimator for θ. The difference

$$\text{Bias}_\theta(T) = \mathsf{E}\, T - \theta,$$

is called the **bias** of the estimator T of the parameter θ. If $\text{Bias}_\theta(T) = 0$, then the estimator T is called **unbiased** for θ, otherwise it is called **biased** for θ.

A desirable property for an estimator is to be unbiased.

Definition 2.3 (Mean squared error)**.**
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and let $T = T(X_1, \ldots, X_n)$ be an estimator for θ. The **mean squared error (MSE)** of the estimator T for the parameter θ is defined by

$$\text{MSE}_\theta(T) = \mathsf{E}\left\{(T - \theta)^2\right\}.$$

The MSE is always non-negative. It is desirable that the MSE be small.

Lemma 2.1.
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and let $T = T(X_1, \ldots, X_n)$ be an estimator for θ. Then

$$\text{MSE}_\theta(T) = \mathsf{Var}\, T + (\text{Bias}_\theta(T))^2$$

Proof. By the definition of MSE, add and subtract $\mathsf{E}\, T$ in the brackets, and note that $\mathsf{E}\, T$ and θ are not random,

$$\begin{aligned}
\text{MSE}_\theta(T) &= \mathsf{E}\left\{(T-\theta)^2\right\} \\
&= \mathsf{E}\left\{(T - \mathsf{E}\, T + \mathsf{E}\, T - \theta)^2\right\} \\
&= \mathsf{E}\left\{(T - \mathsf{E}\, T)^2 + 2(T - \mathsf{E}\, T)(\mathsf{E}\, T - \theta) + (\mathsf{E}\, T - \theta)^2\right\} \\
&= \mathsf{E}\left\{(T - \mathsf{E}\, T)^2\right\} + 2\,\mathsf{E}\left\{(T - \mathsf{E}\, T)(\mathsf{E}\, T - \theta)\right\} + \mathsf{E}\left\{(\mathsf{E}\, T - \theta)^2\right\} \\
&= \mathsf{E}\left\{(T - \mathsf{E}\, T)^2\right\} + 2(\mathsf{E}\, T - \theta)\,\mathsf{E}\left\{(T - \mathsf{E}\, T)\right\} + (\mathsf{E}\, T - \theta)^2 \\
&= \mathsf{Var}\, T + 0 + (\text{Bias}_\theta(T))^2.
\end{aligned}$$

□

According to Lemma 2.1, the MSE incorporates two components, one measuring the variability of the estimator and the other measuring its bias (accuracy). An estimator with low MSE has low combined variance and bias.

Example 2.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta) \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$. The parameter space for μ is $\mathbb{R}$ and for σ^2 is $[0, \infty)$. Then $\bar{X}$ is an estimator for μ because $\bar{X} \in \mathbb{R}$. Its bias is $\text{Bias}_\mu(\bar{X}) = \mathsf{E}\,\bar{X} - \mu = \mu - \mu = 0$ and its variance is $\mathsf{Var}\,\bar{X} = \sigma^2/n$. Therefore, its MSE is $\text{MSE}_\mu(\bar{X}) = \mathsf{Var}\,\bar{X} + \text{Bias}_\mu(\bar{X})^2 = \sigma^2/n$. ▶

Example 2.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. The parameter space for p is $[0, 1]$. Then $\bar{X}$ is an estimator for p because $\bar{X} \in \{0, 1/n, 2/n, \ldots, 1\} \subset [0, 1]$. Its bias is $\text{Bias}_p(\bar{X}) = \mathsf{E}\,\bar{X} - p = p - p = 0$ and its variance is $\mathsf{Var}\,\bar{X} = p(1-p)/n$. Therefore, its MSE is $\text{MSE}_p(\bar{X}) = p(1-p)/n$. Because $p(1-p) \in [0, \frac{1}{4}]$, with the lower bound attained when $p = 0$ or 1 and the upper bound attained when $p = \frac{1}{2}$, $\text{MSE}_p(\bar{X}) \in [0, \frac{1}{4n}]$.

