Inférence Statistique et Vraisemblance

LSTAT2040

Anouar El Ghouch

2024

Contents

Notations and Abbreviations

- Unless otherwise specified, vectors and matrices are emphasized using a **bold** font.
- Unless otherwise specified, *any vector that appears in a mathematical operation (e.g., addition, subtraction, and multiplication) is a* ***column vector***.
- Random variables are assigned to letters from the last part of the alphabet $(X, Y, Z, U, V, \ldots)$, while corresponding observations are assigned to lowercase letters $(x, y, z, u, v, \ldots)$.
- Constants are assigned to letters from the first part of the alphabet $(a, b, c, \ldots)$.
- The index ^{t} denote vector/matrix ***transpose***. For example, $\begin{pmatrix} m_1 & \ldots & m_d \end{pmatrix}^t = \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix}$.
- The symbol $:=$ indicates a ***assignment***. For example, if $x = 2$ and we write $x = y := 2$, then it means that x and y are equal, as y is *defined* to be 2. The symbol $\equiv$ denote ***equivalence***, i.e. things that have exactly the same meaning. These symbols will only be used when they are really relevant to the understanding of a particular concept or situation, otherwise "$=$" will be used.
- The *probability density function* (pdf) of a continuous random variable/vector, or the *probability mass function* (pmf) of a discrete random variable/vector, will be simply referred to as a *probability distribution* (**pd**), and will be typically denoted f.
- rv/rve: Random variable(s)/Random vector(s).
- cdf: Cumulative distribution function; typically denoted by F.
- iid: Independent and identically distributed rv or rve.
- $X \perp\!\!\!\perp Y$: X and Y are independent of each other.
- log and ln will be used interchangeably to refer to the ***natural logarithm***.
- Any integral sign $\int$, without limits, should be understood as $\int_{-\infty}^{\infty}$ in the univariate case and as $\int_{\mathbb{R}^d}$ in the multivariate case.
- The ***indicator function*** of a given set A is denoted by $I(x \in A)$. It takes the value 1 if $x \in A$, and 0 otherwise.
- $a = \arg\min_{x \in I} f(x)$ means that a is a value for which $x \mapsto f(x)$ reaches its minimum on I. The same applies to $\arg\max$.
- For a d-dimensional rve $\boldsymbol{X} = (X_1, \ldots, X_d)^t$. $\boldsymbol{\mu} = E(\boldsymbol{X})$ denote the vector of first moments (means), i.e. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^t$, with $\mu_j = E(X_j)$. The variance-covariance, or simply the variance, of $\boldsymbol{X}$ is the $d \times d$ symmetric matrix denoted by $Var(\boldsymbol{X}) := E\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^t\right) = E(\boldsymbol{X}\boldsymbol{X}^t) - \boldsymbol{\mu}\boldsymbol{\mu}^t$. The (j, k) element of this matrix is nothing but

$Cov(X_j, X_k)$. Note that, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two constant, then $E(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{A} + \boldsymbol{B}E(\boldsymbol{X})$, and $Var(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{B}Var(\boldsymbol{X})\boldsymbol{B}^t$.

- We will use the notation N_d to designate a multivariate normal distribution of dimension d, $d \geq 2$. For an univariate normal ($d = 1$), we simply write N (without any suffix).
- For scalar function $f(x) : \mathbb{R} \to \mathbb{R}$, the first ***derivative*** at $x = a$ is denoted $f'(a) = \frac{df(x)}{dx}\Big|_{x=a}$, and the second derivative is denoted by $f''(a)$. For $k \geq 3$, the k-th order derivative is denoted by $f^{(k)}(a)$.
- For a multivariate function $f(\boldsymbol{x}) = f(x_1, \ldots, x_d) : \mathbb{R}^d \to \mathbb{R}$:
  - The ***partial derivative***, at $\boldsymbol{x} = \boldsymbol{a}$, with respect to x_i is denoted by $\partial_i f(\boldsymbol{a}) = \partial_{x_i} f(\boldsymbol{a}) = \frac{\partial f(\boldsymbol{x})}{\partial x_i}\Big|_{\boldsymbol{x}=\boldsymbol{a}}$. Similarly, we denote the second-order partial derivative by $\partial_{ij} f(\boldsymbol{a}) = \partial_{x_i x_j} f(\boldsymbol{a}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\Big|_{\boldsymbol{x}=\boldsymbol{a}}$, and $\partial_i^2 f(\boldsymbol{a}) = \partial_{x_i}^2 f(\boldsymbol{a}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i^2}\Big|_{\boldsymbol{x}=\boldsymbol{a}}$.
  - The ***gradient***, at a point $\boldsymbol{a}$, is the vector given by $\boldsymbol{\nabla} f(\boldsymbol{a}) = (\partial_1 f(\boldsymbol{a}), \ldots, \partial_d f(\boldsymbol{a}))^t$.
  - The ***Hessian***, at a point $\boldsymbol{a}$, is the matrix given by $Hess f(\boldsymbol{a}) \equiv \boldsymbol{\nabla}^2 f(\boldsymbol{a}) = \left[\partial_{ij} f(\boldsymbol{a})\right]_{i,j=1,\ldots,d}$.

    For example, for $f(x_1, x_2) = x_1^2 + 2x_2^3 + x_1 x_2$, $\boldsymbol{\nabla} f(x_1, x_2) = (2x_1 + x_2, x_1 + 6x_2^2)^t$ and
    $\boldsymbol{\nabla}^2 f(x_1, x_2) = \begin{pmatrix} 2 & 1 \\ 1 & 12x_2 \end{pmatrix}$.
- For a multivariate vector-valued function $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x})) : \mathbb{R}^d \to \mathbb{R}^p$, where $f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, p$, the ***Jacobian*** is the matrix $\boldsymbol{J_f}(\boldsymbol{a}) \equiv \dot{\boldsymbol{f}}(\boldsymbol{a}) = \left[\partial_j f_i(\boldsymbol{a})\right]_{i=1,\ldots,p;j=1,\ldots,d}$.

  For example, for $\boldsymbol{f}(x_1, x_2) = (x_1^3, x_1^2 + 2x_2^3 + x_1 x_2, x_2^2)$, $\dot{\boldsymbol{f}}(x_1, x_2) = \begin{pmatrix} 3x_1^2 & 0 \\ 2x_1 + x_2 & x_1 + 6x_2^2 \\ 0 & 2x_2 \end{pmatrix}$.

Chapter 1

Parametric models and exponential families

1.1 Motivation and formalization

In order to obtain an estimate of an unknown quantity, say μ_0, say, for example, a speed, it is common to take n measurements $x_1, \ldots, x_n$ and calculate their mean:

$$\bar{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

But why should the observations be combined in this way ?

$\rightarrow$ the empirical mean is a relevant measure of the center of the observations, since

$$\bar{x}_n = \arg\min_a \sum_{i=1}^{n} (x_i - a)^2.$$

But (at this level) we can't justify why $\bar{x}_n$ is a good estimate (approximation) of the true value μ_0 since no explicit assumption has been made to connect the data $(x_1, \ldots, x_n)$ and μ_0.

To establish such a connection, we can, for example, presume that:

(i) each x_i is an observed value of a rv X_i, and

(ii) X_i, $i = 1, \ldots, n$, have a common mean μ_0.

Even more specifically, we can, for example, assume the following ***additive error model***

$$X_i = \mu_0 + \epsilon_i,\ \epsilon_i \sim N(0, \sigma_0^2),$$

or equivalently $X_i \sim N(\mu_0, \sigma_0^2)$. In this way, the problem we face is the estimation of $\mu_0 = E(X_i)$ from the sample $(X_1, \ldots, X_n)$.

1.1.1 General formalization

The data $\boldsymbol{x}^t = (x_1, \ldots, x_n)$ that we observe are believed to be generated by a rve $\boldsymbol{X}^t = (X_1, \ldots, X_n)$, which represents our random sample. We assume that $\boldsymbol{X}$ follows some joint distribution which is (partly) unknown. The set of assumptions made about this underlying joint distribution is what we call a ***statistical model***.

$X_1, \ldots, X_n$ are typically assumed to be ***iid*** copies of some population rv, which we denote hereafter by X. In this case the statistical model reduces to the set of assumptions about the distribution of X. To discribe this latter, we usually use a pd (i.e., probability density or mass function) f or a cdf (i.e., cumulative distribution function) F. Under the iid assumption, the joint pd and the joint cdf of $\boldsymbol{X}$, denoted by f_n and F_n, respectively, are given by

$$f_n(\boldsymbol{x}) = f_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \text{ and } F_n(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} F(x_i).$$

The IID hypothesis plays a crucial role

Without the iid assumption the statistical analysis of the data becomes much more complicated. For example, as a statistical model, we could assume that

$$\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with an unknown $\boldsymbol{\theta}^t = (\mu_1, \ldots, \mu_n, \sigma_{11}, \ldots, \sigma_{nn}) \in \mathbb{R}^{n+\frac{n\times(n+1)}{2}}$, where $\mu_i = E(X_i)$ and $\sigma_{ij} = Cov(X_i, X_j)$.

Now, under the iid assumption, $\boldsymbol{\theta}$ reduces to $(\mu_1, \sigma_{11}) \in \mathbb{R}^2$. □

1.1.2 Parametric models

A parametric model or parametric family is a set of distributions indexed by a ***finite*** dimensional parameter $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_d)$, $d \geq 1$. That is to say that the pd of X, *the rv that generated the data*, is known up to the parameter $\boldsymbol{\theta}$, which is unknown. In which case, we denote the pd of X by $f(x; \boldsymbol{\theta})$ and its cdf by $F(x; \boldsymbol{\theta})$, and write $X \sim f(x; \boldsymbol{\theta})$ or $X \sim F(x; \boldsymbol{\theta})$.

The set of possible values for the parameter $\boldsymbol{\theta}$, that we denote by $\boldsymbol{\Theta}$, is called ***the parameter space***. Among all the elements of $\boldsymbol{\Theta}$, the parameter value $\boldsymbol{\theta}_0$ that actually generates the data is referred to as the ***true value***. Now, to keep things simple, when this does not affect understanding, we leave off the subscript 0 in $\boldsymbol{\theta}_0$ and simply use $\boldsymbol{\theta}$ to refer to both the unknown generic parameter of interest and its true value.

Example 1.1.

- The Bernoulli model:

$$f(x; \theta) = \theta^x (1-\theta)^{(1-x)} I(x \in \{0, 1\}),\ \theta \in (0, 1).$$

Thus, $X \sim Ber(\theta_0)$, for some particular but unknown $\theta_0 \in [0,1] = \Theta$.

- The Exponential model:

$$f(x;\theta) = \theta^{-1} e^{-x/\theta} I(x > 0),\ \theta > 0.$$

Thus, $X \sim Exp(\theta_0)$, for some unknown $\theta_0 \in (0,\infty) = \Theta$.

- The Normal model:

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} I(x \in \mathbb{R}),\ \boldsymbol{\theta}^t = (\mu, \sigma^2) \in \mathbb{R} \times (0,\infty).$$

Thus, $X \sim N(\mu_0, \sigma_0^2)$, for some unknown $(\mu_0, \sigma_0^2) \in \mathbb{R} \times (0,\infty) = \boldsymbol{\Theta}$. □

If the distribution of X is not completely determined by a finite number of parameters, then the model is *nonparametric* or *semiparametric* (i.e., a mix of finite- and infinite-dimensional parameters).

Example 1.2.

- X has a pd f, with $\int f''(x)dx < \infty$ and/or $\int x^2 f(x)dx < \infty$.
- X has a symmetric distribution about 0, i.e. it has a pd satisfying $f(-x) = f(x)$, $\forall x$.
- X has a pd f satisfying $f(x;\theta) = f_0(x-\theta)$, where θ is an unknown parameter and f_0 an unknown pd symmetric about 0.

These models/distributions cannot be indexed by a finite dimensional parameter. The first two cases are examples of (fully) nonparametric models, while the last case is an example of a semiparametric model. □

Parametric or nonparametric inference ?

Let us assume that our objective is to estimate $\theta := F(s) = P(X \le s) \in [0,1]$ for a given s.

- Without making any assumption about the distribution of X, a reasonable estimator is the empirical cdf $\hat{F}(s) = n^{-1} \sum_i I(X_i \le s)$.
- If we assume that $X \sim N(\mu, 1)$, then it is more reasonable to estimate θ by $\Phi(s - \hat{\mu})$, where Φ is the cdf of a $N(0,1)$ and $\hat{\mu}$ is any given estimator of μ.

  $\rightarrow$ Incorrect assumptions and restrictions on the underlying distribution of X will lead to wrong or ***biased*** conclusions. But correct assumptions lead to better and more ***efficient*** estimation. □

If there is no $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $X \sim f(x;\boldsymbol{\theta})$, then the model $\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is said to be ***misspecified***. An example of misspecification is when a normal distribution is used for exponential data. A conclusion drawn from a statistical model is ***valid only if the chosen model is correctly specified***. In reality, no model is 100% correct, but some models are more useful than others in approximating the true underlying distribution of the data. Here,

unless explicitly stated, we always assume that any model we fit to the data is correctly specified.

1.1.3 Identifiability

For a given statistical model, a given parameter $\boldsymbol{\theta}$ corresponds to a single distribution. However, this does not rule out the possibility that there may exist $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ such that $f(x;\boldsymbol{\theta}_1) = f(x;\boldsymbol{\theta}_2)$, $\forall x$. In this case, we cannot distinguish between these two parameter values, even if we are given an infinite sample, since both will yield data which are distributed identically. We refer to such situation as an ***identifiability problem***.

Identifiability is an important property of a statistical model, which determines whether the parameter of the model can be recovered (estimated) from the observed data, which is only possible if different values of $\boldsymbol{\theta}$ lead to different distributed samples. Mathematically, this can be formulated by saying that, in a given model $\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, the parameter $\boldsymbol{\theta}$ (or the model) is identifiable, if, $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,

$$f(x;\boldsymbol{\theta}_1) = f(x;\boldsymbol{\theta}_2),\ \forall x \Rightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

Example 1.3.

- The Bernoulli, the Exponential, and the Normal models, as defined above, are identifiable.
- Let $X = \mu_1 + \epsilon$, where $\epsilon \sim N(\mu_2, 1)$ and μ_1 and μ_2 are unknown. Suppose that we observe X (and not ϵ), then $\theta = \mu_1 + \mu_2$ is identifiable but $\boldsymbol{\theta} = (\mu_1, \mu_2)$ is not.
- Let $X = |Y|$, where $Y \sim N(\mu, 1)$ and μ is unknown. Suppose that we observe X, then μ is not identifiable.

Let's verify the identifiability of the last Normal model. For that, observe that $f(x, \boldsymbol{\theta}_1) = f(x, \boldsymbol{\theta}_2)$, $\forall x$, is equivalent to

$$\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2} = 2\log\frac{\sigma_2}{\sigma_1},\ \forall x$$

Since a parabolic function $ax^2 + bx + c$ becomes null if and only if $a = b = c = 0$, and since in our case $a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$, we have that $\sigma_1 = \sigma_2$. As consequence, $f(x;\boldsymbol{\theta}_1) = f(x;\boldsymbol{\theta}_2)$, $\forall x$, is equivalent to

$$(x-\mu_1)^2 - (x-\mu_2)^2 = 0,\ \forall x \iff \mu_1 = \mu_2.$$

To verify the last example above (with $X = |Y|$), observe that $P_\mu(X \leq x) = \Phi(x-\mu) + \Phi(x+\mu) - 1$, where Φ is the cdf of $N(0,1)$. It follows that, $P_1(X \leq x) = P_{-1}(X \leq x)$, $\forall x$, i.e., $\mu = 1$ and $\mu = -1$ lead to the same distribution for X. This demonstrates that μ is not identifiable. □

If a model is not identifiable, it is common to introduce additional constraints/assumptions on it in order to make it identifiable. In that case, the set of these requirements is called the ***identifiability conditions.*** For instance, in our example above with $X = |Y|$, if we assume that $\mu > 0$, i.e., $\Theta = (0, \infty)$, then μ becomes identifiable; can you prove this ?

1.1.4 Purpose of inferential statistics

Statistical inference is the process of learning about a given probability model using observed data. To be more precise, suppose we are given a data set $\boldsymbol{x}$ which we assume to be generated from the model $\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. The aim of parametric statistical inference is to gain knowledge about the unknown parameter $\boldsymbol{\theta}$ from $\boldsymbol{x}$.

There are three major parametric statistical inference procedures:

1. **Point estimation**: A single value is computed from the data $\boldsymbol{x}$ and used as an estimate (approximation) of the the true parameter value $\boldsymbol{\theta}_0$.
2. **Hypothesis testing**: Sets up some specific hypotheses regarding $\boldsymbol{\theta}_0$ and assesses whether or not the data $\boldsymbol{x}$ support these hypothesis.
3. **Confidence set estimation**: Use the observed data $\boldsymbol{x}$ to construct a set of possible values for $\boldsymbol{\theta}_0$. The resulting set must have a high (predetermined) probability of including the true value.

Other well-known topics in statistical inference include *model selection*, *model validation* and *prediction*.

1.2 Exponential family

One important class of statistical models is exponential family models. These models are widely used in statistics and machine learning. They are characterized by a simple and elegant mathematical structure, which makes them analytically tractable and computationally efficient. Exponential family contains most of the standard discrete and continuous distributions that are used for modeling, such as (multivariate) normal, poisson, binomial, multinomial, exponential, and gamma.

The reason for the special status of the exponential family is that a number of important and useful results in inference can be unified within it. This family also forms the basis for an important class of regression models, known as generalized linear models.

1.2.1 One-parameter exponential family

A family of probability distributions that depend on a single (scalar) parameter θ is a one-parameter exponential family if it can be expressed as

$$f(x;\theta) = h(x)\exp(g(\theta)T(x) - B(\theta)),\ \ \forall x.$$

Here x can be a scalar or vector, $h(x) \geq 0$ and $T(x)$ are a functions of x only (*cannot depend on* θ), and $g(\theta)$ and $B(\theta)$ are functions of θ only (*cannot depend on* x). $B(\theta)$ is a normalizing constant, ensuring that $f(x;\theta)$ sums or integrates to 1. The set $\Theta = \{\theta : \int h(x) \exp(g(\theta)T(x))dx < \infty\}$ is the parameter space of the family. $T(X)$ is referred to as ***natural sufficient*** statistic or simply natural statistic.

Note that the above parameterization is not unique since, for example, g could be multiplied by a nonzero constant a if T is divided by a. Also, in many cases $T(x) = x$.

An exponential family can be reparameterized as

$$h(x) \exp(\eta T(x) - A(\eta)).$$

This expression is called the ***canonical (or natural) representation***, and $\eta = g(\theta)$ is the ***canonical parameter***. Here $h(x)$ and $T(x)$ are the same as in the original parameterization, and $A(\eta) = B\big(g^{-1}(\eta)\big)$ is the new normalizing constant (assuming g is invertible). Notice that this parametrization is not unique either.

It's analytically convenient and easier to work with an exponential family in its canonical form. Once a result has been derived for the canonical form, we can rewrite it in terms of the original parameter θ if desired.

To verify that a family of pd's is an exponential family, we must identify all the functions h, T, g, and B (or A). The next example illustrates this.

Example 1.4.

- Poisson:

$$\begin{aligned} \frac{\theta^x e^{-\theta}}{x!} &= \frac{1}{x!} \exp(x \log(\theta) - \theta), \ x = 0, 1, \ldots, \text{ and } \theta > 0 \\ &\equiv \frac{1}{x!} \exp(x\eta - e^{\eta}), \ \eta \in (-\infty, \infty). \end{aligned}$$

- Binomial:

$$\begin{aligned} C_n^x \theta^x (1-\theta)^{n-x} &= C_n^x \exp\left(x \log\left(\frac{\theta}{1-\theta} \right) + n \log(1-\theta) \right), \ x = 0, 1, \ldots, n \text{ and } \theta \in (0,1) \\ &\equiv C_n^x \exp(x\eta - n \log(1 + e^{\eta})), \ \eta \in (-\infty, \infty). \end{aligned}$$

- Normal with a *known* σ ($\theta = \mu$):

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2} \right) \exp\left(\frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2} \right), \ x \in (-\infty, \infty) \text{ and } \mu \in (-\infty, \infty) \\ &\equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2} \right) \exp\left(\eta x - \frac{\eta^2 \sigma^2}{2} \right), \ \eta \in (-\infty, \infty). \ \square \end{aligned}$$

Example 1.5 (Counter-example)**.** The following density

$$\begin{aligned} f(x;\theta) &= \exp(-(x-\theta)),\ \ x \geq \theta \text{ and } \theta \in (-\infty, \infty) \\ &= \exp(-(x-\theta))I(x \geq \theta) \end{aligned}$$

is not an exponential family because of $I(x \geq \theta)$ that cannot be factored into "x" part and an "θ" part. □

In general, for a pd $f(x;\theta)$ to be an exponential family, its *support* $S = \{x : \ f(x;\theta) > 0\}$ *must be free from* θ.

1.2.2 Properties of exponential family

A special fact about the exponential family is that the integral $\int h(x)\exp(g(\theta)T(x) - B(\theta))dx$, in the continuous case, or the sum $\sum_x h(x)\exp(g(\theta)T(x) - B(\theta))$, in the discrete case, can be differentiated, with respect to θ (or with respect η for the canonical representation), any number of times under the integral/sum sign (i.e. the derivatives can always be moved inside the integral/sum). This is the source of many interesting results about the exponential family. One of these results is given in the following.

Proposition 1.1. *With the canonical parameterization, the mean and variance of* $T \equiv T(X)$ *are given by*

$$E_\eta(T) = A^{'}(\eta) \text{ and } Var_\eta(T) = A^{''}(\eta). \tag{1.1}$$

For example, for $N(\mu, \sigma^2)$, given the results above (see Example 1.4), we see that

$$E(X) = \partial_\eta\left(\frac{\eta^2\sigma^2}{2}\right) = \eta\sigma^2 = \mu \ \text{ and } \ Var(X) = \partial_\eta\left(\eta\sigma^2\right) = \sigma^2$$

To show the first equality in (1.1), we can differentiate $\eta \mapsto \int h(x)\exp(\eta T(x) - A(\eta))dx = 1$, with respect to η, and then use the fact that the derivative can be moved inside the integral, which gives

$$\begin{aligned} 0 &= \int \partial_\eta h(x)\exp(\eta T(x) - A(\eta))dx \\ &= \int h(x)(T(x) - A^{'}(\eta))\exp(\eta T(x) - A(\eta))dx \\ &= E_\eta(T) - A^{'}(\eta). \end{aligned}$$

As for the second equality, we can differentiate two times to get

$$\begin{aligned}
0 &= \int \partial_\eta h(x)(T(x) - A^{'}(\eta)) \exp(\eta T(x) - A(\eta))dx \\
&= \int h(x)(-A^{''}(\eta)) \exp(\eta T(x) - A(\eta))dx + \int h(x)(T(x) - A^{'}(\eta))^2 \exp(\eta T(x) - A(\eta))dx \\
&= -A^{''}(\eta) + E_\eta(T - A^{'}(\eta))^2.
\end{aligned}$$

In terms of the original parameterization, with θ, we can write

$$E_\theta(T) = \frac{B^{'}(\theta)}{g^{'}(\theta)} \text{ and } Var_\theta(T) = \frac{\partial_\theta E_\theta(T)}{g^{'}(\theta)}.$$

These can be proven directly from the definition of $f(x;\theta)$ by following the same derivation as we did above for the canonical parameterization. Another way to obtain these results is to apply the chain rule to (1.1). In fact, the definitions given above yield to $A'(\eta) = \frac{B'}{g'}\Big(g^{-1}(\eta)\Big)$.

Attention. In the formulas given above, we used the subscripts θ and η to indicate the parameterization used when calculating the expect values and the variances. This notation will be used wherever relevant hereafter.

Another interesting fact about the exponential family is that its structure is preserved under iid sampling. This is better explained in the following.

Proposition 1.2. *If $X_1, \ldots, X_n$ are iid rv from the exponential family, as defined above, with a natural statistic T then the the joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$:*

$$\left[\prod_{i=1}^{n} h(x_i)\right] \exp\left(\eta \sum_{i=1}^{n} T(x_i) - nA(\eta)\right)$$

is also an exponential family with natural statistic $\sum_{i=1}^{n} T(X_i)$.

1.2.3 Multiparameter exponential family

The multiparameter version of the exponential family is given by

$$\begin{aligned}
f(x;\boldsymbol{\theta}) &= h(x) \exp\left(\sum_{j=1}^{J} g_j(\boldsymbol{\theta}) T_j(x) - B(\boldsymbol{\theta})\right) \\
&\equiv h(x) \exp\big(\boldsymbol{\eta}^t \boldsymbol{T}(x) - A(\boldsymbol{\eta})\big), \qquad \text{(Canonical parametrization)}
\end{aligned}$$

where $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_J)$, $\boldsymbol{\eta}^t = (\eta_1, \ldots, \eta_J)$, with $\eta_j = g_j(\boldsymbol{\theta})$, $\boldsymbol{T}^t(x) = (T_1(x), \ldots, T_J(x))$, and $g_j, A, B : \mathbb{R}^J \to \mathbb{R}$.

The properties that we have seen above for the one-parameter case also apply to the multi-

parameter case. For example, it can be shown that

$$E(T_j(X)) = \partial_{\eta_j} A(\boldsymbol{\eta}), \text{ and } Cov(T_j(X), T_k(X)) = \partial_{\eta_j \eta_k} A(\boldsymbol{\eta}).$$

In matrix form, we can write $E(\boldsymbol{T}) = \boldsymbol{\nabla} A(\boldsymbol{\eta})$ and $Var(\boldsymbol{T}) = \boldsymbol{\nabla}^2 A(\boldsymbol{\eta})$.

Example 1.6. Normal distribution with an unknown μ and σ ($\boldsymbol{\theta} = (\mu, \sigma^2)$):

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma)\right)\right) \\ &\equiv \frac{1}{\sqrt{2\pi}} \exp\left(\eta_1 x + \eta_2 x^2 - A(\boldsymbol{\eta})\right),\end{aligned}$$

where $A(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$ Here $T_1(x) = x$ and $T_2(x) = x^2$.

It follows that,

$$\begin{aligned}\partial_1 A = -\frac{\eta_1}{2\eta_2} = \mu, \quad \partial_2 A = \frac{\eta_1^2 - 2\eta_2}{4\eta_2^2} = \mu^2 + \sigma^2 = E(X^2) \\ \partial_1^2 A = -\frac{1}{2\eta_2} = \sigma^2, \quad \partial_2^2 A = \frac{\eta_2 - \eta_1^2}{2\eta_2^3} = 2\sigma^2(\sigma^2 + 2\mu^2) = Var(X^2),\end{aligned}$$

and $\partial_{1,2} A(\boldsymbol{\eta}) = \frac{\eta_1}{2\eta_2^2} = 2\mu\sigma^2 = Cov(X, X^2)$. □

1.3 Some useful tools

Law of total expectation/variance

For any two rv X and Y,

$$\begin{aligned}E(Y) &= E\left(E[Y|X]\right), \\ Var(Y) &= E\left(Var[Y|X]\right) + Var\left(E[Y|X]\right),\end{aligned}$$

where $Var[Y|X] = E\left[(Y - E[Y|X])^2|X\right] = E[Y^2|X] - (E[Y|X])^2$.

Expected value of a non-negative rv

If X be a non-negative rv, then $E(X) \geq 0$. Moreover, $E(X) = 0$ if and only if $X = 0$ (with probability 1).

Markov-Chebyshev's inequality

If X is a non-negative rv, then $E(X) \geq kP(X \geq k)$, $\forall k \in \mathbb{R}$. As a consequence, for any rv X and any constant $k > 0$, $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$, where $\mu = E(X)$ and $\sigma = \sqrt{Var(X)}$.

To see the first inequality, observe that, $\forall X, \forall k$, $X = XI(X \geq k) + XI(X < k)$. Which, by the fact that $X \geq 0$, implies that, $X \geq kI(X \geq k)$, and hence $E(X) \geq kP(X \geq k)$. Applying this last equality to $(X - \mu)^2$, instead of X, we obtain the second inequality.

Jensen's inequality

If g is a convex function (Reminder: $g''(x) \geq 0, \forall x \in I \Rightarrow g$ is convex in I), then

$$E(g(X)) \geq g(E(X))$$

Moreover, if g is strictly convex ($g'' > 0$), then this inequality is strict unless X is constant. The opposite hold for concave function (Reminder: g is concave if and only if $-g$ is convex).

Example. Let X be a non-constant rv. Since $x \mapsto |x|^a$ is (strictly) convex for $a = (>)1$, we have that, for example, $E|X| \geq |E(X)|$ and $E(X^2) > (E(X))^2$. And since $x \mapsto \sqrt{x}$ and $x \mapsto \log(x)$ are strictly concave in $(0, \infty)$, we have that $E\left(\sqrt{X}\right) < \sqrt{E(X)}$ and $E(\log X) < \log(EX)$, provided that $X > 0$. □

Cauchy-Schwarz's inequality

For any two rv X and Y,

$$(E(XY))^2 \leq E(X^2)E(Y^2)$$

As a consequence, we get the inequality

$$(Cov(X, Y))^2 \leq Var(X)Var(Y).$$

Moreover, if X and Y are not constant, then the last inequality becomes an equality if and only if X and Y are linearly dependent, i.e. if and only if there exists numbers $a \neq 0$ and b such that $Y = aX + b$, with probability one.

Taylor's theorem

Suppose f is a function such that $f^{(n+1)}$ $(n \geq 0)$ is continuous on some interval I. Then, for any $x, a \in I$, there exists a $\theta \in [0, 1]$ such that

$$f(x) = \sum_{i=0}^{n} \frac{(x-a)^i}{i!} f^{(i)}(a) + \frac{(x-a)^{(n+1)}}{(n+1)!} f^{(n+1)}(a + \theta(x-a))$$

We can use this result to approximate the function f, and write that, in a sufficiently small neighbourhood of a,

$$f(x) \approx \sum_{i=0}^{n} \frac{(x-a)^i}{i!} f^{(i)}(a) = f(a) + (x-a)f(a) + \frac{(x-a)^2}{2} f'(a) + \ldots + \frac{(x-a)^n}{n!} f^{(n)}(a).$$

This is called the the nth order Taylor polynomial approximation of f around a.
A similar result holds for functions of several variables. For example, the second-order Taylor polynomial approximation of $f : \mathbb{R}^d \mapsto \mathbb{R}$ around a point $\boldsymbol{a}$ is

$$f(\boldsymbol{x}) \approx f(\boldsymbol{a}) + \boldsymbol{\nabla}^t f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a})^t \boldsymbol{\nabla}^2 f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}).$$

Definite matrix

Definte matrices play a very important role in statistics and optimization ....

Let $\boldsymbol{A}$ be a *symmetric* matrix ($\boldsymbol{A}^t = \boldsymbol{A}$). $\boldsymbol{A}$ is said to be

- *positive definite*, if $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} > 0, \forall \boldsymbol{x} \neq \boldsymbol{0}$.
- *positive semidefinite*, if $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} \geq 0$, $\forall \boldsymbol{x}$.

If the inequalities are reversed, then $\boldsymbol{A}$ is *negative definite* or *negative semidefinite*, respectively.

Here are some examples. The matrix $\boldsymbol{A} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$ is positive definite, since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = 3x_1^2 + 2x_2^2 > 0$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is positive semidefinite since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = (x_1 + x_2)^2 \geq 0$, $\forall \boldsymbol{x}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ is positive definte since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = 2(x_1^2 - x_1 x_2 + x_2^2) > 0$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ is *indefinite* since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = x_1^2 + 4x_1 x_2 + x_2^2$ can be positive or negative.

Different methods exist to check if a matrix is positive definite, such as the Cholesky decomposition, the eigenvalues, or the principal minors. In this course we will not go into the details of these methods. It is sufficient to know that a 2×2 matrix $\boldsymbol{A} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is positive definite if and only if $a > 0$ and $det(\boldsymbol{A}) := ac - b^2 > 0$.

There also are many interesting properties of positive (semi)definite matrices, such as the fact that :

- A positive definite matrix is invertible and its inverse is also positive definite.
- For a positive (semi)definite matrix $\boldsymbol{A}$, there exists a unique positive (semi)definite matrix $\boldsymbol{B}$ such that $\boldsymbol{B} \times \boldsymbol{B} = \boldsymbol{A}$. This $\boldsymbol{B}$ is called the (natural) square-root of $\boldsymbol{A}$ and is denoted by $\boldsymbol{A}^{1/2}$.

Some interesting properties of the Variance-Covariance matrix

Let $\boldsymbol{X}$ be a rve in $\mathbb{R}^d$ and $\boldsymbol{\Sigma} = Var(\boldsymbol{X}) = \big(Cov(X_j, X_k)\big)_{1 \leq j,k \leq d}$. $\boldsymbol{\Sigma}$ is symmetric ($\boldsymbol{\Sigma}^t = \boldsymbol{\Sigma}$) and positive semidefinite. This last property follows directly from the fact that $\boldsymbol{a}^t \boldsymbol{\Sigma} \boldsymbol{a} = Var\big(\boldsymbol{a}^t \boldsymbol{X}\big)$.

$\boldsymbol{\Sigma}$ is positive definite if and only if the components of $\boldsymbol{X}$ are linearly independent (with

probability 1); i.e. it is not possible to express any component of $\boldsymbol{X}$ (i.e. X_1 or X_2, ...) as a linear combination of the others.

Some properties of the multivariate normal

- If $\boldsymbol{B}$ is a $p \times d$ matrix, $\boldsymbol{a}$ is a $p-$dimensional vector, and $\boldsymbol{b}$ a $d-$dimensional vector, then

$$\boldsymbol{b}^t \times N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\boldsymbol{b}^t\boldsymbol{\mu}, \boldsymbol{b}^t\boldsymbol{\Sigma}\boldsymbol{b}\right), \text{ and } \boldsymbol{a} + \boldsymbol{B} \times N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N_p\left(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^t\right).$$

- $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_d(\boldsymbol{0}, \mathbb{1})$, where $\boldsymbol{\Sigma}^{-1/2}$ is the square-root of a of $\boldsymbol{\Sigma}^{-1}$, $\mathbb{1}$ is the identity matrix, and $\boldsymbol{0}$ the $d-$dimensional vector of zeros.
- $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \mathbb{1}) \Rightarrow \boldsymbol{X}^t\boldsymbol{X} \sim \chi^2_d$, where χ^2_d the chi-squared distribution with d degrees of freedom.

Chapter 2

Basic concepts of point estimation

2.1 Statistic, estimator and estimation

Definition 2.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample. Any (measurable) function $T(\boldsymbol{X})$ of $\boldsymbol{X}$, i.e. any quantity that can be calculated solely from the observed data, is called a ***statistic***.

An estimator is any statistic used to estimate a given parameter. Typically, we use the notation $\hat{\theta}_n(\boldsymbol{X}) \equiv \hat{\theta}_n \equiv \hat{\theta}$ to denote an ***estimator*** of θ.

Any realization $\hat{\theta}_n(\boldsymbol{x})$ of $\hat{\theta}_n(\boldsymbol{X})$ is an ***estimation*** (a guess) of θ.

Example 2.1 (Example of statistics)**.**

X_1, $(X_1, \ldots, X_n)$, $\sum_i X_i$, $\overline{X}_n = n^{-1} \sum_i X_i$, $n^{-1} \sum_i X_i^2$, $n^{-1} \sum_i I(X_i \geq 0)$,

$X_{(1)} = \min_i X_i$, $X_{(n)} = \max_i X_i$, $X_{(k)}$ the kth order statistic, i.e. kth-smallest value,

$\tilde{\sigma}_n^2 = n^{-1} \sum_i (X_i - \mu)^2$ (assuming that μ is known), $S_n^2 = (n-1)^{-1} \sum_i (X_i - \overline{X}_n)^2$, $(\overline{X}_n, S_n^2)$,

$\arg\min_a \sum_i |X_i - a|$. □

To assess the usefulness of an estimation procedure, we examine the properties of the estimator used. Being random, any estimator has a distribution that is referred to as the ***sampling distribution***. The properties of the sampling distribution determine which estimator is best for a particular problem and which estimator to avoid.

Example 2.2 (Uniform model)**.**

Consider the uniform density $f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta);\ \theta > 0$.
As an estimator of θ, based on a sample $X_1, \ldots, X_n$, one can consider one of the following

estimators:

$$\hat{\theta}_1 = X_{(n)}, \quad \hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$$
$$\hat{\theta}_3 = X_{(1)} + X_{(n)}, \quad \hat{\theta}_4 = 2\overline{X}_n$$
$$\hat{\theta}_5 = 2\hat{q}_{0.5},$$

where $\hat{q}_{0.5}$ is the sample median, i.e.

$$\hat{q}_{0.5} = \begin{cases} X_{(k+1)} & \text{if } n = 2k+1 \text{ is odd,} \\ \frac{X_{(k)} + X_{(k+1)}}{2} & \text{if } n = 2k \text{ is even. } \square \end{cases}$$

This example suggests questions like:

- If many estimators are available, how can we compare them?
- Are there general methods for constructing estimators?
- How to find the best possible estimator for a given model?

These questions (and many others of the same nature) will be the subject of our next readings.

2.2 Risk and loss function

It seems reasonable that we want an estimate $\hat{\theta}$ which generally comes quite close to the true value of θ, and dislike an estimate $\hat{\theta}$ which generally misses the true value of θ by a large amount. The question is how to make this precise and quantifiable?

We quantify the idea of $\hat{\theta}$ being close to θ, by measuring the ***risk***, that is the *average distance*, between these two quantities. The distance is measured using what is called a ***loss function***.

Examples of loss functions include:

- squared error loss: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
- absolute error loss: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$
- absolute relative loss: $L(\hat{\theta}, \theta) = |\hat{\theta}/\theta - 1|$

Once the loss function is chosen, we calculate the risk as follows

$$E_\theta(L(\hat{\theta}, \theta)) = \int L(\hat{\theta}(\boldsymbol{x}), \theta) f_n(\boldsymbol{x}, \theta) d\boldsymbol{x},$$

where E_θ means that *the expectation is taken under the assumption that θ is the true parameter*; that is, the pd of $\boldsymbol{X}$ is $f_n(\boldsymbol{x}, \theta)$.

2.3 MSE, bias, variance, and relative efficiency

If the squared error loss is used, then the risk is the ***mean squared error***:

$$\begin{aligned} MSE_\theta(\hat{\theta}) &:= E_\theta[(\hat{\theta}-\theta)^2] \\ &= E_\theta\big[(\hat{\theta}-E_\theta(\hat{\theta}))+(E_\theta(\hat{\theta})-\theta)\big]^2 \\ &= Bias_\theta^2(\hat{\theta}) + Var_\theta(\hat{\theta}), \end{aligned}$$

where $Bias_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$ is the bias of the estimator $\hat{\theta}$.

In the following, in order to ease the notation, if no confusion is possible, we drop the index θ and write E, Var and MSE instead of E_θ, Var_θ and MSE_θ, respectively.

A large bias indicates low accuracy ($\hat{\theta}$ lies far from θ, i.e. some systematic error), while a large variance indicates low precision (too much fluctuation). If the bias of $\hat{\theta}$ is always zero, i.e. $Bias(\hat{\theta}) = 0\,\forall\theta \in \Theta$, then $\hat{\theta}$ is called ***unbiased***. This means that ***on average*** the estimator will yield the true value of the unknown parameter (whatever the true value is). *In this case, the MSE reduces to variance.*

The choice of an estimator is very often restricted to the class of unbiased estimators. But there are cases where a small bias is accepted, in particular if the bias converges to zero when the sample size tends to infinity. Moreover, there are cases where no unbiased estimator exists.

Example 2.3 (Unbiased does not necessarily mean a good estimator)**.**

Let X_i, $i = 1, \ldots, n$, be iid rv with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then X_1, $\overline{X}_n$ and $\frac{X_1+\overline{X}_n}{2}$ are unbiased estimators of μ. Which one should we use?

It is clear that all three are unbiased. So to compare these estimators, we have to compare their variances (i.e. their MSE). It easy to see that

$$Var(X_1) = \sigma^2,\ \ Var(\overline{X}_n) = \frac{\sigma^2}{n},\ \text{ and } Var\left(\frac{X_1+\overline{X}_n}{2}\right) = \frac{1}{4}\left(1+\frac{3}{n}\right)\sigma^2.$$

To check the last equality, observe that

$$Var(X_1+\overline{X}_n) = Var\left(\frac{n+1}{n}X_1 + \frac{1}{n}\sum_{i=2}^{n} X_i\right) = \frac{(n+1)^2}{n^2}\sigma^2 + \frac{n-1}{n^2}\sigma^2 = \left(1+\frac{3}{n}\right)\sigma^2.$$

$\rightarrow \overline{X}_n$ is better than the other two. □

Bias and transformation

If $\hat{\theta}$ is an unbiased estimator of θ, then $a + b\hat{\theta}$ is unbiased estimator for $a + b\theta$. In general, however, if $\hat{\theta}$ is an unbiased estimator of θ, then $g(\hat{\theta})$ is not necessarily an unbiased estimator of $g(\theta)$.

For example, we know that $\overline{X}_n$ is an unbiased estimator of μ, but, by Jensen's inequality, $E\left(\overline{X}_n^2\right) > \left(E\left(\overline{X}_n\right)\right)^2 = \mu^2$. Thus, $\overline{X}_n^2$ is a biased estimator of μ^2.