Consider a different estimator given by $T = \dfrac{2\sum X_i + \sqrt{n}}{2n + 2\sqrt{n}}$. Because $\sum X_i \in \{0, 1, \ldots, n\}$, $T \in [0, 1]$ so it is an estimator for p. Its bias is $\text{Bias}_p(T) = \dfrac{2np + \sqrt{n}}{2n + 2\sqrt{n}} - p = (1 - 2p)\dfrac{\sqrt{n}}{2n + 2\sqrt{n}} = \dfrac{\frac{1}{2} - p}{\sqrt{n} + 1}$. So this estimator has no bias if $p = \frac{1}{2}$ but has positive bias (overestimates p) if $p < \frac{1}{2}$ and negative bias (underestimates p) if $p > \frac{1}{2}$. The variance of this estimator is $\mathsf{Var}\, T = \dfrac{np(1-p)}{(n + \sqrt{n})^2} = \dfrac{p(1-p)}{(\sqrt{n}+1)^2}$ so $\text{MSE}_p(T) = \dfrac{p(1-p) + (\frac{1}{2} - p)^2}{(\sqrt{n}+1)^2} = \dfrac{\frac{1}{4}}{(\sqrt{n}+1)^2}$.

If we wish to choose between $\bar{X}$ and T in terms of their MSE, we see that for all $n \geq 1$, $0 < \text{MSE}_p(T) < \frac{1}{4n}$ so it falls between the values of $\text{MSE}_p(\bar{X})$. In particular, if in reality $p = \frac{1}{2}$, then T will always have lower MSE than $\bar{X}$. For some other value of p, say $p = \frac{1}{4}$ then T has lower MSE if $n \leq 41$ but otherwise $\bar{X}$ has lower MSE. ▶

We discuss next a few classical estimation methods.

2.1.1 Method of moments estimator

The method of moments estimation is the simplest method for finding estimators. Consider a population $f(x|\theta)$, $\theta \in \Theta$ and define the rth moment by

$$\mu_r = \mathsf{E}(X^r), \text{ for } r = 1, 2, \ldots,$$

i.e., the expectation of the rth power of X. In the case $r = 1$, $\mu_1 = \mathsf{E}\, X$ corresponds to the mean of the population, while for $r = 2$, $\mu_2 = \mathsf{E}\, X^2$, so $\mathsf{Var}\, X = \mathsf{E}\, X^2 - (\mathsf{E}\, X)^2 = \mu_2 - \mu_1^2$.

A convenient method for computing the moments is through the moment generating function (mgf). Recall $M_X(t) = \mathsf{E} \exp(tX)$ and

$$\mu_r = \frac{d^r}{dt^r} M_X(t)|_{t=0}.$$

Example 2.3. Let $X \sim \text{Exponential}(\mu)$, i.e., $f(x|\mu) = (1/\mu)\exp(-x/\mu)$. Then

$$\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx} \frac{1}{\mu} e^{-\frac{x}{\mu}}\, \mathrm{d}x \\
&= \frac{1}{\mu} \int_0^\infty e^{-x(\frac{1}{\mu} - t)}\, \mathrm{d}x \\
&= \frac{1}{\mu} (\frac{1}{\mu} - t)^{-1} \left[-e^{-x(\frac{1}{\mu} - t)} \right]_0^\infty \\
&= (1 - t\mu)^{-1}, \text{ assuming } t < 1/\mu. \\
\Rightarrow M_X^{(1)}(t) &= \frac{d}{dt} M_X(t) = \mu(1 - t\mu)^{-2} \\
\Rightarrow \mu_1 &= M_X^{(1)}(0) = \mu \\
\Rightarrow M_X^{(2)}(t) &= \frac{d}{dt} M_X(t) = 2\mu^2(1 - t\mu)^{-3} \\
\Rightarrow \mu_2 &= M_X^{(2)}(0) = 2\mu^2
\end{aligned}$$

▶

It is apparent that the rth moment is a function of the parameter θ which we write as $\mu_r(\theta)$ to make this dependence explicit. Note that θ may be a scalar or a κ-dimensional vector $\theta = (\theta_1, \ldots, \theta_\kappa)$.