It is sometimes easy to adjust a biased estimator and transform it into an unbiased one. For example, if we known that $E(\hat{\theta}) = a\theta + b$ then $(\hat{\theta} - b)/a$ is unbiased for θ.

Example 2.4. Let X_i, $i = 1, \ldots, n$, be iid rv with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let's find an unbiased estimator for μ^2. To do so, observe that

$$E\left(\overline{X}_n^2\right) = \left(E\left(\overline{X}_n\right)\right)^2 + Var\left(\overline{X}_n\right) = \mu^2 + \frac{\sigma^2}{n}.$$

This means that $Bias\left(\overline{X}_n^2\right) = \sigma^2/n$. This also implies that an unbiased estimator of μ^2 is given by

$$\overline{X}_n^2 - \frac{S_n^2}{n},$$

provided that S_n^2 is an unbiased estimator of σ^2 (see the next example). □

Example 2.5 (The sample variance)**.** Let X_i, $i = 1, \ldots, n$, be iid rv with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. A natural estimator of this latter is given by $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. We have that

$$\begin{aligned} E(\hat{\sigma}_n^2) &= \frac{1}{n}\sum_{i=1}^n \left(E(X_i^2) + E\left(\overline{X}_n^2\right) - \frac{2}{n}\sum_{j=1}^n E(X_i X_j) \right) \\ &= (\sigma^2 + \mu^2) + \left(\frac{\sigma^2}{n} + \mu^2\right) - \frac{2}{n}\left(\sigma^2 + \mu^2 + (n-1)\mu^2\right) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

With the correction factor $\frac{n}{n-1}$ we obtain an unbiased estimator, namely the well-known empirical variance (or sample variance):

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2. \square$$

We can compare the quality of two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ by looking at the ratio of their MSE and we call this quantity the ***relative efficiency*** of $\hat{\theta}_1$ to $\hat{\theta}_2$:

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}.$$

This ratio is particularly meaningful if both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased or have about same bias. For unbiased estimators, RE reduces to

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}.$$

If this quantity is less than 1 $\forall\theta \in \Theta$, then $\hat{\theta}_1$ has a larger variance than $\hat{\theta}_2$, and the latter is said to be *more efficient* than the former.

Example 2.6.

Let $X_1, \ldots, X_n$, be an iid sample form $Unif[0, \theta]$. Let $\hat{\theta}_1 = 2\overline{X}_n$ and $\hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$, two estimators of θ.

Recall that an uniform distribution in $[a, b]$ is characterized by its cdf

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b. \end{cases}$$

Its mean and variance are given by

$$E(X) = \frac{a+b}{2} \text{ and } Var(X) = \frac{(b-a)^2}{12}.$$

From this, it follows that $\hat{\theta}_1$ is unbiased and $Var(\hat{\theta}_1) = \frac{\theta^2}{3n}$.

To find the expectation and variance of $\hat{\theta}_2$ first observe that

$$P(X_{(n)} \leq x) = (P(X_1 \leq x))^n.$$

So the cdf of $X_{(n)}$ is given by

$$F_{X_{(n)}}(x) = \begin{cases} 0, & \text{if } x < 0 \\ (x/\theta)^n, & \text{if } 0 \leq x \leq \theta \\ 1, & \text{if } x > \theta. \end{cases}$$

Hence, the pd of $X_{(n)}$ is given by $f_{X_{(n)}}(x) = n\frac{x^{n-1}}{\theta^n} I(0 \leq x \leq \theta)$. It follows that $E(X_{(n)}) = \frac{n}{n+1}\theta$ and $E\left(X_{(n)}^2\right) = \frac{n}{n+2}\theta^2$. Therefore, $\hat{\theta}_2$ is unbiased and its variance is $Var(\hat{\theta}_2) = \frac{\theta^2}{n(n+2)}$. Finally, the relative efficiency is

$$\frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} = \frac{3}{n+2}.$$

Indicating that for $n > 1$, $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$. □

Note that in the above example, the relative efficiency does not depend on θ (the parameter of interest), but this is not the case in general; more on this to follow in the next section.

In practice, when it is difficult or impossible to obtain an explicit/exact formula for MSE, we

use asymptotic expansions or simulations to estimate it. The R code below illustrates the second approach in the context of Example 2.6.

```
mse <- function(estimator, truth) mean((estimator - truth)^2)
bias <- function(estimator, truth) mean(estimator) - truth
var <- function(estimator) mean((estimator - mean(estimator))^2)

sim <- function(n, theta, estFun, N = 5000) {
  est <- replicate(N, {
    x <- runif(n, min = 0, max = theta)
    estFun(x) } )
  c(mse = mse(est, theta), bias = bias(est, theta), var = var(est))
}

restht1 <- replicate(500, sim(n = 10, theta = 3, estFun = \(x) 2 * mean(x)))
restht2 <- replicate(500,
                     sim(n <- 10, theta = 3, estFun = \(x) ((n + 1) / n) * max(x)))

restht1 |> rowMeans()
```

```
      mse      bias       var
 0.299688 -0.000347  0.299622
```

```
restht2 |> rowMeans()
```

```
      mse      bias       var
 7.51e-02 -8.49e-05  7.51e-02
```

These results closely match the theory which tells us that, for $n = 10$, the MSE of $\hat{\theta}_1$ is $3^2/(3 \times 10) = 0.3$ and that the MSE of $\hat{\theta}_2$ is $3^2/(10 \times (10+2)) = 0.075$.

2.4 The best unbiased estimator (MVUE)

It seems very natural to prefer estimators that have a small MSE. An estimator $\hat{\delta}$ whose MSE is uniformly larger than another estimator (i.e. $\exists \hat{\theta} : \ MSE_\theta(\hat{\theta}) \leq MSE_\theta(\hat{\delta})$, $\forall \theta \in \Theta$) is called ***inadmissible***. Actually, it would be best if we could find an estimator $\hat{\theta}$ that has the smallest MSE among all possible estimators for each possible value of θ. *Unfortunately, this is impossible*. Because in such a case $\hat{\theta}$ should satisfies $E_\theta\big(\hat{\theta} - \theta\big)^2 = 0$, $\forall \theta \in \Theta$ which is simply not possible.

Typically, two estimators $\hat{\theta}$ and $\hat{\delta}$ are not (uniformly) comparable. We can always find $\theta_1, \theta_2 \in \Theta$ such that

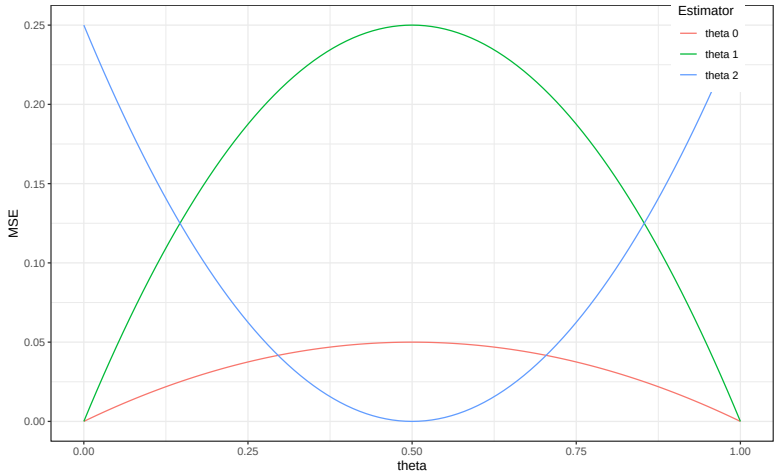
$$MSE_{\theta_1}(\hat{\theta}) < MSE_{\theta_1}(\hat{\delta}) \text{ and } MSE_{\theta_2}(\hat{\theta}) > MSE_{\theta_2}(\hat{\delta}).$$

Example 2.7.

Suppose an iid simple $X_i, \ldots, X_n$ from a Bernoulli distribution with an unknown parameter θ, $0 < \theta < 1$. Let $\hat{\theta}_0 = \overline{X}_n$, $\hat{\theta}_1 = X_1$ and $\hat{\theta}_2 = 1/2$. It is easy to see that

$$MSE(\hat{\theta}_0) = \frac{\theta(1-\theta)}{n}, \quad MSE(\hat{\theta}_1) = \theta(1-\theta), \quad MSE(\hat{\theta}_2) = (\theta - 1/2)^2$$

These three MSE (quadratic risk) functions, of θ, are plotted below (for $n = 5$).



$\hat{\theta}_1$ is inadmissible as it is less efficient than $\hat{\theta}_0$. The estimators $\hat{\theta}_0$ and $\hat{\theta}_2$ are not uniformly comparable. Near $\theta = 1/2$, $\hat{\theta}_2$ is the best, and away from $\theta = 1/2$, $\hat{\theta}_0$ is the best. □

In view of the fact that there is no uniformly minimum MSE (best) estimator, statisticians adopt other strategies. One of this strategies is to restrict attention to the *class of unbiased estimators*, and then to search for the best estimator in this restricted group.

Definition 2.2 (MVUE)**.** An *unbiased* estimator $\hat{\theta}$ of $\theta \in \Theta$ is the uniformly ***Minimum Variance Unbiased Estimator*** (MVUE) if for any other *unbiased* estimator $\hat{\delta}$

$$Var_\theta(\hat{\theta}) \leq Var_\theta(\hat{\delta}), \forall \theta \in \Theta.$$

In other words, the MVUE is the best (***most efficient***) *unbiased* estimator that can be found.

Facts to know

- MVUE may not exist (even an unbiased estimator may not exist!), but when it does, it's unique.
- In terms of MSE, the MVUE is not necessarily the best estimator as there may be *biased estimators* that achieve lower MSE than the MSE of the MVUE. In fact, sometimes a small increase in bias is associated with a large decrease in variance, overall decreasing the MSE.

The question now is how to find the MVUE (when it exists). To answer this question, different techniques exist in the literature. One of these techniques is based on a variance

inequality known as the ***Cramér-Rao bound***. Before presenting this method, we need to introduce the concept of ***Fisher information*** which plays a crucial role in statistical inference.

Chapter 3

Fisher information and Cramer-Rao bound

3.1 Score and Fisher information

Let X be a rv (or a rve) with pd $f(x;\theta)$ indexed by an unknown parameter $\theta \in \Theta \subset \mathbb{R}$. The question of interest here is the following: how much information can be obtained about θ as X get observed?

To answer this question, let's assume that f is differentiable with respect to θ, and define the ***score*** function associated with f to be

$$S(\theta, x) := \partial_\theta \log f(x;\theta) = \frac{\partial_\theta f(x;\theta)}{f(x;\theta)}.$$

Obverse that, by definition, for any given $\theta_0 \in \Theta$,

$$S(\theta_0, x) = \lim_{\epsilon \to 0} \frac{\frac{1}{\epsilon}\left[f(x, \theta_0 + \epsilon) - f(x;\theta_0)\right]}{f(x;\theta_0)}.$$

Thus, we can interpret this score as the relative (instantaneous) rate of change of $\theta \mapsto f(x;\theta)$ at the point θ_0. In particular, a function f that varies rapidly in the neighborhood of θ_0 should produce a large (absolute) score, and, in opposite, if f is flat the score should be small. In other words, a large value of $|S(\theta_0, x)|$ indicate that we can easily distinguish θ_0 from its neighboring values.

A remarkable property of the score function is given in the following proposition.

Proposition 3.1. *Suppose that*

(I) the support of f, i.e. the set $\{x : \ f(x;\theta) > 0\}$, does not depend on θ, and

(II) the operations of integration (or summation) and differentiation by θ can be interchanged in $\int f(x;\theta)dx$. Thus, $\partial_\theta \int f(x;\theta)dx = \int \partial_\theta f(x;\theta)dx$. (more about this condition can be found here)

Then

$$E_\theta S(\theta, X) = 0, \ \forall \theta.$$

Assumptions (I) and (II), given above, are both known to *hold for the exponential family*. And from now on, *unless otherwise stated*, we will always assume that (I) and (II) hold.

Taking the square (of S) and averaging we obtain $I_X(\theta)$:

$$I_X(\theta) := E_\theta \left[S^2(\theta, X) \right] = Var_\theta \left[S(\theta, X) \right]$$

which is known as the (expected) ***Fisher information*** (**FI**) that X contains about θ.

The Fisher information attempts to *quantify the average sensitivity of the random variable X to the value of the parameter θ*. If small changes in θ result in large changes in the values of X, then observing the latter can tell us a lot about θ. In this case the FI would be quite large. In other words, FI attempts to quantify how easy one can guess the θ that produced the observed X.

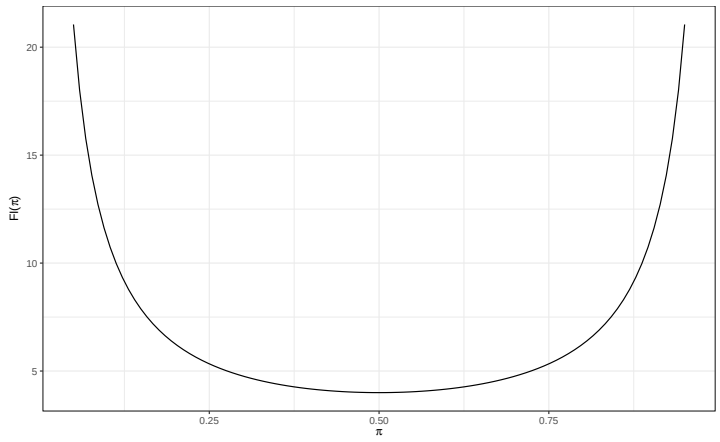
Remark. Note that the "X" in $I_X(\theta)$ is only a symbol used to indicate that the FI corresponds to the rv X. It does not mean that the FI it self is random! □

Example 3.1 (Calculating Fisher Information 1)**.**

- Bernoulli distribution: $X \sim Be(\pi)$ with $\pi \in (0, 1)$.

$$\begin{aligned} f(x; \pi) &= \pi^x (1-\pi)^{1-x}, \ x = 0, 1 \\ S(\pi, x) &= \partial_\pi \left\{ x \log(\pi) + (1-x) \log(1-\pi) \right\} = \frac{x - \pi}{\pi(1-\pi)} \\ I(\pi) &= E_\pi \left[S^2(\pi, X) \right] = \frac{1}{\pi(1-\pi)}. \end{aligned}$$

In this case, the FI is the reciprocal of the variance. This is not a unusual situation. Actually, as we will see latter, the FI is typically inversely proportional to the variance. The greater the variation (thus the smaller the FI), the more difficult it is to recover the parameter of interest.



- Normal distribution: $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ (the parameter of interest) and $\sigma^2 > 0$.

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$
$$S(\mu, x) = \partial_\mu \left\{ -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x-\mu)^2 \right\} = \frac{x-\mu}{\sigma^2}$$
$$I(\mu) = E\left[S^2(\mu, X)\right] = \frac{1}{\sigma^2}.$$

  In this case, the FI about μ does not depend on μ but only on σ^2. It decreases with σ^2. □

Proposition 3.2. *Assume thta $f(x;\theta)$ is twice differentiable, with respect to θ, and double integration (or summation) and differentiation under the integral sign can be interchanged, thus $\partial_\theta^2 \int f(x;\theta)dx = \int \partial_\theta^2 f(x;\theta)dx$. Then,*

$$I_X(\theta) = -E_\theta\left[\partial_\theta S(\theta, X)\right].$$

This is equivalent to say that $I_X(\theta) = -E_\theta\left[\partial_\theta^2 \log f(X;\theta)\right]$.

To see why this proposition is true, consider the following calculation where we have simplify the notation by writing S and f instead of $S(\theta, X)$ and $f(X;\theta)$, respectively.

$$\partial_\theta S = \partial_\theta \frac{\partial_\theta f}{f} = \frac{\partial_\theta^2 f \times f - (\partial_\theta f)^2}{f^2} = \frac{\partial_\theta^2 f}{f} - S^2$$

But $E_\theta\left(\frac{\partial_\theta^2 f(X;\theta)}{f(X;\theta)}\right) = \int \partial_\theta^2 f(x;\theta)dx = \partial_\theta^2 \int f(x;\theta)dx = 0$. Thus, $E_\theta(\partial_\theta S) = -E_\theta S^2$.

Example 3.2 (Calculating Fisher Information 2)**.**

- Bernoulli distribution: $X \sim Be(\pi)$ with $\pi \in (0,1)$.

$$I(\pi) = -E[\partial_\pi S(\pi, X)] = -\frac{1}{\pi^2(1-\pi)^2} E\left[-\pi(1-\pi) - (\pi(1-\pi))'(X-\pi)\right] = \frac{1}{\pi(1-\pi)}.$$

- Normal distribution: $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

$$I(\mu) = -E[\partial_\mu S(\mu, X)] = -E\left[\frac{-1}{\sigma^2}\right] = \frac{1}{\sigma^2}. \square$$

3.2 FI contained in a statistic

The above definitions of the score and FI can be directly applied to any statistics. In fact, let $T \equiv T(\boldsymbol{X}) \equiv T(X_1, \ldots, X_n)$ be a statistic whose pd is given by $h_n(t;\theta)$. The score associated with h_n and its corresponding FI are given by

$$S(\theta, t) := \partial_\theta \log h_n(t;\theta)$$
$$I_T(\theta) := E\left[S^2(\theta, T)\right] = Var\{S(\theta, T)\}$$

$I_T(\theta)$ is the (Fisher) information about θ that we can extract form T.
Assuming the interchangeability of integration and differentiation twice, this FI can also be expressed as

$$I_T(\theta) = -E\left[\partial_\theta S(\theta, T)\right] = -E\left[\partial^2_\theta \log h_n(T;\theta)\right].$$

Example 3.3.

- Let $X_i, i = 1, \ldots, n$, be an iid sample from a Bernoulli distribution $Be(\pi)$. Let define the statistic $T = \sum_{i=1}^n X_i$. Since $T \sim Bin(n, \pi)$, the pd of T is given by $h_n(t;\pi) = p_\pi(T = t) = C_n^t \pi^t (1-\pi)^{n-t}$. It follows that

$$I_T(\pi) = E\left[S^2(\theta, T)\right] = E\left[\frac{T - n\pi}{\pi(1-\pi)}\right]^2 = \frac{n}{\pi(1-\pi)}.$$

- Let $X_i, i = 1, \ldots, n$, be an iid sample from a Normal distribution $N(\mu, \sigma^2)$. Since $\overline{X}_n \sim N(\mu, \sigma^2/n)$, it follows that

$$I_{\overline{X}_n}(\mu) = \frac{n}{\sigma^2}. \square$$

An important fact about FI is its ***additivity***. Let T_1 and T_2 be two statistics with pd h_1 and h_2, and with FI $I_{T_1}(\theta)$ and $I_{T_2}(\theta)$, respectively. If T_1 and T_2 are *independent*, i.e. if $h(t_1, t_2; \theta) = h_1(t_1;\theta)h_2(t_2;\theta)$, $\forall t$, with h beging the joint pd of (T_1, T_2), then

$$\begin{aligned} I_{(T_1,T_2)}(\theta) &= E(\partial_\theta \log h(T_1, T_2;\theta))^2 = E(\partial_\theta \log h_1(T_1;\theta) + \partial_\theta \log h_2(T_2;\theta))^2 \\ &= E(\partial_\theta \log h_1(T_1;\theta))^2 + E(\partial_\theta \log h_2(T_2;\theta))^2 + 2E(\partial_\theta \log h_1(T_1;\theta)\partial_\theta \log h_2(T_2;\theta)) \\ &= I_{T_1}(\theta) + I_{T_2}(\theta). \end{aligned}$$

As consequence we have the following result.

Proposition 3.3. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample with joint pd $f_n(\boldsymbol{x};\theta) = \prod_i f(x_i, \theta)$, $I_{X_i}(\theta) = E(\partial_\theta \log f(X_i, \theta))^2$ be the FI contained in X_i about θ, and $I_{\boldsymbol{X}}(\theta) = E(\partial_\theta \log f_n(\boldsymbol{X}, \theta))^2$ be the FI contained in $\boldsymbol{X}$ about θ. Then, $I_{\boldsymbol{X}}(\theta) = nI_{X_i}(\theta)$.*

Attention. From now on, if no confusion is possible, we reserve the notation I_n for the FI contained in the entire sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, i.e. $I_n = I_{\boldsymbol{X}_n}$, and reserve the notation I for the FI contained in one sample unit, i.e. $I = I_{X_i} = I_{X_1}$. We can therefore write the equality above as $I_n(\theta) = nI(\theta)$.

For any (vector) statistic $T = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_d(\boldsymbol{X}))$, $d \geq 1$, it can be shown that

$$0 \leq I_T(\theta) \leq I_n(\theta).$$

Thus, the information on θ contained in any statistic T, derived from a sample $\boldsymbol{X}$, cannot exceed the information, on θ, contained in the sample $\boldsymbol{X}$ itself .

It may happen that $I_T(\theta) = I_n(\theta)$, in which case T *contains the same amount of information about* θ *as the whole sample*. Such a statistic is called a ***sufficient statistic***. Formally, the (original) definition of a sufficient statistic is the following.

Definition 3.1 (Sufficiency)**.** A statistic $T \equiv T(\boldsymbol{X})$ is sufficient for θ if the conditional distribution of $\boldsymbol{X}$ given T does not depend on θ.

Put another way, given a sufficient statistic T for θ, the sample $\boldsymbol{X}$ provides no additional information about θ. A sufficient statistic is particularly interesting if it is of *smaller dimension than the sample size*; i.e. $d << n$.

Example 3.4.

- From the examples 3.1 and 3.3, we learned that, for the Bernoulli distribution, $I_{\sum_{i=1}^n X_i}(\pi) = I_n(\pi)$. Thus, the one-dimensional statistic $\sum_{i=1}^n X_i$ contains as much information about π as the whole sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. So, $\sum_{i=1}^n X_i$ is sufficient for π.
- The same remark applies to the normal distribution where, from the examples 3.1 and 3.3, we can see that $I_{\overline{X}_n}(\mu) = I_n(\mu)$ and so $\overline{X}_n$ is sufficient for μ. □

A sufficient statistic is not unique. In fact, if $U = k(T)$, T is a sufficient statistic, and k is bijective (or at least one-to-one), then U is also sufficient. As an example, if we consider the case of the normal distribution, we can say that $\sum_{i=1}^n X_i$ is sufficient for μ since we known that $\overline{X}_n$ is.

The following result, known as the ***Factorization Theorem***, makes it very easy to identify sufficient statistics.

Theorem 3.1 (Factorization Theorem)**.** *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be a random sample with joint pd* $f_n(\boldsymbol{x}; \theta)$*. A statistic* $T(\boldsymbol{X})$ *is sufficient for* θ *if and only if* f_n *can be factorized as*

$$f_n(\boldsymbol{x}; \theta) = \phi(T(\boldsymbol{x}); \theta) h(\boldsymbol{x}), \ \forall \boldsymbol{x}, \theta.$$

where ϕ *is a function that depends on* $\boldsymbol{x}$ *only through* $T(\boldsymbol{x})$ *and* h *is a function that does not depend on* θ*.*

As a consequence of this theorem, we have the following result that allows one to obtain sufficient statistics when data come from an exponential family.

Proposition 3.4. *If* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *is an iid sample from a one-parameter exponential family with pd*

$$h(x) \exp\left(\eta(\theta) T(x) - B(\theta)\right),$$

then the statistic $\sum_{i=1}^n T(X_i)$ *is sufficient for* θ*.*

Example 3.5.

- In the case of $N(\mu, \sigma^2)$, with a known σ^2, based on the above proposition and what we

have learned previously about this distribution, we can conclude that $\sum_i X_i$ is sufficient for μ.

- Let $X_1, \ldots, X_n$ be an iid sample from $f(x;\theta) = \theta x^{\theta-1}$, where $x \in (0,1)$ and $\theta > 0$. Let's show that $\prod_{i=1}^n X_i$ is sufficient for θ. To see this, we can write $f(x;\theta)$ as

$$f(x;\theta) = I(0 < x < 1)x^{-1}\exp(\theta \log x + \log \theta),$$

which is an exponential family with natural statistic $T = \log X$. It follows that $L = \sum_i \log(X_i)$ is sufficient for θ. And since $\prod_{i=1}^n X_i = \exp(L)$ is a bijective function of L, it is also sufficient for θ. □

3.3 FI and re-parametrization

We have already seen that statistical models can be parameterized in different ways. It is important to realize that FI depends on the chosen parameterization.

Proposition 3.5 (FI re-parametrization)**.** *Let $\eta : \theta \mapsto \eta(\theta)$ be a real differentiable function of θ. Denote by $f^*(x;\eta)$ the pd of X parameterized with η so that $f^*(x;\eta) = f(x;\theta)$, $\forall x$. Let $I(\theta) = E[\partial_\theta \log f(X;\theta)]^2$ be the FI about θ (when the parameterization in θ is used) and $I(\eta) = E\left[\partial_\eta \log f^*(X;\eta)\right]^2$ be the FI about η (when the parameterization in η is used). Then $I(\theta) = I(\eta) \times (\partial_\theta \eta(\theta))^2$.*

The proof of this proposition is straightforward and is a direct consequence of the chain rule :

$$I(\theta) = E[\partial_\theta \log f(X;\theta)]^2 = E[\partial_\theta \log f^*(X;\eta)]^2 = E\left[\partial_\theta \eta(\theta)\partial_\eta \log f^*(X;\eta)\right]^2 = (\eta'(\theta))^2 I(\eta).$$

In the following, for a given model $\{f(x;\theta);\ \theta \in \Theta\}$, every time we write $I(\eta(\theta))$, we will be referring to the FI with the reparameterization $\theta \mapsto \eta(\theta)$. □

Example 3.6.

- Let $X \sim N(\mu, \sigma^2)$. $f(x;\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

$$I(\sigma) = -E\{\partial_\sigma^2 \log f\} = -E\left\{\frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4}\right\} = \frac{2}{\sigma^2}.$$

And, with the (re)parametrization $\theta = \sigma^2$,

$$I(\sigma^2) = -E\{\partial_\theta^2 \log f\} = -E\left\{\frac{1}{2\theta^2} - \frac{(X-\mu)^2}{\theta^3}\right\} = \frac{1}{2\sigma^4}$$

We observe that

$$I(\sigma^2) = \frac{1}{4\sigma^2} I(\sigma).$$

This last result can be obtained directly by applying Proposition 3.5. Indeed, we have that

$$I(\sigma^2) = I(\sigma) \times \left(\partial_t \sqrt{t} \Big|_{t=\sigma^2} \right)^2 = I(\sigma) \left(\frac{1}{2\sqrt{\sigma^2}} \right)^2.$$

Or, equivalently,

$$I(\sigma) = I(\sigma^2) \times \left(\partial_t \, t^2 \Big|_{t=\sigma} \right)^2 = I(\sigma^2) \times (2\sigma)^2.$$

- Let $X \sim Pois(\theta)$. $f(x;\theta) = \frac{\theta^x}{x!} e^{-\theta}$, $x = 0, 1, \ldots$ and $\theta > 0$. Direct calculation leads to

$$I(\theta) = -E\{\partial_\theta^2 \log f\} = -E\left\{ -\frac{X}{\theta^2} \right\} = \frac{1}{\theta}$$

Now, let's consider the parametrization with $\eta = \log(\theta)$: $f^*(x;\eta) = \frac{e^{x\eta}}{x!} e^{-e^\eta}$, $\eta \in (-\infty, \infty)$.

$$I(\eta) = -E\left\{ \partial_\eta^2 \log f^* \right\} = -E\{-e^\eta\} = e^\eta$$

Thus, $I(\log(\theta)) = \theta$. This same result can be obtained by directly applying Proposition 3.5 as follows:

$$I(\log(\theta)) = \frac{I(\theta)}{((\log(\theta))')^2} = \frac{1/\theta}{1/\theta^2} = \theta. \square$$

This last example suggests that, in Poisson model, it is easier to estimate $\eta = \log(\theta)$ than θ when the latter is large. Let's check this out. A natural estimator of θ is $\hat{\theta} := \overline{X}$ and a natural estimator of η is $\hat{\eta} := \log(\hat{\theta}) = \log(\overline{X})$. We have that $MSE(\hat{\theta}) = \frac{\theta}{n}$ and, by first order Taylor polynomial approximation, i.e. $\log(\hat{\theta}) \approx \log(\theta) + (\hat{\theta} - \theta)\frac{1}{\theta}$, we can write $MSE(\hat{\eta}) \approx \frac{1}{n\theta}$. These mean-square errors cannot be compared, as they are of different scales/magnitudes. However, we can see that as θ increases, the MSE performance of $\hat{\theta}$ becomes worse and worse compared to that of $\hat{\eta}$. The following simulation confirms this fact.

```
th <- 10 # 20, 40, 80, 100, ...
eta <- log(th)

hat.th <- replicate(5000, rpois(100, th) |> mean())
hat.eta <- log(hat.th)

# relative MSE
mse(hat.th, th) / mse(hat.eta, eta)
```

```
[1] 100
```

```r
# relative "mean absolute relative error"
mean(abs(hat.th / th - 1)) / mean(abs(hat.eta / eta - 1))
```

```
[1] 2.3
```

Remark. We have seen that, in general, $I(g(\theta)) \neq I(\theta)$. We may ask the question what happen with FI when the data itself, or a statistic from it, are transformed. The answer depends on the type of transformation used. For example in the case of strictly monotonic and differentiable transformation of the data, FI does change. More precisely, if $U = k(T)$, where k is a differentiable and strictly monotonic function that does not depend on θ, then $I_T(\theta) = I_T(\theta)$. This is a direct consequence of the *Change of Variable(s) Formula*: $f_T(T;\theta) = f_U(U;\theta)\,|k'(U)|$. □

3.4 Information Inequality: The Cramer-Rao Lower bound (CRLB)

We will develop a lower bound for the variance of any given statistic, which can be mainly used (i) as a benchmark for comparing estimator performance, and (ii) to find the MVUE (Minimum Variance Unbiased Estimator). The bound we are interested in is called the Cramer-Rao Lower Bound (**CRLB**), and is given in the following theorem.

Theorem 3.2 (Information Inequality)**.** *Let $\boldsymbol{X} \equiv \boldsymbol{X}_n = (X_1, \ldots, X_n)$ be an iid sample with joint pd $f_n(\boldsymbol{x};\theta)$, $\theta \in \Theta$. Assume that assumptions (I) and (II) as given above (see Proposition 3.1) hold for $f_n(\boldsymbol{x};\theta)$, the joint dp of $\boldsymbol{X}$. Let $T \equiv T(\boldsymbol{X})$ be a statistic. Assume that (III) $\partial_\theta E_\theta(T)$ exists and can be obtained by differentiating under the integral sign. i.e., $\partial_\theta \int T(\boldsymbol{x}) f_n(\boldsymbol{x};\theta) d\boldsymbol{x} = \int T(\boldsymbol{x}) \partial_\theta f_n(\boldsymbol{x};\theta) d\boldsymbol{x}$. Then*

$$Var_\theta(T) \geq \frac{(\partial_\theta E_\theta(T))^2}{I_n(\theta)}, \forall \theta \in \Theta.$$

The inequality above is a direct consequence of the Cauchy-Schwarz inequality. The proof goes as follows. Let $S_n \equiv S(\theta;\boldsymbol{X}) = \partial_\theta \log f_n(\boldsymbol{X};\theta)$ be the Score associated with f_n. By the Cauchy-Schwarz inequality,

$$[Cov(T, S_n)]^2 = [E(TS_n) - E(T)E(S_n)]^2 = [E(TS_n)]^2 \leq Var(T)Var(S_n) = Var(T)I_n(\theta).$$

The final result is the consequence of the fact that $E(TS_n) = \int T(\boldsymbol{x}) \partial_\theta \log f_n(\boldsymbol{X};\theta) f_n(\boldsymbol{x};\theta) d\boldsymbol{x} = \int T(\boldsymbol{x}) \partial_\theta f_n(\boldsymbol{x};\theta) d\boldsymbol{x} = \partial_\theta E(T)$.

As a corollary of the theorem above, we can say that, if $\delta \equiv \delta(\boldsymbol{X})$is an *unbiased* estimator of θ, then $Var_\theta(\delta) \geq \frac{1}{I_n(\theta)}$. Or, more generally, if δ is an unbiased estimator of $g(\theta)$, then

$$Var_\theta(\delta) \geq \frac{\left(g'(\theta)\right)^2}{I_n(\theta)} = \frac{1}{I_n(g(\theta))}.$$

The right hand side of this inequality is called the ***CRLB*** *for* $g(\theta)$. This bound is the

minimum possible variance that any unbiased estimator of $g(\theta)$ can achieve. To put it more precisely, in a *parametric model in which Assumption (III) above is fulfilled for any statistic T (including $T = 1$)*, the variance of any unbiased estimator of $g(\theta)$ is at least equal to the $CRLB(g(\theta)) := I_n^{-1}(g(\theta))$. Note that Assumption (III) *holds for the exponential family for any statistic T*.

An unbiased estimator δ of $g(\theta)$ is called ***efficient*** if its variance equals the CRLB for $g(\theta)$, i.e. if $Var(\delta(\boldsymbol{X})) = I_n^{-1}(g(\theta))$; otherwise its (absolute) efficiency is defined to be

$$Eff(\delta) = \frac{\left(g^{'}(\theta)\right)^2}{I_n(\theta)Var(\delta)}.$$

$Eff(\delta) \in [0,1]$, and equals 1 if and only if δ is efficient.

By definition, an efficient estimator is (1) unbiased and (2) its variance is uniformly lower than (or equal to) the variance of any other unbiased estimator. Thus, *an efficient estimator, when it exists, is the uniformly minimum variance unbiased estimator (MVUE)*. Note that *efficiency is a stronger requirement than being MVUE*. In fact, there are many cases where the CRLB is not attainable and where a MVUE exists.

Example 3.7. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $N(\mu, \sigma^2)$ (exponential family).

- Suppose that μ is our parameter of interest. We have seen that $I_n(\mu) = n/\sigma^2$. So for any unbiased estimator $\hat{\mu}_n$ of μ,

  $$Var(\hat{\mu}_n) \geq \frac{\sigma^2}{n}.$$

  Now, since $\overline{X}_n$ is an unbiased estimator of μ and $Var(\overline{X}_n) = \frac{\sigma^2}{n}$, we conclude that $\overline{X}_n$ is efficient for μ. And so, $\overline{X}_n$ is the MVUE of μ.

- Suppose that σ^2 is our parameter of interest and μ is known. We have seen that $I(\sigma^2) = \frac{1}{2\sigma^4}$. So for any unbiased estimator $\hat{\sigma}_n^2$ of σ^2,

  $$Var(\hat{\sigma}_n^2) \geq 2\sigma^4/n.$$

  On the other hand, we known that $\tilde{\sigma}_n^2 = n^{-1}\sum_i (X_i - \mu)^2$ is an unbiased estimator of σ^2. And, using the fact that $E(X-\mu)^4 = 3\sigma^4$, we have that $Var\left(\tilde{\sigma}_n^2\right) = n^{-1}Var(X-\mu)^2 = n^{-1}\left(E(X-\mu)^4 - \sigma^4\right) = 2\sigma^4/n$. So, we conclude that $\tilde{\sigma}_n^2$ is efficient for σ^2. And so it is the MVUE of σ^2.

- Suppose that μ^2 is our parameter of interest. By the information inequality, with $g : \mu \mapsto \mu^2$, we have that

  $$Var(\hat{\delta}) \geq \frac{(2\mu)^2}{n/\sigma^2} = \frac{4\mu^2\sigma^2}{n}.$$

  for any unbiased estimator $\hat{\delta}$ of μ^2. But, for now, we cannot say if this limit is attainable

or not and thus if there is an efficient estimator for μ^2 or not. □

Example 3.8 (Importance of assumptions)**.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $Unif(0, \theta)$, $\theta > 0$. Thus $f(x;\theta) = \frac{1}{\theta}$, $0 < x < \theta$. Since $\partial_\theta \log f(x;\theta) = -1/\theta$, if we apply Theorem 3.2, we could conclude that, for any unbiased estimator $\hat{\theta}$ of θ,

$$Var(\hat{\theta}) \geq \frac{\theta^2}{n}.$$

But, we learned earlier that the estimator $\hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$ is unbiased and its variance is $\frac{\theta^2}{n(n+2)}$, which is uniformly smaller than θ^2/n ! The problem here is that the support of f depends on θ, which means that the required Assumption (III) is not fulfilled. □

3.5 Efficiency in exponential families

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from a one-parameter exponential family with pd (in canonical form, with η begin the canonical parameter) : $f(x;\eta) = h(x) \exp(\eta T(x) - A(\eta))$. For $T = T(X_1)$, we have seen that $E(T) = A^{'}(\eta)$ and $Var(T) = A^{''}(\eta)$. The FI contained in T about η is

$$I(\eta) = Var\big(\partial_\eta \log f\big) = Var(T) = A^{''}(\eta).$$

Note that this result can also be obtained from

$$I(\eta) = -E\Big(\partial_\eta^2 \log f\Big) = -E\Big(-A^{''}(\eta)\Big) = A^{''}(\eta).$$

Now, let $\overline{T} = n^{-} \sum_i T(X_i)$. We have that, $E\big(\overline{T}\big) = E(T) = A^{'}(\eta)$, and

$$Var\big(\overline{T}\big) = \frac{Var(T)}{n} = \frac{A^{''}(\eta)}{n} = \frac{\Big(A^{''}(\eta)\Big)^2}{I_n(\eta)}.$$

This demonstrates that $\overline{T}$ is efficient for $A^{'}(\eta)$.

Let's now consider the original parameterization with θ : $f(x;\theta) = h(x) \exp(\eta(\theta) T(x) - B(\theta))$. Remember that, in terms of θ, we can express $E(T)$ and $Var(T)$ as $E(T) = B^{'}(\theta)/\eta^{'}(\theta)$ and $Var(T) = \partial_\theta E(T)/\eta^{'}(\theta)$. The FI contained in T about θ is

$$I(\theta) = Var(\partial_\theta \log f) = Var\Big(\eta^{'}(\theta) T\Big) = \eta^{'}(\theta) \partial_\theta E(T).$$

Again, we have that $E\big(\overline{T}\big) = E(T)$, and

$$Var(\overline{T}) = \frac{Var(T)}{n} = \frac{\partial_\theta E(T)}{n\eta^{'}(\theta)} = \frac{(\partial_\theta E(T))^2}{I_n(\theta)}.$$

So, $\overline{T}$ is efficient for $B^{'}(\theta)/\eta^{'}(\theta)$.

Thus, we can conclude that, *in an exponential family, the sample mean of the natural statistic is always efficient for the expected value of the latter*.

Example 3.9. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from the exponential distribution. Thus, for some $\theta > 0$, the pd of X_i is given by

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} e^{-\frac{x}{\theta}} I(x > 0) \\ &= I(x > 0) \exp\left(-\frac{1}{\theta} x - \log(\theta) \right) \qquad \text{(exponential family)} \end{aligned}$$

Since $E(X_1) = \theta$, $n^{-1} \sum_{i=1}^{n} X_i$ is efficient for θ. □

3.6 CRLB Attainment

A natural question to ask is under what conditions a given unbiased estimator, say $T(\boldsymbol{X})$, of $g(\theta)$ can attain the CRLB? It turns out that the CRLB is achieved only when the definition of the estimator $T(\boldsymbol{X})$ has the special form given in the following theorem.

Theorem 3.3 (CRLB Attainment)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample with a joint pd $f_n(\boldsymbol{x};\theta), \theta \in \Theta$. Suppose that Assumption (III), as given above, holds for any statistic. Then, there exists an efficient estimator $T(\boldsymbol{X})$ for $g(\theta)$ if and only if*

$$\partial_\theta \log f_n(\boldsymbol{x};\theta) = a_n(\theta)[T(\boldsymbol{x}) - g(\theta)], \ \forall \theta \in \Theta, \tag{3.1}$$

for some function $a_n(\theta) \neq 0$, $\forall \theta \in \Theta$. In addition, if $f_n(\boldsymbol{x};\theta)$ satisfies (3.1), then (i) $a_n(\theta)$ equals $I_n(\theta)/g'(\theta)$, and (ii) $f_n(\boldsymbol{x};\theta)$ is a one-parameter exponential family.

Let's proof this result. Let $S_n = \partial_\theta \log f_n(\boldsymbol{X};\theta)$ be the Score associated with f_n. Remember that $E(S_n) = 0$, $Var(S_n) = I_n(\theta)$ and, for any statistic T, $Cov(S_n, T) = \partial_\theta E(T)$. Suppose that $T(\boldsymbol{X})$ is an efficient estimator of $g(\theta)$. Then, by definition, (i) $E(T) = g(\theta)$, and (ii) $Var(T) = (g'(\theta))^2 / I_n(\theta)$, $\forall \theta \in \Theta$. This implies that $Var(T)Var(S_n) = Cov^2(S_n, T)$. Since Cauchy-Schwarz inequality become an equality only in the case of linear dependence, we conclude that $\exists a \equiv a_n(\theta) \neq 0$ and $b \equiv b_n(\theta)$, such that $S_n = aT + b$. But since $E(S_n) = 0$, we have that $b = -ag(\theta)$. Thus, $S_n = a(T - g(\theta))$. Conversely, suppose that, $\forall \theta \in \Theta$, $\exists a \equiv a_n(\theta) \neq 0$ such that $S_n = a(T - g(\theta))$. The fact that $E(S_n) = 0$ and $Var(S_n) = I_n(\theta)$ implies that $E(T) = g(\theta)$ and $I_n(\theta) = a^2 Var(T)$, receptively. And $Cov(S_n, T) = \partial_\theta E(T)$ implies that $\partial_\theta E(T) = aVar(T)$, which in turn implies that $a = I_n(\theta)/g'(\theta)$. Thus, $Var(T) = (g'(\theta))^2 / I_n(\theta)$, which concludes the proof.