Now suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and consider the rth *sample* moment

$$m_r = \frac{1}{n} \sum_{i=1}^n X_i^r, \text{ for } r = 1, 2, \ldots.$$

In particular $m_1 = \bar{X}$ and $m_2 = \frac{1}{n}\sum X_i^2$. The sample moments are functions of the sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ and we write $m_r(\mathbf{X})$ to make this dependence explicit.

The method of moments estimates the rth moment by the corresponding sample moment, i.e., the method of moments estimator (MoM) for θ, which we denote by $\hat{\theta}$, is given by the solution of the following system of equations,

$$\mu_r(\hat{\theta}) = m_r(\mathbf{X}), \text{ for } r = 1, 2, \ldots.$$

Because there are κ unknown parameters, we need κ equations to be able to identify $\hat{\theta}$ uniquely. These are selected among those equations corresponding to the lowest moments up to as many as needed to be able to solve for $\hat{\theta}$.

Example 2.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\mu)$, $\mu > 0$. Then $\mu_1 = \mu$ so $\hat{\mu} = \bar{X}$ is the method of moments estimator. ▶

Example 2.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, i.e., normal with known mean 0 and unknown variance $\sigma^2 > 0$. Then $\mu_1 = 0$ and $\mu_2 = \sigma^2$. Note that the first moment does not depend on the parameter so the first equation, $\mu_1 = \bar{X}$, is not helpful for estimating σ^2. Using the second equation we have $\hat{\sigma}^2 = m_2$. ▶

Example 2.6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{U}(0, \theta)$, i.e., uniform with known lower bound 0 and unknown upper bound $\theta > 0$. The pdf is $f(x|\theta) = \theta^{-1}$, $x \in (0, \theta)$, so $\mu_1 = \int_0^\theta x\theta^{-1}\,\mathrm{d}x = \theta^{-1}\left[\frac{x^2}{2}\right]_0^\theta = \theta/2$. Using the first moment equation we have $\hat{\theta}/2 = \bar{X} \Rightarrow \hat{\theta} = 2\bar{X}$. ▶

Example 2.7. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, i.e., gamma with shape $\alpha > 0$ and rate $\beta > 0$. In this case there are two parameters to estimate, i.e., $\kappa = 2$. The mgf of this distribution is given by $M_X(t) = (1 - \frac{t}{\beta})^{-\alpha}$. Then $\mu_1 = \alpha/\beta$ and $\mu_2 = \alpha/\beta^2 + \alpha^2/\beta^2$. This leads to the following system of equations

$$\hat{\alpha}/\hat{\beta} = m_1, \qquad \hat{\alpha}/\hat{\beta}^2 + \hat{\alpha}^2/\hat{\beta}^2 = m_2.$$

By substituting $\hat{\alpha} = \hat{\beta} m_1$ from the first equation into the second, we have $m_1/\hat{\beta} + m_1^2 = m_2 \Rightarrow \hat{\beta} = m_1/(m_2 - m_1^2)$ and $\hat{\alpha} = m_1^2/(m_2 - m_1^2)$.

An obvious question to ask is are these actual estimators? In other words, are $\hat{\alpha}$ and $\hat{\beta} > 0$? To check this we need to check whether $m_2 > m_1^2$ for all possible samples. But $0 < \sum(X_i - \bar{X})^2 = \sum(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2$. So $\sum X_i^2 > n\bar{X}^2 \Rightarrow \frac{1}{n}\sum X_i^2 > n\bar{X}^2 \Rightarrow m_2 > m_1^2$ as required. ▶

2.1.2 Maximum likelihood estimator

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\theta \in \Theta$. Then, the joint density/mass function of $\mathbf{X} = (X_1, \ldots, X_n)$ is given by

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{2.1}$$

In (2.1), we see the parameter θ as fixed and evaluate the function at a given $\mathbf{x}$. If instead (2.1) is viewed as a function of θ for a given sample $\mathbf{x}$, then it is called a **likelihood function** and is denoted by $L(\theta|\mathbf{x})$. We have the following definition.

Definition 2.4 (Likelihood function)**.**
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\theta \in \Theta$. Then,

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta),$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is the observed value for $(X_1, \ldots, X_n)$.