The CRLB attainment theorem leads to an explicit constructive procedure for deriving the (efficient) MVUE of $g(\theta)$ *when it exists*. Namely, put

$$T = g(\theta) + \frac{g'(\theta)}{I_n(\theta)} \partial_\theta \log f_n(\boldsymbol{X};\theta).$$

If the expression on the right hand side of the equality above *does not depend on* θ, i.e. T as defined above is a statistic, then T is efficient for $g(\theta)$, and so it is also the MVUE.

Example 3.10. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from the exponential distribution. Thus, for some $\theta > 0$, the pd of X_i is given by $f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, for $x > 0$.

- Let's try to find the efficient estimator for θ. To do so, put

$$T = \theta + \partial_\theta \log f_n(\boldsymbol{X};\theta)/I_n(\theta).$$

We have that $\partial_\theta \log f_n(\boldsymbol{X};\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n X_i$, and $\partial_\theta^2 \log f_n(\boldsymbol{X};\theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^n X_i$. So, $I_n(\theta) = \frac{n}{\theta^2}$, and thus

$$T = \theta + \frac{\theta^2}{n}\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n X_i\right) = n^{-1}\sum_{i=1}^n X_i.$$

$\to n^{-1}\sum_i X_i$ is the desired estimator.

- Let's now try to find an efficient estimator for $\delta = \frac{1}{\theta}$. Following the same procedure, let

$$\begin{aligned} T &= \frac{1}{\theta} + \frac{\left(\frac{1}{\theta}\right)'}{\frac{n}{\theta^2}}\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n X_i\right) \\ &= \frac{2}{\theta} - \frac{1}{\theta^2} n^{-1}\sum_{i=1}^n X_i. \end{aligned}$$

The latter is not a statistic, so there is no efficient estimator for $1/\theta$. Note, however, that this does not mean that there is no MVUE. □

3.7 Multiparameter case

The above theory can be extended to the case of parametric models with a pd $f(x;\boldsymbol{\theta})$ that depends on several parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. We assume in the sequel that the following two regularity conditions are satisfied :

(I) The set $\{x : \ f(x;\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$, and

(II) The operations of integration (or summation) and differentiation by θ_j can be interchanged in $\int f(x;\boldsymbol{\theta})dx$. i.e., $\partial_{\theta_j}\int f(x;\boldsymbol{\theta})dx = \int \partial_{\theta_j} f(x;\boldsymbol{\theta})dx$, $\forall j = 1, \ldots, d$.

- The ***Score vector*** of f is defined as the gradient of $\boldsymbol{\theta} \mapsto \log f(X;\boldsymbol{\theta})$, i.e.

$$\boldsymbol{S} := \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta}) = (S_1, \ldots, S_d)^t,$$

where $S_j = \partial_{\theta_j} \log f(X;\boldsymbol{\theta})$ is the score for θ_j. It is easy to see that $E(S_j) = 0$, $\forall j$. Thus, $E(\boldsymbol{S}) = \boldsymbol{0}$.

- The ***FI matrix*** contained in X about $\boldsymbol{\theta}$ is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}) := E(\boldsymbol{S}\boldsymbol{S}^t) = Var\,(\boldsymbol{S})\,.$$

  Thus, $\boldsymbol{I}(\boldsymbol{\theta}) = \left[I_{jk}(\boldsymbol{\theta})\right]_{1\le j,k\le d}$, where $I_{jk}(\boldsymbol{\theta}) := E\left(S_j \times S_k\right) = Cov\left(S_j, S_k\right)$.

- Any FI matrix is symmetric and positive semidefinite. Actually, if the considered model is irredundant (i.e. not overparameterized), then its FI matrix is positive definite and therefore invertible. From now on, we'll always assume that the FI matrices we are considering are invertible.

- If $f(x;\boldsymbol{\theta})$ is twice differentiable and double integration and differentiation under the integral sign can be interchanged, i.e., $\partial_{\theta_k\theta_j}\int f(x;\boldsymbol{\theta})dx = \int \partial_{\theta_k\theta_j} f(x;\boldsymbol{\theta})dx$, $\forall j,k = 1,\ldots,d$, then $I_{jk}(\boldsymbol{\theta}) = -E\left(\partial_{\theta_k} S_j\right)$. Thus,

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\left(\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta})\right),$$

  where $\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta})$ denotes the Hessian of $\boldsymbol{\theta} \mapsto \log f(X;\boldsymbol{\theta})$.

- To be more explicit about the formulas given above. Let's consider the special case of a two-parameter model with $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $S_1 = \partial_{\theta_1} \log f(X;\boldsymbol{\theta})$, and $S_2 = \partial_{\theta_2} \log f(X;\boldsymbol{\theta})$. Under the regularity assumptions stated above, we can write the FI matrix in any of the following equivalent expressions:

$$\boldsymbol{I}(\theta_1,\theta_2) = E\begin{pmatrix} S_1^2 & S_1S_2 \\ S_1S_2 & S_2^2 \end{pmatrix} = \begin{pmatrix} Var(S_1) & Cov(S_1,S_2) \\ Cov(S_1,S_2) & Var(S_2) \end{pmatrix} = -E\begin{pmatrix} \partial_{\theta_1}S_1 & \partial_{\theta_2}S_1 \\ \partial_{\theta_1}S_2 & \partial_{\theta_2}S_2 \end{pmatrix}.$$

- Let $\boldsymbol{X} = (X_1,\ldots,X_n)$ be an iid sample with joint density $f_n(\boldsymbol{x};\boldsymbol{\theta})$. The Score vector of f_n is defined as $\boldsymbol{S}_n = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f_n(\boldsymbol{X};\boldsymbol{\theta})$, and the FI matrix contained in $\boldsymbol{X}$ about $\boldsymbol{\theta}$ is given by

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = E\left(\boldsymbol{S}_n\boldsymbol{S}_n^t\right).$$

  It can be shown that $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$.

Example 3.11. Normal distribution $N(\mu,\sigma^2)$ with pd $f(x;\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

$$S_1 := \partial_\mu \log f = \frac{X-\mu}{\sigma^2} \quad \text{and} \quad S_2 := \partial_{\sigma^2}\log f = -\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4}, \text{ so}$$

$$I_{11} := -E(\partial_\mu S_1) = -E\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}\ , \quad I_{12} := -E(\partial_{\sigma^2}S_1) = -E\left(-\frac{X-\mu}{\sigma^4}\right) = 0,$$

$$I_{21} := -E(\partial_\mu S_2) = 0 \text{ , and } \quad I_{22} := -E(\partial_{\sigma^2}S_2) = -E\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right) = \frac{1}{2\sigma^4}.$$

$$\Rightarrow \boldsymbol{I}(\mu,\sigma^2) := \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \text{ and } \boldsymbol{I}_n(\mu,\sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}. \square$$

Proposition 3.6 (FI reparametrization – Multiparameter case)**.** *If we rewrite our model in terms of some other parameter* $\boldsymbol{\eta} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta}))$*, and denote by* $\boldsymbol{I}(\boldsymbol{\eta}) \equiv \boldsymbol{I}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ *the FI matrix of the new model/parametrization, then the FI for the two parametrizations (original and new) are related by*

$$\boldsymbol{I}(\boldsymbol{\theta}) = \dot{\boldsymbol{\eta}}^t(\boldsymbol{\theta}) \boldsymbol{I}(\boldsymbol{\eta}) \dot{\boldsymbol{\eta}}(\boldsymbol{\theta}), \tag{3.2}$$

where $\dot{\boldsymbol{\eta}}$ *denotes the the* $d \times d$ *Jacobian matrix of* $\boldsymbol{\eta}$*, whose* $(j,k)-$*th element is* $\partial_{\theta_k} \eta_j(\boldsymbol{\theta})$*.*

Example 3.12. Normal distribution $N(\mu, \sigma^2)$.

$$\boldsymbol{I}(\mu, \sigma) = \dot{\boldsymbol{\eta}}^t(\mu, \sigma) \boldsymbol{I}(\mu, \sigma^2) \dot{\boldsymbol{\eta}}(\mu, \sigma),$$

where $\dot{\boldsymbol{\eta}} := (\eta_1, \eta_2)$, with $\eta_1 : (\mu, \sigma) \mapsto \mu$ and $\eta_2 : (\mu, \sigma) \mapsto \sigma^2$.

$$\dot{\boldsymbol{\eta}} = \begin{pmatrix} \partial_\mu \eta_1 & \partial_\sigma \eta_1 \\ \partial_\mu \eta_2 & \partial_\sigma \eta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix}.$$

$$\Rightarrow \boldsymbol{I}(\mu, \sigma) = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^3} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \square$$

Theorem 3.4 (Information Inequality – Multiparameter CRLB)**.** *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be an iid sample with joint pd* $f_n(\boldsymbol{x}; \boldsymbol{\theta})$*. Assume that assumptions (I) and (II) as given above (see the beginning of the current section) hold for* $f_n(\boldsymbol{x}; \boldsymbol{\theta})$*. Let* $\boldsymbol{T} \equiv (T_1(\boldsymbol{X}), \ldots, T_p(\boldsymbol{X}))$ *be a statistic. Asuume that (III)* $\partial_{\theta_k} E_{\boldsymbol{\theta}}(T_j)$ *exists and can be obtained by differentiating under the integral sign. If* $\boldsymbol{T}$ *is an unbiased estimator of* $\boldsymbol{g}(\boldsymbol{\theta}) = \big(g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta})\big)$*, i.e. if* $E_{\boldsymbol{\theta}}(T_j) = g_j(\boldsymbol{\theta})$*,* $\forall j, \boldsymbol{\theta}$*, then*

$$Var_{\boldsymbol{\theta}}(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta})),$$

where $\boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta})) = \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) \dot{\boldsymbol{g}}^t(\boldsymbol{\theta})$*,* $\boldsymbol{I}_n^{-1}$ *is the inverse matrix of* $\boldsymbol{I}_n$*, and* $\dot{\boldsymbol{g}}$ *is the* $p \times d$ *Jacobian matrix of* $\boldsymbol{g}$*, whose* $(j,k)-$*th element is* $\partial_{\theta_k} g_j(\boldsymbol{\theta})$*. In particular, if* $\boldsymbol{T}$ *is unbiased for* $\boldsymbol{\theta}$*, then* $Var_{\boldsymbol{\theta}}(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$*.*

Above, the notation $\boldsymbol{A} \succeq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite matrix. Consequently, writing $Var(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}$ is equivalent to say that $Var(\boldsymbol{a}^t \boldsymbol{T}) \geq \boldsymbol{a}^t \boldsymbol{I}_n^{-1} \boldsymbol{a}$, $\forall \boldsymbol{a}$.

To see an interesting implication of this result, let's consider the two-parameters case ($d = 2$), with $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Let's write down the FI matrix and its inverse as follows

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = n \begin{pmatrix} Var(S_1) & Cov(S_1, S_2) \\ Cov(S_2, S_1) & Var(S_2) \end{pmatrix} \equiv n \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}.$$

$$\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) = \frac{1}{n(I_{11} I_{22} - I_{12}^2)} \begin{pmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{pmatrix}.$$

Assume that θ_1 is our *primary parameter of interest* and θ_2 is a *nuisance parameter*. In the normal model, for example, we might be interested only in the mean μ, while σ^2 is a a "nuisance", which is there only to make the model correct. In such a context, we can think of two situations:

- θ_2 *is known*: in this case θ_1 is the only (unknown) parameter of our model, and the (univariate) information inequality (see Theorem 3.2) tells us that, for any unbiased estimator T_1 of θ_1,
$$Var(T_1) \geq \frac{1}{nI_{11}}.$$
- θ_2 *is unknown*: in this case, by applying Theorem 3.4 with $g(\theta_1, \theta_2) = \theta_1$, the information inequality tells us that, for any unbiased estimator T_1 of θ_1,
$$Var(T_1) \geq (1,0) I_n^{-1}(\boldsymbol{\theta})(1,0)^t = \frac{1}{nI_{11}^*},$$
where $I_{11}^* = \dfrac{I_{11}I_{22} - I_{12}^2}{I_{22}} = I_{11}\left(1 - Cor^2(S_1, S_2)\right)$.

Observe that $I_{11}^* \leq I_{11}$, with equality if and only if S_1 and S_2 are uncorrelated (i.e. *diagonal FI matrix*).
$\rightarrow$ *nuisance parameter(s) cause loss of information*, resulting in an increase in the lower bound of the variance (of other model parameters).

Example 3.13. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown. We have seen that
$$\boldsymbol{I}_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$
Notice that this is a *diagonal* matrix.

- Suppose that μ is our parameter of interest. The information inequality tells us that $Var(\hat{\mu}) \geq \frac{\sigma^2}{n}$, for any unbiased estimator $\hat{\mu}$ of μ, *whether* σ^2 *is known or not*. Since $E(\overline{X}_n) = \mu$ and $Var(\overline{X}_n) = \sigma^2/n$, $\overline{X}_n$ is efficient for μ, whether σ^2 is known or not.
- Suppose that σ^2 is our parameter of interest. The information inequality tells us that $Var(\hat{\sigma}^2) \geq \frac{2\sigma^4}{n}$, for any unbiased estimator $\hat{\sigma}^2$ of σ^2, *whether* μ *is known or not*. Let $\tilde{\sigma}_n^2 = n^{-1}\sum_i (X_i - \mu)^2$. We have seen that $E(\tilde{\sigma}_n^2) = \sigma^2$ and $Var(\tilde{\sigma}_n^2) = 2\sigma^4/n$, so $\tilde{\sigma}_n^2$ is efficient for σ^2 *when* μ *is known*. When μ is unknown, $\tilde{\sigma}_n^2$ is not an estimator. In this case, it is "natural" to estimate σ^2 by $S_n^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$. This latter is unbiased but (marginally) not efficient since $Var(S_n^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$. Note that it can be shown that S_n^2 is the MVUE of σ^2.
- Suppose that $\lambda = \sigma/\mu$ is our parameter of interest. The information inequality, with $g(\mu, \sigma^2) = \sigma/\mu$, tells us that
$$Var(\hat{\lambda}) \geq \left(\frac{-\sigma}{\mu^2}, \frac{1}{2\mu\sigma}\right) \boldsymbol{I}_n^{-1}(\mu, \sigma^2) \left(\frac{-\sigma}{\mu^2}, \frac{1}{2\mu\sigma}\right)^t = \frac{\sigma^4}{n\mu^4}\left(1 + \frac{\mu^2}{2\sigma^2}\right),$$

for any unbiased estimator $\hat{\lambda}$ of σ/μ. □

Example 3.14. Let $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, be iid rve from the trinomial distribution $Mult(m, (\pi_1, \pi_2, \pi_3))$, with joint pd

$$f(\boldsymbol{x}; \boldsymbol{\pi}) = \frac{m!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

where $m \geq 1$, $\boldsymbol{x} = (x_1, x_2)$, and $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is unknown; with $x_1 = 0, \ldots, m$, $x_2 = 0, \ldots, m$, $x_3 = m - x_1 - x_2$, $\pi_1 \in (0,1)$, $\pi_2 \in (0,1)$, and $\pi_3 = 1 - \pi_1 - \pi_2$. We have that

$$S_1 := \partial_{\pi_1} \log f(\boldsymbol{X}; \boldsymbol{\pi}) = \frac{X_1}{\pi_1} - \frac{X_3}{\pi_3} \text{ and } S_2 := \partial_{\pi_2} \log f(\boldsymbol{X}; \boldsymbol{\pi}) = \frac{X_2}{\pi_2} - \frac{X_3}{\pi_3}$$

The score vector is given by $\boldsymbol{S} = (S_1, S_2)^t$, and, knowing that $E(X_k) = m\pi_k$, $k = 1, 2, 3$, the FI matrix is given by

$$\boldsymbol{I}(\pi_1, \pi_2) = -E \begin{pmatrix} \partial_{\pi_1} S_1 & \partial_{\pi_2} S_1 \\ \partial_{\pi_1} S_2 & \partial_{\pi_2} S_2 \end{pmatrix} = m \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_3} & \frac{1}{\pi_3} \\ \frac{1}{\pi_3} & \frac{1}{\pi_2} + \frac{1}{\pi_3} \end{pmatrix}.$$

And its inverse is given by

$$\boldsymbol{I}^{-1}(\pi_1, \pi_2) = m^{-1}(\pi_1 \pi_2 \pi_3) \begin{pmatrix} \frac{1}{\pi_2} + \frac{1}{\pi_3} & -\frac{1}{\pi_3} \\ -\frac{1}{\pi_3} & \frac{1}{\pi_1} + \frac{1}{\pi_3} \end{pmatrix} = m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}.$$

Suppose that π_1 is our parameter of interest and π_2 is a nuisance parameter. According to the FI matrix above and the information inequality, we can say that, for any unbiased estimator $\hat{\pi}_1$ of π_1,

$$Var(\hat{\pi}_1) \geq n^{-1}m^{-1} \left(\frac{1}{\pi_1} + \frac{1}{\pi_3} \right)^{-1} = \frac{\pi_1(1 - \pi_1 - \pi_2)}{nm(1 - \pi_2)}, \text{ if } \pi_2 \text{ is known.}$$
$$Var(\hat{\pi}_1) \geq \frac{\pi_1(1-\pi_1)}{nm}, \text{ if } \pi_2 \text{ is unknown. } \square$$

Theorem 3.5 (CRLB Attainment–Multiparameter case)**.** *Under the regularity conditions stated above, if the "statistic" $\boldsymbol{T}$ as defined by*

$$\boldsymbol{T} = \boldsymbol{g}(\boldsymbol{\theta}) + \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f_n(\boldsymbol{X}; \boldsymbol{\theta}),$$

do not depend on $\boldsymbol{\theta}$, then $\boldsymbol{T}$ is efficient for $\boldsymbol{g}(\boldsymbol{\theta})$.

Example 3.15. Consider the trinomial distribution of our previous example. Let's see if we can find an efficient estimator for $\boldsymbol{\pi}^t = (\pi_1, \pi_2)$. We have that

$$\boldsymbol{\pi} + \boldsymbol{I}_n^{-1}(\boldsymbol{\pi}) \boldsymbol{\nabla}_{\boldsymbol{\pi}} \log f_n(\boldsymbol{X}; \boldsymbol{\pi}) = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} + n^{-1}m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix} \begin{pmatrix} \frac{\sum_i X_{1i}}{\pi_1} - \frac{\sum_i X_{3i}}{\pi_3} \\ \frac{\sum_i X_{2i}}{\pi_2} - \frac{\sum_i X_{3i}}{\pi_3} \end{pmatrix} = \begin{pmatrix} \frac{\sum_i X_{1i}}{nm} \\ \frac{\sum_i X_{2i}}{nm} \end{pmatrix}.$$

So $\left(\frac{\overline{X}_1}{m}, \frac{\overline{X}_2}{m}\right)$ is efficient for (π_1, π_2). $\square$

Chapter 4

Asymptotic evaluations

All the criteria we have examined so far are finite sample size criteria. In contrast, we might consider asymptotic properties, properties describing the behavior of a procedure *as the sample size becomes infinite* or very large. Describing the distribution of an estimator for a fixed (finite) sample size n is usually very difficult (if not impossible). But in the limit (or asymptotic) regime, things become more structured and generally lead to practical and powerful solutions.

4.1 Convergence in probability and consistency

Definition 4.1 (Convergence in probability)**.** A sequence of random variables X_n, $n = 1, 2, \ldots$, is said to converge to X in probability, in symbols $X_n \xrightarrow{p} X$, if for every $\epsilon > 0$,

$$P(|X_n - X| \geq \epsilon) \to 0 \quad \text{when} \quad n \to \infty.$$

Roughly speaking, $X_n \xrightarrow{p} X$ means that, as n increases, X_n and X get closer to each other (*with a probability approaching 1*), i.e. $P(X_n \approx X) \approx 1$, when n becomes large. In most practical situations, X is a constant (non-random).

Example 4.1.

- Let X_n be a sequence of random variables such that $X_n \sim Ber(1/n)$, $n = 1, 2, \ldots$.

  $$P(|X_n - 0| \geq \epsilon) = P(X_n \geq \epsilon) = P(X_n = 1) = 1/n \to 0.$$

  So $X_n \xrightarrow{p} 0$.

- Let X_n be a sequence of random variables such that $X_n = (1 + 1/n)X$, $n = 1, 2, \ldots$, with $P(|X| < 10) = 1$.

  $$P(|X_n - X| \geq \epsilon) = P(|X| \geq n\epsilon) \leq P(|X| \geq 10) = 0, \forall n \geq 10\epsilon^{-1}.$$

So $X_n \xrightarrow{p} X$. □

Note also that the definition above says nothing about the rate of convergence (how fast X_n converges to X). The figure below illustrates the concept of convergence of probability and convergence speed.

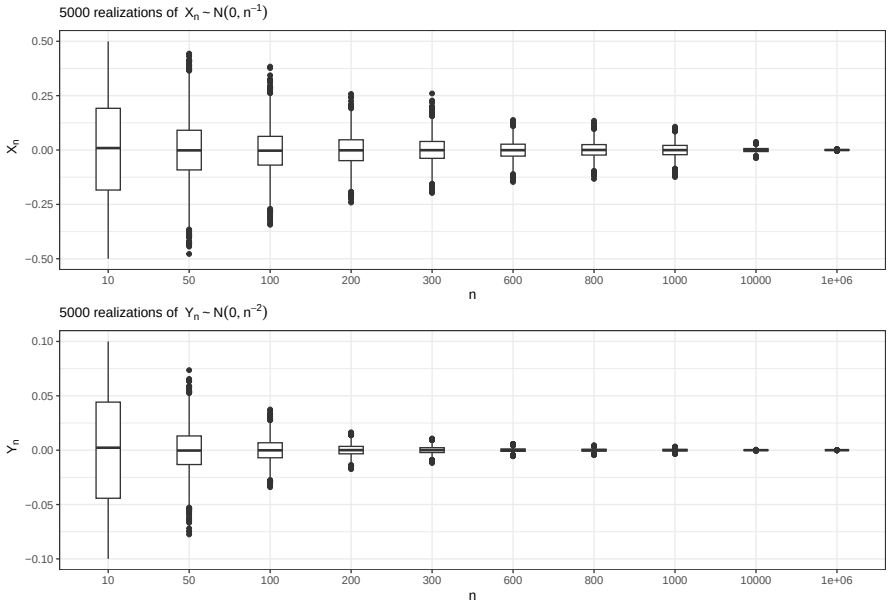


Figure 4.1: Both X_n and $Y_n \xrightarrow{p} 0$ but Y_n goes faster to 0.

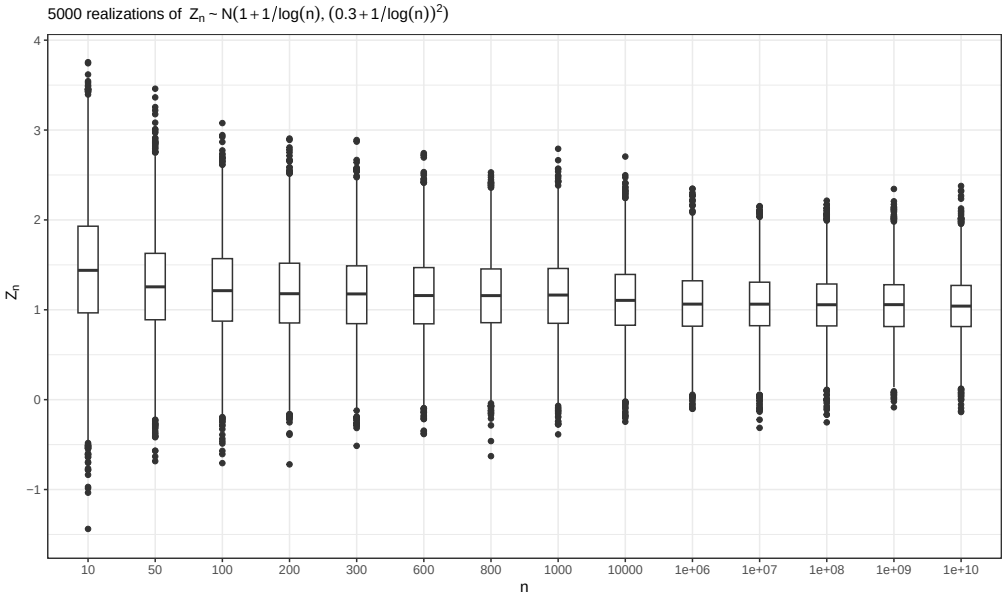


Figure 4.2: $Z_n \xrightarrow{p} N(1, 0.3^2)$.

Facts to know

Convergence in probability is closed under all arithmetic operations. Thus, if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then

- $a_n X_n + b_n \xrightarrow{p} aX + b$, if $a_n \to a$ and $b_n \to b$.
- $X_n \pm Y_n \xrightarrow{p} X \pm Y$.
- $X_n Y_n \xrightarrow{p} XY$.
- $1/X_n \xrightarrow{p} 1/X$, if $X \neq 0$ (with probability one).

For example, let X_n be a sequence of random variables such that $X_n \sim N(\mu_n, \sigma_n^2)$, $n = 1, 2, \ldots$, with $\sigma_n^2 > 0$ and $\mu_n \to \mu$. Let $Z \sim N(0, 1)$. Applying the first of the above properties, we can immediately conclude that if $\sigma_n^2 \to \sigma^2 \geq 0$, then $X_n \xrightarrow{p} X := \mu + \sigma Z$, with $X \sim N(\mu, \sigma^2)$. In short, we write $X_n \xrightarrow{p} N(\mu, \sigma^2)$. If $\sigma = 0$, then $X_n \xrightarrow{p} \mu$.

The concept of convergence in probability is easily generalized to the multivariate case as follows

$$(X_n, Y_n) \xrightarrow{p} (X, Y) \iff X_n \xrightarrow{p} X \text{ and } Y_n \xrightarrow{p} Y.$$

Definition 4.2 (Consistency)**.** Let $\hat{\theta}_n \equiv \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of a parameter θ, $\theta \in \Theta \subset \mathbb{R}$. $\hat{\theta}_n$ is said to be consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$, $\forall \theta \in \Theta$.

In the case of *multiple parameters*, the consistency of a parameter vector estimator is defined as the consistency of each of its components. Thus, for example, we say $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{1n}, \hat{\theta}_{2n})$ is consistent for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ if $\hat{\theta}_{1n}$ is consistent for θ_1, and $\hat{\theta}_{2n}$ is consistent for θ_2.

The following theorem states one of the most useful result for verifying the consistency of an estimator provided its MSE exists.

Theorem 4.1 (Consistent in mean square error)**.** *$\hat{\theta}_n$ is a consistent estimator for θ if*

$$MSE_\theta(\hat{\theta}_n) := E_\theta(\hat{\theta}_n - \theta)^2 \xrightarrow[n \to \infty]{} 0, \ \forall \theta \in \Theta.$$

Note that, by definition, $MSE(\hat{\theta}_n) = Bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n)$. Thus, this theorem can be equivalently restated as $E(\hat{\theta}_n) \to \theta$ and $Var(\hat{\theta}_n) \to 0$ or as $E(\hat{\theta}_n) \to \theta$ and $E(\hat{\theta}_n^2) \to \theta^2$. An estimator such that $MSE_\theta(\hat{\theta}_n) \to 0$ is referred to as *consistent in mean square* or **MSE-consistent**. And so the above theorem states that the MSE-consistency implies consistency.

Example 4.2.

Let $X_1, \ldots, X_n$, be an iid sample from $Unif(0, \theta)$, $\theta > 0$. We have seen previously that the cdf of $X_{(n)}$ is given by

$$F_{X_{(n)}}(x) = \begin{cases} 0, & \text{if } x < 0 \\ (x/\theta)^n, & \text{if } 0 \leq x \leq \theta \\ 1, & \text{if } x > \theta. \end{cases}$$

So, for any $\epsilon > 0$,

$$P(|X_{(n)} - \theta| \geq \epsilon) = P(X_{(n)} \leq \theta - \epsilon) = \begin{cases} 0 & \text{if } \epsilon > \theta \\ \left(1 - \frac{\epsilon}{\theta}\right)^n \to 0 & \text{if } \epsilon \leq \theta. \end{cases}$$

Thus, we conclude that $X_{(n)}$ is a consistent estimator for θ.

Instead of using the definition, we can prove the consistency of $X_{(n)}$ by using the above Theorem 4.1. In fact, we have seen that $E(X_{(n)}) = \frac{n}{n+1}\theta$ and $E\left(X_{(n)}^2\right) = \frac{n}{n+2}\theta^2$. And since $E(X_{(n)}) \to \theta$ and $E(X_{(n)}^2) \to \theta^2$, $X_{(n)}$ is a MSE-consistent estimator for θ. Another way to get to this result is to observe that

$$MSE\left(X_{(n)}\right) = \frac{2\theta^2}{(n+1)(n+2)} \to 0.$$

Another estimator for θ that we previously studied is $\hat{\theta}_1 = 2\overline{X}_n$. We have seen that

$$MSE(\hat{\theta}_1) = Var(\hat{\theta}_1) = \frac{\theta^2}{3n} \to 0.$$

So, $\hat{\theta}_1$ is also a consistent for θ.
But clearly $X_{(n)}$ is better as its MSE converges to 0 faster than the one of $\hat{\theta}_1$ (rate of $1/n^2$ vs $1/n$). □

Statisticians are often interested in the limiting/asymptotic properties of estimators that can be expressed as *arithmetic means or functions of means*. The most basic and most famous result of this type is the following.

Theorem 4.2 (Weak Law of Large Numbers (WLLN))**.** *Suppose X_i, $i = 1, \ldots, n$, is a sequence of iid random variables with finite mean μ, then $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i \xrightarrow{p} \mu$.*

Note that above we made the implicit assumption that $\mu = E(X_i)$ excsites and is finite.

Example 4.3 (Consistency of the empirical distribution function)**.** Let X_i, $i = 1, \ldots, n$, be an iid sample from a cdf $F(x)$. The empirical distribution function is given by $F_n(x) = n^{-1}\sum_{i=1}^n I(X_i \leq x)$. By the WLLN, we have that

$$F_n(x) \xrightarrow{p} E(I(X_1 \leq x)) = F(x). \square$$

Example 4.4 (Consistency of the sample variance)**.** Let X_i, $i = 1, \ldots, n$, be an iid sample with $\mu = E(X_1)$ and $\sigma^2 = Var(X_1)$. The empirical variance is given by

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_i (X_i - \mu)^2 - (\overline{X}_n - \mu)^2\right),$$

By the WLLN, $\frac{1}{n}\sum_i (X_i - \mu)^2 \xrightarrow{p} \sigma^2$, and $\overline{X}_n \xrightarrow{p} \mu$, the fact that the convergence in probability is closed under the arithmetic operations, and $\frac{n}{n-1} \to 1$, we get that $S_n^2 \xrightarrow{p} \sigma^2$. □

4.2 Convergence in law and asymptotic distribution

Convergence in law, also called *convergence in distribution*, is another widely used concept in statistical inference. The definition is given here for the bivariate case, the same applies to any dimension (univariate or multivariate).

Definition 4.3. Let (X_n, Y_n), $n = 1, 2, \ldots$, be a sequence of rv's with cdf $F_n(x, y) = P(X_n \leq x, Y_n \leq y)$. If there exists a *cdf* $F(x, y)$ such that

$$F_n(x, y) \xrightarrow[n\to\infty]{} F(x, y), \ \forall (x, y) \text{ at which } F \text{ is continuous}$$

then F is called *the limiting cdf* of (X_n, Y_n).

Letting (X, Y) have the cdf F, i.e. $F(x, y) = P(X \leq x, Y \leq y)$, then we say that (X_n, Y_n) converges in distribution (or converges in law) to (X, Y), and we denote this by $(X_n, Y_n) \xrightarrow{d} (X, Y)$ or $(X_n, Y_n) \xrightarrow{L} (X, Y)$.

The convergence in distribution is a property of the distribution of X_n rather than X_n itself. $X_n \xrightarrow{d} X$ means that $P(X_n \leq x) \approx P(X \leq x)$, for large n, but *this does not imply convergence in probability*, i.e. $X_n \xrightarrow{d} X \not\Longrightarrow X_n \xrightarrow{p} X$. In other words, $X_n \xrightarrow{d} X$ *does not imply that* X_n *gets closer to* X *as* n *goes to infinity*.

Example 4.5. Let $X \sim Ber(1/2)$, i.e. $P(X = 0) = P(X = 1) = 1/2$, and let $X_n = 1 + \frac{1}{n} - X$, $n \geq 1$. We have that

$$F_n(y) = P(X_n \leq x) = P(1 - X \leq x - 1/n) = \begin{cases} 0, & \text{if } x < 1/n \\ 1/2, & \text{if } 1/n \leq x < 1 + 1/n \\ 1, & \text{if } x \geq 1 + 1/n, \end{cases}$$

$$\xrightarrow[n\to\infty]{} \begin{cases} 0, & \text{if } x < 0 \\ 1/2, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } x \geq 1. \end{cases}$$

This latter is the cdf of $Ber(1/2)$. So, we can write that $X_n \xrightarrow{d} Ber(1/2)$ or, equivalently, that $X_n \xrightarrow{d} X$. Observe that $|X_n - X| = |\pm 1 + 1/n| \geq 1/2$, $\forall n \geq 2$. Hence, $X_n \not\xrightarrow{p} X$. □

Example 4.6. Let X_n be a rv with cdf $F_n(x) = (1 - 1/x^n) I(x \geq 1)$. Put $Y_n = nX_n - n$. We have that

$$G_n(y) := P(Y_n \leq y) = P(X_n \leq 1 + y/n) = \left(1 - \frac{1}{(1 + y/n)^n}\right) I(y \geq 0).$$

L'Hôpital's rule can be used to verify that $(1+y/n)^n \to e^y$. So,

$$G_n(y) \xrightarrow[n\to\infty]{} \begin{cases} 0, & \text{if } y < 0 \\ 1-e^{-y}, & \text{if } y \geq 0. \end{cases}$$

This latter is the cdf of the exponential distribution with rate 1. We therefore conclude that $Y_n \xrightarrow{d} Expo(1)$. □

Example 4.7. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Unif[0,1]$. Put $Y_n = nX_{(1)}$, we have that

$$P(Y_n \leq y) = 1 - (1 - P(X_1 \leq y/n))^n = \begin{cases} 1-(1-0)^n = 0, & \text{if } y < 0 \\ 1-(1-y/n)^n, & \text{if } 0 \leq y \leq n \\ 1-(1-1)^n = 1, & \text{if } y > n. \end{cases}$$
$$\xrightarrow[n\to\infty]{} \begin{cases} 0, & \text{if } y < 0 \\ 1-e^{-y}, & \text{if } y \geq 0. \end{cases}$$

So, $Y_n \xrightarrow{d} Expo(1)$. □

Joint versus marginal convergence in distribution

- If $(X_n, Y_n) \xrightarrow{d} (X, Y)$, then $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$.
- But, unlike convergence in probability, marginal convergence in distribution does not (generally) imply joint convergence in distribution. In other words,

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} Y \not\Longrightarrow (X_n, Y_n) \xrightarrow{d} (X, Y).$$

- Sometimes, marginal convergence in distribution does imply joint convergence in distribution. This is for example the case if X (or Y) is a constant or $X_n \perp\!\!\!\perp Y_n$, $\forall n$.

Example 4.8. Let $X_n = X + 1/n$ with $X \sim N(0,1)$, and $Y_n = -X_n$. Clearly, $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} X$, but $(X_n, Y_n) \not\xrightarrow{d} (X, X)$. □

Convergence in probability is stronger than convergence in distribution. Thus,

$$\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{X} \Rightarrow \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}.$$

A special case of convergence in distribution occurs when the limiting distribution degenerates, thus when $F(x) = I(x \geq c)$, for some $c \in \mathbb{R}$, i.e. X is constant with $P(X = c) = 1$. In this case, X_n converges in distribution to a constant, and we write $X_n \xrightarrow{d} c$, i.e. $P(X_n \leq x) \to 1$, $\forall x > c$ *and* $P(X_n \leq x) \to 0$, $\forall x < c$. It can be shown that

$$\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{c} \Leftrightarrow \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{c}.$$

Example 4.9. Let $X_n \sim Unif(0, 1/n)$. We have that

$$P(X_n \leq x) = \begin{cases} 0, & \text{if } x < 0 \\ nx, & \text{if } 0 \leq x \leq 1/n \\ 1, & \text{if } x > 1/n, \end{cases}$$

$$\xrightarrow[n\to\infty]{} \begin{cases} 0, & \text{if } x \leq 0 \\ 1, & \text{if } x > 0. \end{cases}$$

$$= I(x \geq 0), \forall x \neq 0.$$

So, $X_n \xrightarrow{d} 0$. Thus, $X_n \xrightarrow{p} 0$. In fact, $P(|X_n - 0| \geq \epsilon) = P(X_n \geq \epsilon) = 1 - P(X_n \leq \epsilon) = 0$, $\forall n > 1/\epsilon$. □

Convergence in probability or in distribution does not imply convergence in mean. Thus,

$$X_n \xrightarrow{p \text{ or } d} X \not\Rightarrow E(X_n) \to E(X).$$

Example 4.10. Let X_n be a sequence of rv such that $P(X_n = 0) = 1 - 1/n$ and $P(X_n = n) = 1/n$, $n = 1, 2, \ldots$. For any $\epsilon > 0$, we have that $P(|X_n| \geq \epsilon) = P(X_n = n) = 1/n \to 0$. So $X_n \xrightarrow{p} 0$. But $E(X_n) = 1$, $\forall n$. □

A sufficient condition to ensure that $X_n \xrightarrow{p \text{ or } d} X \Rightarrow E(X_n) \to E(X)$ is that $\sup_n E(|X_n|^{1+\delta}) < \infty$, for some $\delta > 0$.

The behavior of the sample mean is very important, and especially its limiting distribution. In this respect, one of the most remarkable theorems in statistics is the central limit theorem (CLT).

Theorem 4.3 (CLT)**.**

- *(Univariate case). Let X_i, $i = 1, \ldots, n$, be an iid rv's with mean μ and with variance $\sigma^2 \in (0, \infty)$. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

- *(Multivariate case). Let $\boldsymbol{X}_i$, $i = 1, \ldots, n$, be a d-dimensional iid rve's with mean $\boldsymbol{\mu}$ and with a variance-covariance matrix $Var(\boldsymbol{X}_1) = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ has finite entries. Then*

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} N_d(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

So, according to the (univariate) CLT, using the definition of convergence in distribution, with some abuse of notations we can write that

$$P(\overline{X}_n \leq x) \approx P\left(N(\mu, \sigma^2/n) \leq x\right), \text{ for large } n.$$

$N(\mu, \sigma^2/n)$ is called the ***asymptotic distribution*** of $\overline{X}_n$, and we write this down as

$$\overline{X}_n \sim_a N(\mu, \sigma^2/n), \text{ or } \overline{X}_n \sim AN(\mu, \sigma^2/n).$$

Note that $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ implies that $\overline{X}_n \xrightarrow{p} \mu$. So CLT is stronger than WLLN. $\overline{X}_n \xrightarrow{p} \mu$ means that $\overline{X}_n$ gets closer to μ as n increases but it says nothing about how $\overline{X}_n$ varies around μ which is what CLT is about.

Example 4.11. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Pois(\lambda)$ with pd $f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x = 0, 1, \ldots$, and $\lambda > 0$. We have that $E(X_1) = Var(X_1) = \lambda$, $E(I(X_1 = 0)) = e^{-\lambda}$, $Var(I(X_1 = 0)) = e^{-\lambda}(1 - e^{-\lambda})$, and $Cov(X_1, I(X_1 = 0)) = -\lambda e^{-\lambda}$. Let's find the asymptotic distribution of $\overline{X}_n$, the asymptotic distribution of $\overline{Z}_n$, where $Z_i = I(X_i = 0)$, and the joint asymptotic distribution of $(\overline{X}_n, \overline{Z}_n)$.

By the (univariate) CLT, $\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{d} N(0, \lambda)$ and $\sqrt{n}(\overline{Z}_n - e^{-\lambda}) \xrightarrow{d} N\left(0, e^{-\lambda}(1 - e^{-\lambda})\right)$. Thus, $\overline{X}_n \sim_a N(\lambda, \lambda/n)$ and $\overline{Z}_n \sim_a N\left(e^{-\lambda}, e^{-\lambda}(1 - e^{-\lambda})/n\right)$.