Intuitively, the likelihood function tells us how likely the observed data are for that value of θ. Therefore it makes sense to estimate θ by that value which makes the observed data appear

more likely. Therefore we define the **maximum likelihood estimator** for the parameter θ, the value $\hat{\theta}$ for which $L(\theta|\mathbf{x})$ is maximised, i.e.,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta|\mathbf{x}).$$

In practice, it is usually easier to maximise the logarithm of the likelihood function instead of the likelihood function itself. We define the **log-likelihood function**, $\ell(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x})$. In this case the **maximum likelihood estimator** (MLE) becomes

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta|\mathbf{x}).$$

Note that θ could be a vector, i.e., $\theta = (\theta_1, \ldots, \theta_\kappa)$. In some special cases the MLE can be obtained by solving a system of equations

$$\frac{\partial}{\partial \theta_r} \ell(\theta|\mathbf{x}) = 0, \; r = 1, \ldots, \kappa,$$

but note that the MLE is not always obtained in this way.

Remark. It is custom when writing the log-likelihood function to omit additive constants which do not depend on the parameters. This makes the expression for the log-likelihood brief and does not affect the MLE. It is important however to remain consistent throughout our calculations to avoid errors.

Example 2.8. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\mu)$, $\mu > 0$. Then $f(x|\mu) = (1/\mu)\exp(-x/\mu)$ so $L(\mu|\mathbf{x}) = \prod\{(1/\mu)\exp(-x_i/\mu)\} = (1/\mu^n)\exp(-\sum x_i/\mu)$ and $\ell(\mu|\mathbf{x}) = -n\log\mu - \sum x_i/\mu$.

In this case we can find the MLE by solving $\frac{d\ell}{d\mu} = 0$:
$\frac{d\ell}{d\mu} = -n/\hat{\mu} + \sum x_i/\hat{\mu}^2 = 0 \Rightarrow -n + \sum x_i/\hat{\mu} = 0 \Rightarrow \hat{\mu} = \sum x_i/n = \bar{x}$. Note that this is identical to the MoM estimator in Example 2.4. ▶

Example 2.9. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, i.e., normal with known mean 0 and unknown variance $\sigma^2 > 0$. Then
$L(\sigma^2|\mathbf{x}) = \prod(2\pi\sigma^2)^{-\frac{1}{2}}\exp(-\frac{x_i^2}{2\sigma^2}) = (2\pi\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2\sigma^2}\sum x_i^2)$, so
$\ell(\sigma^2|\mathbf{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum x_i^2$.

Again we solve for $\frac{d\ell}{d\sigma^2} = 0$:
$\frac{d\ell}{d\sigma^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum x_i^2 = 0 \Rightarrow -n + \frac{1}{\hat{\sigma}^2}\sum x_i^2 = 0 \Rightarrow \hat{\sigma}^2 = \sum x_i^2/n$. Note that this is identical to the MoM estimator in Example 2.5. ▶

Example 2.10. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{U}(0, \theta)$, i.e., uniform with known lower bound 0 and unknown upper bound $\theta > 0$. The pdf is $f(x|\theta) = \theta^{-1}$ for $x \in (0, \theta)$. The pdf can be written using the indicator function as $f(x|\theta) = \theta^{-1}\mathbf{1}_{(0,\theta)}(x)$, where $\mathbf{1}_{(0,\theta)}(x) = 1$ if $x \in (0, \theta)$ and 0 otherwise. The reason for using this notation is because the value of x is apparent in the function. Then, the likelihood is
$L(\theta|\mathbf{x}) = \prod \theta^{-1}\mathbf{1}_{(0,\theta)}(x_i) = \theta^{-n}\prod \mathbf{1}_{(0,\theta)}(x_i)$. The term $\prod \mathbf{1}_{(0,\theta)}(x_i)$ will equal 1 if $x_i \in (0, \theta)$ for all $i = 1, \ldots, n$ and 0 otherwise. If all $x_i > 0$, then $\prod \mathbf{1}_{(0,\theta)}(x_i) = 1$ if and only if $x_i < \theta$ for all i, or equivalently, $\max\{x_i : i = 1, \ldots, n\} < \theta$.