By the (multivariate) CLT

$$\sqrt{n}\left(\begin{pmatrix} \overline{X}_n \\ \overline{Z}_n \end{pmatrix} - \begin{pmatrix} \lambda \\ e^{-\lambda} \end{pmatrix}\right) = \begin{pmatrix} \sqrt{n}(\overline{X}_n - \lambda) \\ \sqrt{n}(\overline{Z}_n - e^{-\lambda}) \end{pmatrix} \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda & -\lambda e^{-\lambda} \\ -\lambda e^{-\lambda} & e^{-\lambda}(1 - e^{-\lambda}) \end{pmatrix}\right). \tag{4.1}$$

Thus,

$$\begin{pmatrix} \overline{X}_n \\ \overline{Z}_n \end{pmatrix} \sim_a N_2\left(\begin{pmatrix} \lambda \\ e^{-\lambda} \end{pmatrix}, n^{-1}\begin{pmatrix} \lambda & -\lambda e^{-\lambda} \\ -\lambda e^{-\lambda} & e^{-\lambda}(1 - e^{-\lambda}) \end{pmatrix}\right). \square$$

Note that Equation (4.1) can also be written as

$$\begin{pmatrix} \sqrt{n}(\overline{X}_n - \lambda) \\ \sqrt{n}(\overline{Z}_n - e^{-\lambda}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} U \\ V \end{pmatrix}, \text{ with } \begin{pmatrix} U \\ V \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda & -\lambda e^{-\lambda} \\ -\lambda e^{-\lambda} & e^{-\lambda}(1 - e^{-\lambda}) \end{pmatrix}\right),$$

implying that $\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{d} N(0, \lambda)$ and $\sqrt{n}(\overline{Z}_n - e^{-\lambda}) \xrightarrow{d} N\left(0, e^{-\lambda}(1 - e^{-\lambda})\right)$, which match the results obtained above when we applied the univariate CLT.

Exercise 4.1. Let X_i, $i = 1, \ldots, n$, be an iid sample from the pd

$$f(x; \tau) = (2\tau)^{-1} e^{-|x|/\tau}, \ \tau > 0.$$

Find the joint asymptotic distribution of $(\overline{X}_n, \overline{Y}_n)$, where $Y_i = |X_i|$.

4.3 Tools for proving asymptotic results

There are many techniques available to check the consistency of an estimator and to find its asymptotic distribution. We present the main useful techniques below.

4.3.1 Continuous mapping theorem

Theorem 4.4 (Continuous Mapping Theorem (CMT))**.** *Let $\boldsymbol{X}_n$ be a sequence of d-dimensional random vectors and $\boldsymbol{X}$ a d-dimensional random vector. Let $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^p$ be a function* ***continuous*** *on some $\boldsymbol{I} \subset \mathbb{R}^d$ for which $P(\boldsymbol{X} \in \boldsymbol{I}) = 1$. Then,*

$$\boldsymbol{X}_n \to \boldsymbol{X} \Rightarrow \boldsymbol{g}(\boldsymbol{X}_n) \to \boldsymbol{g}(\boldsymbol{X}).$$

Above, $\to$ can be either the convergence in probability ($\xrightarrow{p}$) or in distribution ($\xrightarrow{d}$).

The CMT allows the function $\boldsymbol{g}$ to be discontinuous but the probability of $\boldsymbol{X}$ being at a discontinuity point of $\boldsymbol{g}$ must be zero. For example, suppose $X_n \to X$. Since the function $x \mapsto 1/x$ is discontinuous at 0, we can apply the CMT to conclude that $1/X_n \to 1/X$, provided that $P(X = 0) = 0$.

CMT implies that the convergence in probability (or in distribution) is closed under all arithmetic operations. So, if $(X_n, Y_n) \to (X, Y)$, then $aX_n + bY_n + c \to aX + bY + c$, $\forall a, b, c$, $X_nY_n \to XY$, and $X_n/Y_n \to X_n/Y$, if $P(Y = 0) = 0$.

Example 4.12.

From the examples studied above, we can conclude that $S_n \xrightarrow{p} \sigma$, $S_n/\overline{X}_n \xrightarrow{p} \sigma/\mu$ (if $\mu \neq 0$), $n\frac{(\overline{X}_n - \mu)^2}{\sigma^2} \xrightarrow{d} \chi^2_1$, and, from Example 4.11, that $\frac{\overline{X}_n - \lambda}{\overline{Z}_n - e^{-\lambda}} \xrightarrow{d} \frac{U}{V}$, where (U, V) is bivariate-normal with mean and variance as defined above. □

4.3.2 Slutsky's theorem

Theorem 4.5 (Slutsky)**.** *If $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{C}$, where $\boldsymbol{C}$ is a constant, then*

$$\begin{pmatrix} \boldsymbol{X}_n \\ \boldsymbol{Y}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{C} \end{pmatrix}$$

This theorem along with the CMT implies that if $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{C}$, then $\boldsymbol{X}_n \pm \boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{X} \pm \boldsymbol{C}$, $\boldsymbol{X}_n\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{X}\boldsymbol{C}$, and $\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{C}^{-1}\boldsymbol{X}$ (when $\boldsymbol{C} \neq \boldsymbol{0}$). People often call such results (CMT + Slutsky) "Slutsky's theorem".

Example 4.13.

- By CLT we have that, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. And we known that $S_n^2 \xrightarrow{p} \sigma^2$. So, by CMT + Slutsky, we deduce that

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \xrightarrow{d} N(0, 1).$$

- Let X_i, $i = 1, \ldots, n$ be an iid sample from $Ber(\pi)$. Put $\hat{\pi}_n = n^{-1}\sum_{i=1}^n X_i$. The same

reasoning as above leads to

$$\sqrt{n}\frac{\hat{\pi}_n - \pi_n}{\sqrt{\hat{\pi}_n(1-\hat{\pi}_n)}} \xrightarrow{d} N(0,1).$$

- In Example 4.4, we have seen that $S_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2$, where $\hat{\sigma}_n^2 := n^{-1}\sum_i (X_i - \overline{X}_n)^2 = \tilde{\sigma}_n^2 - (\overline{X}_n - \mu)^2$, with $\tilde{\sigma}_n^2 = n^{-1}\sum_i (X_i - \mu)^2$. By CLT, we known that

$$\begin{aligned}
&\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0,\sigma^2), \text{ and} \\
&\sqrt{n}(\tilde{\sigma}_n^2 - \sigma^2) \xrightarrow{d} N(0,\nu^2), \text{ with } \nu^2 := Var(X_1 - \mu)^2.
\end{aligned}$$

  Hence, $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \sqrt{n}(\tilde{\sigma}_n^2 - \sigma^2) - \sqrt{n}(\overline{X}_n - \mu)^2 \xrightarrow{d} N(0,\nu^2)$. And since,

$$\sqrt{n}(S_n^2 - \sigma^2) = \frac{n}{n-1}\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) + \frac{\sqrt{n}}{n-1}\sigma^2,$$

  we conclude that $\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0,\nu^2)$. □

4.3.3 Delta method

Theorem 4.6 (Delta method). *Let $\boldsymbol{g} : \mathbb{R}^d \mapsto \mathbb{R}^p$ such that $\dot{\boldsymbol{g}}$ is continuous in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^d$ and let $a_n \xrightarrow[n\to\infty]{} \infty$. Then,*

$$a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{W} \Rightarrow a_n\left(\boldsymbol{g}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{g}(\boldsymbol{\theta})\right) \xrightarrow{d} \dot{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{W},$$

where $\dot{\boldsymbol{g}}$ is the $p \times d$ Jacobian matrix of $\boldsymbol{g}$, whose $(i,j)-$th element is $\partial_{\theta_j} g_i(\boldsymbol{\theta})$. In particular, if $\boldsymbol{W} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_p\big(\dot{\boldsymbol{g}}(\boldsymbol{\theta})\mu, \dot{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{\Sigma}\dot{\boldsymbol{g}}^t(\boldsymbol{\theta})\big)$.

As a consequence, we have the following results:

- **Scalar case with real-valued function**: $\theta \in \mathbb{R}$ and $g(\theta) : \mathbb{R} \mapsto \mathbb{R}$. Suppose that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0,\sigma^2)$. Then

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} g'(\theta) \times N(0,\sigma^2) = N\left(0, (g'(\theta))^2\sigma^2\right).$$

- **Scalar case with vector-valued function**: $\theta \in \mathbb{R}$ and $\boldsymbol{g}(\theta) := (g_1(\theta), g_2(\theta)) : \mathbb{R} \mapsto \mathbb{R}^2$. Suppose that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0,\sigma^2)$. Then

$$\sqrt{n}\left(\begin{pmatrix} g_1(\hat{\theta}) - g_1(\theta) \\ g_2(\hat{\theta}) - g_2(\theta) \end{pmatrix}\right) \xrightarrow{d} \begin{pmatrix} \partial_\theta g_1(\theta) \\ \partial_\theta g_2(\theta) \end{pmatrix} \times N(0,\sigma^2) = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\boldsymbol{\Lambda}\right),$$

  where $\boldsymbol{\Lambda} = \begin{pmatrix} \partial_\theta g_1(\theta) \\ \partial_\theta g_2(\theta) \end{pmatrix}\begin{pmatrix} \partial_\theta g_1(\theta) & \partial_\theta g_2(\theta) \end{pmatrix} = \begin{pmatrix} (\partial_\theta g_1(\theta))^2 & \partial_\theta g_1(\theta)\partial_\theta g_2(\theta) \\ \partial_\theta g_1(\theta)\partial_\theta g_2(\theta) & (\partial_\theta g_2(\theta))^2 \end{pmatrix}$.

- **Bivariate case with real-valued function**: $\boldsymbol{\theta} := (\theta_1, \theta_2) \in \mathbb{R}^2$ and $g(\boldsymbol{\theta}) = g(\theta_1, \theta_2) : \mathbb{R}^2 \mapsto \mathbb{R}$. Suppose that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{n}\begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \end{pmatrix} \xrightarrow{d} N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\sigma_{ij}]_{i,j=1,2}$. Then,
$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{d} \begin{pmatrix} \partial_{\theta_1} g(\boldsymbol{\theta}) & \partial_{\theta_2} g(\boldsymbol{\theta}) \end{pmatrix} \times N_2(\mathbf{0}, \boldsymbol{\Sigma}) = N(0, \sigma^2),$$
where $\sigma^2 = (\partial_{\theta_1} g(\boldsymbol{\theta}))^2 \sigma_{11} + (\partial_{\theta_2} g(\boldsymbol{\theta}))^2 \sigma_{22} + 2\partial_{\theta_2} g(\boldsymbol{\theta}) \partial_{\theta_1} g(\boldsymbol{\theta}) \sigma_{12}$.
- **Bivariate case with vector-valued function**: $\boldsymbol{\theta} := (\theta_1, \theta_2) \in \mathbb{R}^2$ and $\boldsymbol{g}(\boldsymbol{\theta}) := (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta})) : \mathbb{R}^2 \mapsto \mathbb{R}^2$. Suppose that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_2(\mathbf{0}, \boldsymbol{\Sigma})$. Then,
$$\sqrt{n}\begin{pmatrix} g_1(\hat{\boldsymbol{\theta}}) - g_1(\boldsymbol{\theta}) \\ g_2(\hat{\boldsymbol{\theta}}) - g_2(\boldsymbol{\theta}) \end{pmatrix} \xrightarrow{d} \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \times N_2(\mathbf{0}, \boldsymbol{\Sigma}) = N_2(\mathbf{0}, \boldsymbol{\Lambda}),$$
where $\boldsymbol{\Lambda} = \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \boldsymbol{\Sigma} \dot{\boldsymbol{g}}^t(\boldsymbol{\theta})$, with $\dot{\boldsymbol{g}}(\boldsymbol{\theta}) = \begin{pmatrix} \partial_{\theta_1} g_1(\boldsymbol{\theta}) & \partial_{\theta_2} g_1(\boldsymbol{\theta}) \\ \partial_{\theta_1} g_2(\boldsymbol{\theta}) & \partial_{\theta_2} g_2(\boldsymbol{\theta}) \end{pmatrix}$.

Example 4.14.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Pois(\lambda)$ with pd $f(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x = 0, 1, \ldots$, and $\lambda > 0$. By CLT, $\sqrt{n}\left(\overline{X}_n - \lambda\right) \xrightarrow{d} N(0, \lambda)$. So, by Delta method,
$$\sqrt{n}\left(e^{-\overline{X}_n} - e^{-\lambda}\right) \xrightarrow{d} N\left(0, \lambda((e^{-\lambda})')^2\right) = N\left(0, \lambda e^{-2\lambda}\right)$$
- Let X_i, $i = 1, \ldots, n$ be an iid sample from $Ber(\pi)$. Let $\hat{\pi} \equiv \hat{\pi}_n = n^{-1} \sum_i X_i$. By CLT, we know that $\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} N(0, \pi(1 - \pi))$. Put $O = \dfrac{\pi}{1 - \pi}$ (this is called the odds) and $\hat{O} = \dfrac{\hat{\pi}}{1 - \hat{\pi}}$. By Delta method,
$$\sqrt{n}(\hat{O} - O) \xrightarrow{d} N\left(0, \pi(1 - \pi)\left(\left(\frac{\pi}{1 - \pi}\right)'\right)^2\right) = N\left(0, \frac{\pi}{(1 - \pi)^3}\right) = N\left(0, O(1 + O)^2\right).$$
- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Pois(\lambda)$ with pd $f(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x = 0, 1, \ldots$, and $\lambda > 0$. By CLT, $\sqrt{n}\left(\overline{X}_n - \lambda\right) \xrightarrow{d} N(0, \lambda)$. So, by the Delta method, with $g : \lambda \mapsto (\lambda, e^{-\lambda})$,
$$\sqrt{n}\begin{pmatrix} \overline{X}_n - \lambda \\ e^{-\overline{X}_n} - e^{-\lambda} \end{pmatrix} \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \lambda \begin{pmatrix} 1 \\ -e^{-\lambda} \end{pmatrix}\begin{pmatrix} 1 & -e^{-\lambda} \end{pmatrix}\right) = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda & -\lambda e^{-\lambda} \\ -\lambda e^{-\lambda} & \lambda e^{-2\lambda} \end{pmatrix}\right).$$
- Let (X_i, Y_i), $i = 1, \ldots, n$, be an iid sample of (X, Y). Let $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = Var(X)$, $\sigma_2^2 = Var(Y)$, and $\sigma_{12} = Cov(X, Y)$. By CLT
$$\sqrt{n}\left(\begin{pmatrix} \overline{X}_n \\ \overline{Y}_n \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right) \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right)$$

So, by the Delta method, with $g:\ (\mu_1,\mu_2)\mapsto \mu_1-\mu_2$,

$$\sqrt{n}((\overline{X}_n-\overline{Y}_n)-(\mu_1-\mu_2)) \xrightarrow{d} N(0,\sigma_1^2+\sigma_2^2-2\sigma_{12}).$$

Using again the Delta method, but this time with $g:\ (\mu_1,\mu_2)\mapsto \mu_1/\mu_2$, we get

$$\sqrt{n}(\overline{X}_n/\overline{Y}_n-\mu_1/\mu_2) \xrightarrow{d} N(0,\nu^2),$$

with $\nu^2=\frac{1}{\mu_2^2}\left(\sigma_1^2+\frac{\mu_1^2}{\mu_2^2}\sigma_2^2-2\frac{\mu_1}{\mu_2}\sigma_{12}\right)$.

- Let X_i, $i=1,\ldots,n$, be an iid sample from $Ber(\pi_1)$ and Y_i, $i=1,\ldots,n$, be another iid sample from $Ber(\pi_2)$. Suppose that these two samples are independent. Let $\hat{\pi}_1=n^{-1}\sum_i X_i$, and $\hat{\pi}_2=n^{-1}\sum_i Y_i$. By CLT,

$$\sqrt{n}\left(\begin{pmatrix}\hat{\pi}_1\\ \hat{\pi}_2\end{pmatrix}-\begin{pmatrix}\pi_1\\ \pi_2\end{pmatrix}\right) \xrightarrow{d} N_2\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\pi_1(1-\pi_1) & 0\\ 0 & \pi_1(1-\pi_1)\end{pmatrix}\right)$$

So, by the Delta method, with $g:\ (\pi_1,\pi_2)\mapsto(\pi_1-\pi_2,\pi_1/\pi_2)$,

$$\sqrt{n}\left(\begin{pmatrix}\hat{\pi}_1-\hat{\pi}_2\\ \hat{\pi}_1/\hat{\pi}_2\end{pmatrix}-\begin{pmatrix}\pi_1-\pi_2\\ \pi_1/\pi_2\end{pmatrix}\right) \xrightarrow{d} N_2\left(\begin{pmatrix}0\\ 0\end{pmatrix},\mathbf{\Lambda}\right),$$

where $\mathbf{\Lambda}=\begin{pmatrix}1 & -1\\ 1/\pi_2 & -\pi_1/\pi_2^2\end{pmatrix}\begin{pmatrix}\pi_1(1-\pi_1) & 0\\ 0 & \pi_2(1-\pi_2)\end{pmatrix}\begin{pmatrix}1 & 1/\pi_2\\ -1 & -\pi_1/\pi_2^2\end{pmatrix}$. □

4.4 Asymptotic efficiency

In general, for a given estimator $\hat{\theta}_n$ of a parameter θ, if one can show that $\dfrac{\hat{\theta}_n-\theta_n}{\sigma_n} \xrightarrow{d} N(0,1)$, then we say that $\hat{\theta}_n$ is *asymptotically distributed as* $N(\theta_n,\sigma_n^2)$, or that $\hat{\theta}_n$ is ***asymptotically normal*** with ***asymptotic mean*** $Amean(\hat{\theta}_n):=\theta_n$ and ***asymptotic variance*** $Avar(\hat{\theta}_n):=\sigma_n^2$, and write

$$\hat{\theta}_n \sim_a N(\theta_n,\sigma_n^2),\ \text{or}\ \hat{\theta}_n\sim AN(\theta_n,\sigma_n^2).$$

The quantity $Abias(\hat{\theta}_n):=\theta_n-\theta$ is called the ***asymptotic bias***. If $\theta_n=\theta$, then we say that $\hat{\theta}_n$ is *asymptotically unbiased*.

For two asymptotically unbiased estimators of θ, the *asymptotic relative efficiency* of $\hat{\theta}_1$ to $\hat{\theta}_2$ is the ratio of the asymptotic variance of $\hat{\theta}_2$ to the asymptotic variance of $\hat{\theta}_1$:

$$ARE(\hat{\theta}_1,\hat{\theta}_2)=\frac{Avar(\hat{\theta}_2)}{Avar(\hat{\theta}_1)}.$$

Example 4.15. Let X_i, $i=1,\ldots,n$, be an iid sample from $Pois(\lambda)$ with pd $f(x,\lambda)=\dfrac{\lambda^x}{x!}e^{-\lambda}$,

$x = 0, 1, \ldots$, and $\lambda > 0$. Let's say that our parameter of interest is $\delta := P(X = 0) = e^{-\lambda} \in (0, 1)$. Put $\hat{\delta}_1 = n^{-1} \sum_{i=1}^{n} I(X_i = 0)$, and $\hat{\delta}_2 = e^{-\overline{X}_n}$. By CLT,

$$\sqrt{n}\left(\hat{\delta}_1 - \delta\right) \xrightarrow{d} N\left(0, \delta(1-\delta)\right).$$

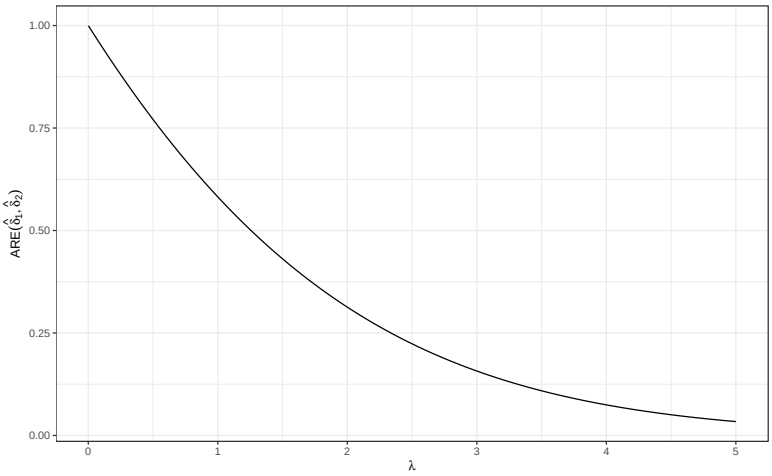
Thus, $\hat{\delta}_1 \sim_a N\left(\delta, e^{-\lambda}(1 - e^{-\lambda})/n\right)$. On the other hand, we have that (see Example 4.14)

$$\sqrt{n}\left(e^{-\overline{X}_n} - e^{-\lambda}\right) \xrightarrow{d} N(0, \lambda e^{-2\lambda}).$$

Thus, $\hat{\delta}_2 \sim_a N(\delta, \lambda e^{-2\lambda}/n)$. We conclude that

$$ARE(\hat{\delta}_1, \hat{\delta}_2) = \frac{\lambda}{e^{\lambda} - 1}.$$

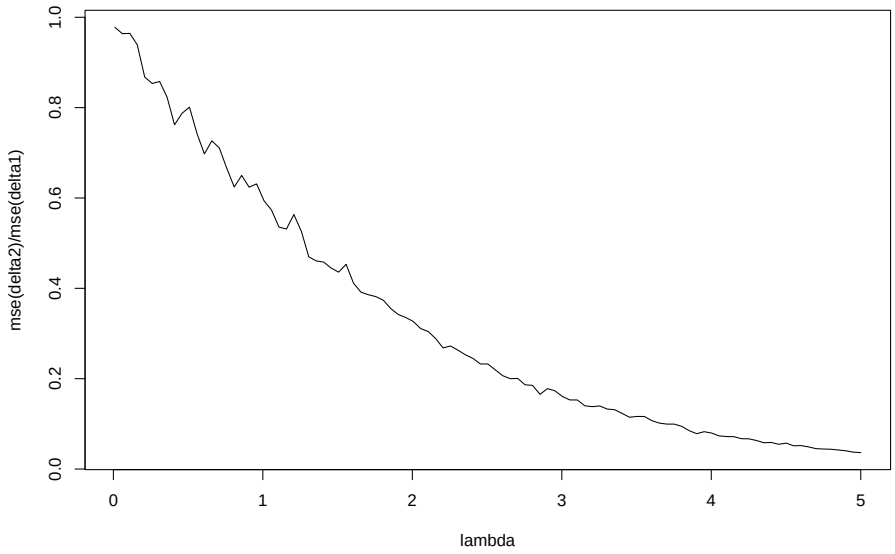
This is a strictly decreasing function of λ with a maximum of 1.