Let $x_{(n)} = \max\{x_i : i = 1, \ldots, n\}$. Then

$$L(\theta|\mathbf{x}) = \begin{cases} 0 & \text{if } \theta < x_{(n)}, \\ \theta^{-n} & \text{if } \theta \geq x_{(n)}. \end{cases}$$

Because θ^{-n} is a decreasing function in θ, the likelihood is then maximised when $\hat{\theta} = x_{(n)}$. This can be verified by the following plot ▶

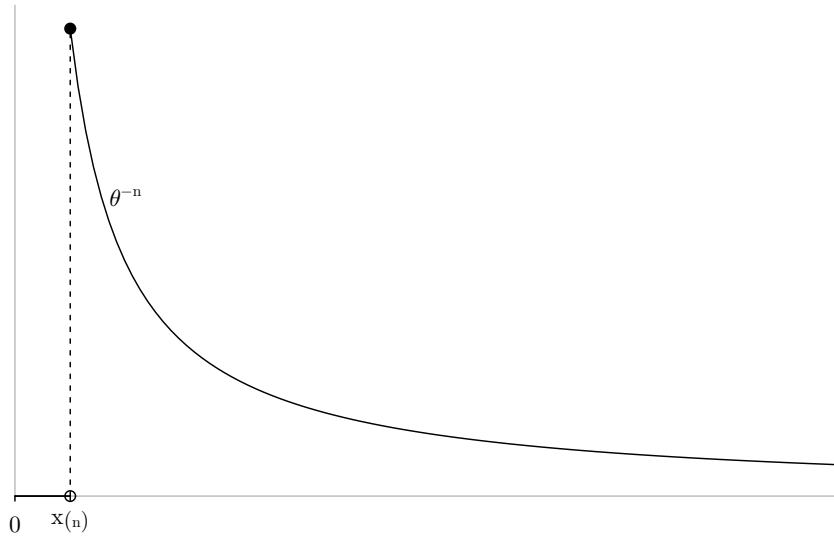

Figure 2.1: Demonstration of the MLE for Example 2.10.

Example 2.11. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, i.e., gamma with shape $\alpha > 0$ and rate $\beta > 0$. In this case there are two parameters to estimate, i.e., $\kappa = 2$. The pdf is given by

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Then,

$$L(\alpha, \beta|\mathbf{x}) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod x_i\right)^{\alpha-1} e^{-\beta \sum x_i},$$

so

$$\ell(\alpha, \beta|\mathbf{x}) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \log \left(\prod x_i\right) - \beta \sum x_i.$$

In this case we have a system of 2 equations: $\frac{d\ell}{d\alpha} = 0$ and $\frac{d\ell}{d\beta} = 0$:
$\frac{d\ell}{d\beta} = \frac{n\alpha}{\beta} - \sum x_i = 0$, and
$\frac{d\ell}{d\alpha} = n \log \beta - n\psi(\alpha) + \log\left(\prod x_i\right) = 0$, where $\psi(\alpha)$ denotes the *digamma function*, $\psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha)$.
From the first equation we have $\beta = (1/\alpha)\bar{x}$ so if α were known we could estimate β in this way. If α were unknown, we could substitute the expression for β into the second equation to get an equation in terms of α only. This becomes
$\log \bar{x} - \log \alpha - \psi(\alpha) + \sum \log x_i / n = 0$ which does not have a closed form solution. In this case the MLE for α can be obtained numerically. Once the solution is computed, say $\hat{\alpha}$, then it is plugged in the expression for β to get $\hat{\beta} = \bar{x}/\hat{\alpha}$. ▶

2.2 Confidence intervals

Consider a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$. An estimator $T = T(X_1, \ldots, X_n)$ of the parameter θ, whatever its properties, will provide only a point estimate, $\hat{\theta}$ which is likely to differ from the true value of θ. The point estimator does not provide any information about the deviation of our estimator from the true parameter value. Ideally we would like to provide a range of values which we believe to contain the true parameter value with some known probability. This range of values is called a **confidence interval** and the probability that the interval contains the parameter is called the **confidence level**.

Definition 2.5 (Confidence interval, confidence level)**.**
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\theta \in \Theta$. The random interval $[L, U]$ with bounds the statistics $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ such that $L \leq U$ and $L, U \in \Theta$ is called a **confidence interval** for the parameter θ. The number $1 - \alpha$, $\alpha \in (0, 1)$ is called the **confidence level** of the interval if for all $\theta \in \Theta$, the probability

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha.$$

Remark. Although in Definition 2.5 we define the confidence interval as a closed interval $[L, U]$, it will sometimes be more natural to quote the open interval (L, U) when the random variables L and U are continuous and $L < U$.

A general procedure for constructing a confidence interval for a given confidence level $1 - \alpha$, which is applicable in many problems can be summarised in the following steps.

1. Derive a point estimator $T = T(X_1, \ldots, X_n)$ of the parameter θ and come up with a **pivot function**
$$Y = g(T, \theta)$$
of T and θ whose distribution does not depend of θ.

2. Using the distribution of Y, derive two quantiles, c_1 and c_2 with $c_1 < c_2$ such that
$$\mathbb{P}(c_1 \leq Y \leq c_2) = 1 - \alpha,$$
i.e., the probability that Y falls within c_1 and c_2 is $1 - \alpha$, or equivalently, the probability that Y falls outside c_1 and c_2 is α, i.e.,
$$\mathbb{P}(Y < c_1) + \mathbb{P}(Y > c_2) = \alpha.$$
Note that the choice of c_1 and c_2 is not unique. We usually choose them so that
$$\mathbb{P}(Y < c_1) = \mathbb{P}(Y > c_2) = \alpha/2.$$

3. Rearrange the inequality $c_1 \leq g(T, \theta) \leq c_2$ with θ in the middle in the form $L \leq \theta \leq U$, where $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ do not depend on θ but do depend on c_1 and c_2. Then,
$$\mathbb{P}(L \leq \theta \leq U) = \mathbb{P}(c_1 \leq Y \leq c_2) = 1 - \alpha,$$
so $[L, U]$ is a confidence interval for θ with significance level $1 - \alpha$.

Example 2.12. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\mu)$, $\mu > 0$. As shown in Example 2.8, the MLE for μ is $\bar{X}$. To get a confidence interval, note that $\bar{X} = \sum X_i/n$ and the distribution of

$W = \sum X_i$ is Gamma(n, μ), i.e., with shape n and scale μ, so $Y = W/\mu \sim$ Gamma$(n, 1)$.

For a given significance level $1 - \alpha$, let c_1 and c_2 be the $\alpha/2$ and $1 - \alpha/2$ quantiles of Gamma$(n, 1)$ respectively which can be obtained in Python using
`scipy.stats.gamma.ppf([alpha/2,1-alpha/2],a=n,scale=1)`.

Then, $c_1 \leq \dfrac{W}{\mu} \leq c_2 \Rightarrow \dfrac{1}{c_2} \leq \dfrac{\mu}{W} \leq \dfrac{1}{c_1} \Rightarrow \dfrac{W}{c_2} \leq \mu \leq \dfrac{W}{c_1} \Rightarrow \dfrac{n\bar{X}}{c_2} \leq \mu \leq \dfrac{n\bar{X}}{c_1}$ is the confidence interval for μ. ▶

Example 2.13. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ and we wish to derive a confidence interval for μ. We will consider the case where σ^2 has a known value and where it is estimated.

Suppose σ^2 has a known value. An estimator for μ is $\bar{X}$. The distribution of $\bar{X}$ is $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$ so

$$Y = \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1),$$

which is a pivot quantity. Let z_p denote the argument in the CDF of the $\mathrm{N}(0, 1)$ distribution, $\Phi(z)$ such that $\Phi(z_p) = p$ (see Figure 2.2). This can by obtained using `scipy.stats.norm.ppf` in Python. Note that, because of the symmetry of the standard normal distribution around 0, $z_p = -z_{1-p}$.

If we let $c_1 = z_{\alpha/2} = -z_{1-\alpha/2}$, $c_2 = z_{1-\alpha/2}$, then
$-z_{1-\alpha/2} \leq \frac{\mu-\bar{X}}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \Rightarrow \bar{X} - z_{1-\alpha/2} \times \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \times \sigma/\sqrt{n}$ is a level $1 - \alpha$ confidence interval for μ with given value for σ^2.

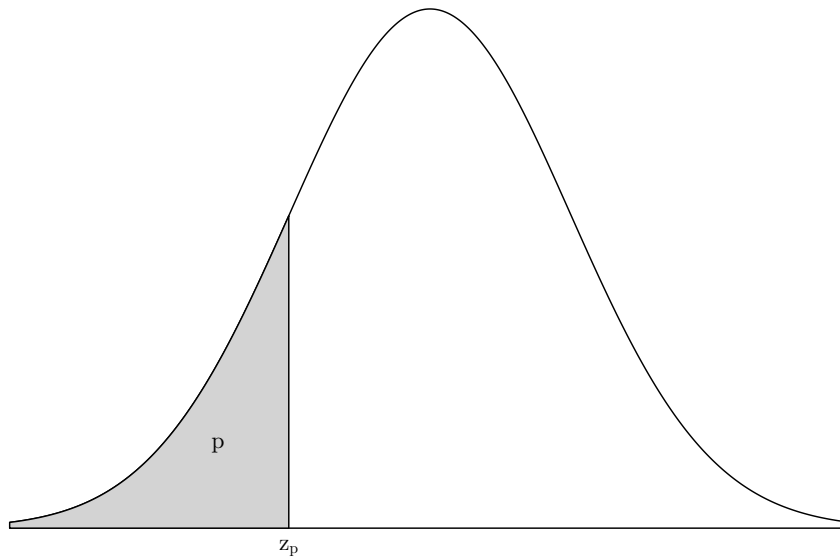

Figure 2.2: Illustration of the standard normal quantile z_p corresponding to left-tail probability p. The plotted curve corresponds to the $\mathrm{N}(0, 1)$ pdf and for given p, z_p satisfies $p = \Phi(z_p)$ where $\Phi(z)$ is the CDF of $\mathrm{N}(0, 1)$.

Suppose now that σ^2 has an unknown value. Recall that S^2 can be used as its estimator

and $(n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$. We consider the following statistic

$$Y = \frac{\mu - \bar{X}}{S/\sqrt{n}},$$

i.e., the same as before but with σ replaced by $S = \sqrt{S^2}$. To derive the distribution of Y, we use the following definition.

Definition 2.6 (Student's t_ν distribution)**.**
Let $Z \sim \mathrm{N}(0,1)$ and let $W \sim \mathcal{X}^2_\nu$ and suppose that Z and W are independent. Then the distribution of the random variable $Y = \frac{Z}{\sqrt{W/\nu}}$ is called the Student's t distribution with ν degrees of freedom, written as t_ν.

Proposition 2.2.
The t_ν distribution has similar shape as the $\mathrm{N}(0,1)$ distribution. In particular it is symmetric around 0 and converges to $\mathrm{N}(0,1)$ as $\nu \to \infty$.

In Python this distribution is given by `scipy.stats.t`. We can apply this definition in our problem. We know that $Z = \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ and that $W = (n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$ and that they are independent, so

$$\begin{aligned} Y &= \frac{Z}{\sqrt{W/(n-1)}} \\ &= \frac{\frac{\mu - \bar{X}}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} \\ &= \frac{\mu - \bar{X}}{S/\sqrt{n}} \sim \mathrm{t}_{n-1}. \end{aligned}$$

Proceeding similarly with the known-variance case, we let $c_1 = t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$ and $c_2 = t_{n-1;1-\alpha/2}$, i.e., the $\alpha/2$ and $1-\alpha/2$ quantiles of t_{n-1}, then $\bar{X} - t_{n-1;1-\alpha/2} \times S/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2} \times S/\sqrt{n}$ is a level $1-\alpha$ confidence interval for μ. ▶

2.3 Exercises

1. Consider the method of moments estimator of Example 2.6. Identify potential drawbacks of this estimator.

2. Verify the formula from the mgf of the gamma distribution in Example 2.7 and use it to derive its mean and variance.

3. Explain why the MLE of Example 2.10 is biased but do not derive its bias.

4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0,1)$. Derive the MLE for θ.

5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown parameters. Derive a level $1-\alpha$ confidence interval for σ^2.

6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{U}(0,\theta)$, $\theta > 0$. By considering an appropriate pivot construct a level $1-\alpha$ confidence interval for μ.