So, $\hat{\delta}_2$ *is asymptotically uniformly more efficient than* $\hat{\delta}_1$. And, the larger λ is, the better $\hat{\delta}_2$ is.

The following simulation confirms the calculations above (for $n = 100$).

```
REpois <- function(lambda, n = 100, rep = 5000) {
  delta <- exp(-lambda)
  hat.delta1 <- replicate(rep, mean(rpois(n, lambda) == 0))
  hat.delta2 <- replicate(rep, exp(-mean(rpois(n, lambda))))
  mse(hat.delta2, delta) / mse(hat.delta1, delta)   # Relative MSE
}
curve(Vectorize(REpois)(x), from = 0.01, to = 5, xlab = "lambda",
      ylab = "mse(delta2)/mse(delta1)")
```



□

We have seen that if X_i, $i = 1, \ldots, n$, is an iid sample from a pd $f(x, \theta)$ and $\hat{\theta}$ is any *unbiased estimator* of θ, then

$$Var(\hat{\theta}) \geq \frac{1}{I_n(\theta)}. \tag{4.2}$$

Under some regularity conditions (see the previous chapter), this property holds *for any finite sample size* n. The bound $I_n^{-1}(\theta)$ is attainable if $f(x, \theta)$ belongs to the exponential family (see the CRLB Attainment theorem). Although this family is very rich, this considerably limits the applicability of such a result.

In asymptotic regime, under some some regularity conditions, it can be shown that for *any asymptotically normal and asymptotically unbiased* estimator $\hat{\theta}$ of θ,

$$Avar(\hat{\theta}) \geq \frac{1}{I_n(\theta)}. \tag{4.3}$$

The regularity conditions that guarantee the validity of this asymptotic result are less restrictive than those required for the finite-sample variant (Equation (4.2)). Furthermore, the number of models/families for which the limit in (4.3) is attainable is much larger.

An asymptotically normal and asymptotically unbiased estimator that attains the lower bound in (4.3) is said to be *asymptotically efficient* for θ. More precisely, $\hat{\theta}$ is asymptotically efficient for θ if $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$, i.e. if $\hat{\theta} \sim_a N(\theta, I_n^{-1}(\theta))$. The same concept applies to the case of multiple parameters, resulting in the following general definition: $\hat{\boldsymbol{\theta}}$ is asymptotically efficient for $\boldsymbol{\theta} \in \mathbb{R}^d$ if

$$\hat{\boldsymbol{\theta}} \sim_a N_d(\boldsymbol{\theta}, \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})).$$

An interesting property of "asymptotic efficiency" is

$$\hat{\boldsymbol{\theta}} \text{ is asymptotically efficient for } \boldsymbol{\theta} \implies \boldsymbol{g}(\boldsymbol{\theta}) \text{ is asymptotically efficient for } \boldsymbol{g}(\boldsymbol{\theta}).$$

$\boldsymbol{g} : \mathbb{R}^d \mapsto \mathbb{R}^p$ is any function with continuous Jacobian. The reverse is also true if $\boldsymbol{g}$ is bejective. This result is a direct consequence of the Delta method. In fact, if $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$, then, by Delta method,

$$\sqrt{n}\big(\boldsymbol{g}(\hat{\boldsymbol{\theta}}) - \boldsymbol{g}(\boldsymbol{\theta})\big) \xrightarrow{d} N_p\Big(\mathbf{0}, \dot{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})\dot{\boldsymbol{g}}^t(\boldsymbol{\theta})\Big).$$

Thus, $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \sim_a N_p\big(\boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta}))\big)$.

Example 4.16. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Pois(\lambda)$ with pd $f(x;\lambda) = \dfrac{\lambda^x}{x!}e^{-\lambda}$, $x = 0, 1, \ldots,$ and $\lambda > 0$.
We have seen that $\sqrt{n}(\overline{X} - \lambda) \xrightarrow{d} N(0, \lambda)$ and that $I(\lambda) = 1/\lambda$. So, $\overline{X}$ is asymptotically efficient for λ (actually, $\overline{X}$ is efficient for λ).
As consequence, we can, for example, say that $e^{-\overline{X}}$ is asymptotically efficient for $e^{-\lambda}$. In fact, by Delta method, $\sqrt{n}(e^{-\overline{X}} - e^{-\lambda}) \xrightarrow{d} N(0, \lambda)(e^{-\lambda})' = N(0, I^{-1}(\lambda))$. Note that it can be shown that $e^{-\overline{X}}$ is not efficient for $e^{-\lambda}$.
The asymptotic efficiency of $\overline{X}$ also implies that of, for example, $(\overline{X}, e^{-\overline{X}})$ as an estimator of $(\lambda, e^{-\lambda}) = (E(X), P(X = 0))$. □

Chapter 5

Estimation methods

In this section, we examine some general methods that can be used to construct estimators that often have good properties.

5.1 The method of moments (MoM)

Let X_i, $i = 1, \ldots, n$, be an iid sample of X. Let $\mu_j = E(X^j)$ denote the $j-$th moment of X, and $\hat{\mu}_j = n^{-1} \sum_{i=1}^n X_i^j$ the $j-$th sample moment.
To apply the method of moments to the problem of estimating a parameter $\boldsymbol{\theta}$, we need to be able to express $\boldsymbol{\theta}$ as a function of $\mu_1, \mu_2, \ldots$. Thus, we need to find a (known) function, say $\boldsymbol{g}$, such that

$$\boldsymbol{\theta} = \boldsymbol{g}(\mu_1, \mu_2, \ldots).$$

A simple estimation method consists in replacing the μ_j's in the equation above by their empirical versions. This leads to the following MoM estimator:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{g}(\hat{\mu}_1, \hat{\mu}_2, \ldots).$$

MoM are not necessarily the best estimators, but they are typically easy to obtain and, under reasonable conditions, they are consistent and are asymptotically normal. In fact,

- By the WLLN + CMT, a MoM estimator is consistent provided (i) the population moments exist, and (ii) $\boldsymbol{g}$ is a continuous.
- By the CLT + Delta method, a MoM estimator can, typically, be shown to be asymptotically normal.

Example 5.1.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from the exponential distribution with pd $f(x; \lambda) = \lambda e^{-\lambda x} I(x \geq 0)$, $\lambda > 0$. Since $E(X_1) = 1/\lambda$, the MoM estimator of λ is $\hat{\lambda} = 1/\overline{X}_n$. This is a consistent estimator. Moreover, since $Var(X_1) = 1/\lambda^2$, $\hat{\lambda}$ is asymptotically normal with limiting distribution given by $N(\lambda, \lambda^2/n)$.

- Let X_i, $i = 1, \ldots, n$, be an iid sample of X. Let $\mu_j = E(X^j)$, $\sigma_{jk} = Cov(X^j, X^k)$, $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, and $\overline{X_n^2} = n^{-1}\sum_{i=1}^n X_i^2$. Let's find the MoM estimator of $\sigma^2 \equiv \sigma_{11} = Var(X)$ and its asymptotic distribution. Since $\sigma^2 = \mu_2 - \mu_1^2$, the MoM estimator of σ^2 is

$$\hat{\sigma}_n^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \overline{X_n^2} - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

  By CLT,

$$\sqrt{n}\left(\begin{pmatrix} \overline{X}_n \\ \overline{X_n^2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right) \xrightarrow{d} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right)$$

  So, by the Delta method, with $g(x, y) = y - x^2$,

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{d} N(0, v^2),$$

  where $v^2 := 4\mu_1^2\sigma_{11} + \sigma_{22} - 4\mu_1\sigma_{12} = Var(X - \mu_1)^2$.

- Let Y_i, $i = 1, \ldots, n$, be an iid sample from the Log-normal distribution (i.e. $log(Y) \sim N(\mu, \sigma^2)$) with parameters μ and σ^2. Y_i has the pd

$$f(y; \mu, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) I(y > 0),\ \mu \in (-\infty, \infty),\ \text{and}\ \sigma > 0.$$

  Since

$$\mu_1 := E(Y_1) = \exp(\mu + \sigma^2/2) \text{ and } \mu_2 := E(Y_1^2) = \exp(2\mu + 2\sigma^2),$$

  the MoM estimator of μ and σ^2 are

$$\hat{\mu} = 2\log(\hat{\mu}_1) - \frac{1}{2}\log(\hat{\mu}_2) \text{ and } \hat{\sigma}^2 = \log(\hat{\mu}_2) - 2\log(\hat{\mu}_1).$$

  Can you prove that $(\hat{\mu}, \hat{\sigma}^2)$ is consistent? What about asymptotic normality? □

5.2 The method of maximum likelihood (ML)

The method of maximum likelihood is, by far, the most popular technique for deriving estimators and performing inference. It has an intuitive motivation and usually has fairly very good properties (at least asymptotically).

5.2.1 Likelihood: definition and meaning

Definition 5.1 (The likelihood function)**.** Let $f_n(\cdot; \theta)$ denote the joint pd of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $\theta \in \Theta$ is the parameter of interest (θ may be vector valued; we don't bold it here for ease of notation).
Given that $\boldsymbol{X} = \boldsymbol{x} := (x_1, \ldots, x_n)$ is observed, *the function of* θ defined by

$$L_n(\theta|\boldsymbol{x}) = f_n(\boldsymbol{x}; \theta),$$

is called the likelihood function of θ (given the observation $\boldsymbol{X} = \boldsymbol{x}$).

The parameter θ is listed first in L_n because, unlike f_n, L_n is viewed as a function of θ, for a given sample point $\boldsymbol{x}$ of $\boldsymbol{X}$.

If $X_1, \ldots, X_n$ are **iid** with marginal pd $f(x;\theta)$, the likelihood factorizes into

$$L_n(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta)$$

For the discrete case, $L_n(\theta|\boldsymbol{x}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$, where P_θ means that *the probability is taken under the assumption that θ is the true parameter*. So, if we compare the likelihood function at two parameter points, say θ_1 and θ_2, and find, for example, that

$$L_n(\theta_1|\boldsymbol{x}) = P_{\theta_1}(\boldsymbol{X} = \boldsymbol{x}) > L_n(\theta_2|\boldsymbol{x}) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x}),$$

then the sample $\boldsymbol{x}$ we actually observe is *more likely* to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. This can be interpreted by saying that, given the data $\boldsymbol{x}$, θ_1 is a more plausible value for θ than θ_2.
In general, we can think of $L_n(\theta|\boldsymbol{x})$ as a measure of how "likely" θ has produced the observed $\boldsymbol{x}$. A similar interpretation applies to the continuous case.

Example 5.2. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Ber(\pi)$, where π is our parameter of interest. Suppose that we have $n = 6$ observations $x_i = 1, 1, 0, 1, 1, 0$, and $\pi \in \{0.2, 0.3, 0.7, 0.8, 0.9\} = \Theta$.
The likelihood function of the observed data is

$$L_n(\pi|\boldsymbol{x}) = P(X_1 = 1, X_2 = 1, \ldots, X_6 = 0) = \prod_{i=1}^{6} P(X_i = x_i) = \pi^4(1-\pi)^2.$$

π	$L_n(\pi)$
0.2	0.001
0.3	0.004
0.7	0.0216
0.8	0.0164
0.9	0.0066

Given the actual sample, we can say that 0.7 is the most plausible value of π among $\{0.2, 0.3, 0.7, 0.8, 0.9\}$. □

Definition 5.2 (Maximum Likelihood Estimator)**.** A statistic $\hat{\theta} \equiv \hat{\theta}(\boldsymbol{X})$ is called a maximum likelihood estimator (MLE) of $\theta \in \Theta$, if (i) $\hat{\theta} \in \Theta$, and (ii) for each sample point $\boldsymbol{x}$,

$$L_n(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) \geq L_n(\theta|\boldsymbol{x}), \ \forall \theta \in \Theta.$$

In other words, a MLE $\hat{\theta}(\boldsymbol{x})$ is a parameter point at which the likelihood $L_n(\theta|\boldsymbol{x})$, as a function of θ, attains its maximum: $L_n(\hat{\theta}(\boldsymbol{x})|\boldsymbol{x}) = \max_{\theta\in\Theta} L_n(\theta|\boldsymbol{x})$. In mathematical notation, we write

$$\hat{\theta}_n(\boldsymbol{x}) = \arg\max_{\theta\in\Theta} L_n(\theta|\boldsymbol{x}),$$

where arg max is a shortcut for "the arguments of the maxima", i.e. any point at which $\theta \mapsto L_n(\theta|\boldsymbol{x})$ is maximized over Θ.

Facts to know

- MLE *may not exist* and *may not be unique*.
- By construction, the range of MLE coincides with Θ, the range of the parameter.
- MLE is a judicious choice in that it represents the parameter point that is most likely to have generated the data. In general, this is a good point estimator in that it possesses some *optimal properties* that will be examined later.
- However, this method has the inherent drawback of having to find the maximum of the likelihood function, which is often a difficult problem. In fact, it can be challenging to find (analytically or numerically) a *global maximizer* and to ensure that it is indeed a global (and not local) maximizer, especially in the multi-parameter case; more on this later.

5.2.2 MLE implementation

For the sake of generality, we consider here the multi-parameter case, where the parameter of interest $\boldsymbol{\theta}$ is a vector of dimension d.

Maximum likelihood estimators are often found by maximizing the log-likelihood function

$$\ell_n(\boldsymbol{\theta}|\boldsymbol{x}) = \log(L_n(\boldsymbol{\theta}|\boldsymbol{x})) = \sum_{i=1}^{n} \log f(x_i;\boldsymbol{\theta}).$$

Since the logarithmic transformation is strictly monotonically increasing, it does not make any difference to maximize ℓ_n or L_n.

If the likelihood function is differentiable, possible candidates for the MLE are the $\boldsymbol{\theta}$'s that solve the ***likelihood equation***:

$$\partial_{\theta_k}\ell_n(\boldsymbol{\theta}|\boldsymbol{x}) = 0,\ \forall k = 1,\dots,d.$$

This is equivalent to solve the ***Score equation*** $S_n(\boldsymbol{\theta},\boldsymbol{x}) = \mathbf{0}$, where $S_n(\boldsymbol{\theta},\boldsymbol{x})$ is the ***score function*** associated with $f_n(\boldsymbol{x};\boldsymbol{\theta})$, i.e. $S_n(\boldsymbol{\theta},\boldsymbol{x}) = (\partial_{\theta_1}\ell_n(\boldsymbol{\theta}|\boldsymbol{x}),\dots,\partial_{\theta_d}\ell_n(\boldsymbol{\theta}|\boldsymbol{x}))^t$.

Let's first focus on the one-parameter case. When looking for a maximum using derivatives, remember that :

- ***Stationary points*** (i.e. any θ for which $\ell_n'(\theta|\boldsymbol{x}) = 0$) may be local or global minimizer of $\theta \mapsto \ell_n(\theta|\boldsymbol{x})$, local or global maximizer, or inflection points (i.e. points at which the concavity changes). Our goal is to find a ***global maximizer***.

- The ***second derivative test*** can be used to determine the nature of a stationary point (min/max/inflection). In fact, *if the second derivative at a stationary point is negative, than this point is a* ***local maximizer***. Furthermore, if $\ell_n''(\theta|\boldsymbol{x}) < 0$, $\forall\theta \in \Theta$, then ℓ_n is ***strictly concave***, and hence any stationary point that can be found will be the unique global maximizer.
- Maximum (or minimum) can occur where the derivative does not exist. $\rightarrow$ *Check all the points where the derivative dose note exists*.
- The zeros of the first derivative locate only extremum points *in the interior of the domain* of a function (here Θ). If the extremum occurs on the boundary, the first derivative may not be 0. $\rightarrow$ *Check the endpoints of* Θ.
- When all maxima candidates (if any) have been identified, the one(s) with the highest likelihood is/are the MLE.

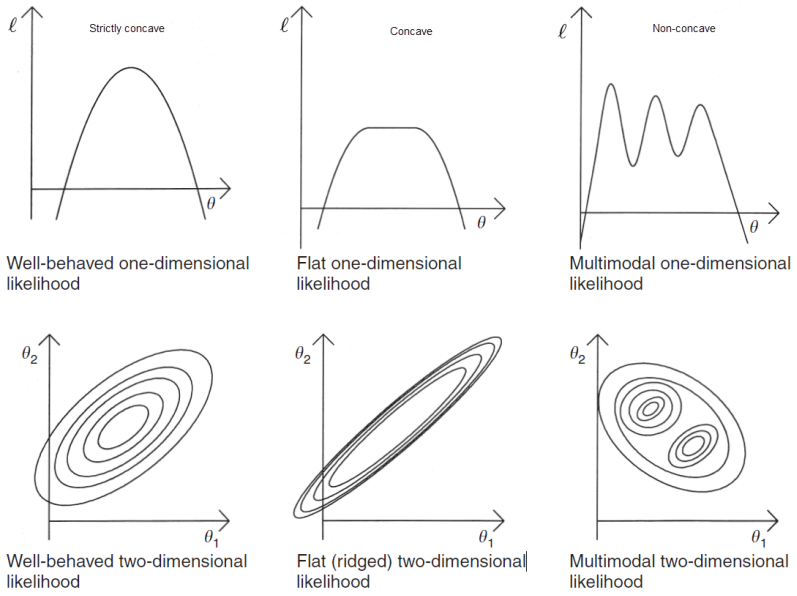
For most of the models considered here after, only one maximum exists, and it corresponds to the solution of the score equation.

Basically, the same approach can be applied for the multi-parameter case but the computational complexity increases with the number of parameters to be estimated. For example, in the case of a model with two parameters, i.e. $\boldsymbol{\theta} = (\theta_1, \theta_2)$, a stationary point, say $\boldsymbol{\theta}^*$, is a ***local maxima*** if *the Hessian matrix*

$$\nabla^2 \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) = \begin{pmatrix} \partial^2_{\theta_1} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) & \partial_{\theta_1\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) \\ \partial_{\theta_1\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) & \partial^2_{\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) \end{pmatrix}$$

is *negative definite*. This is the case if and only if (i) $\det\left(\nabla^2 \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x})\right) := \partial^2_{\theta_1} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) \partial^2_{\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) - \left(\partial_{\theta_1\theta_2} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x})\right)^2 > 0$, and (ii) $\partial^2_{\theta_1} \ell_n(\boldsymbol{\theta}^*|\boldsymbol{x}) < 0$.

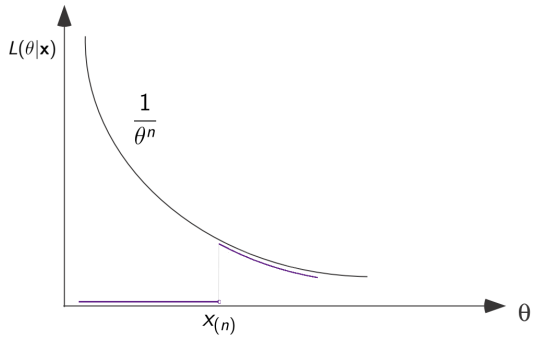
The figure below shows some log-likelihood functions and their extremum values/points in the case of one and two parameters.



Example 5.3.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Unif[0, \theta]$, $\theta > 0$.

$$L_n(\theta|\boldsymbol{x}) = \frac{1}{\theta^n} I(x_{(n)} \leq \theta).$$



  $\implies$ $X_{(n)}$ is the MLE of θ. We have seen that $X_{(n)}$ is biased, so it is not efficient.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Unif[\theta, \theta + 1]$, $\theta > 0$.

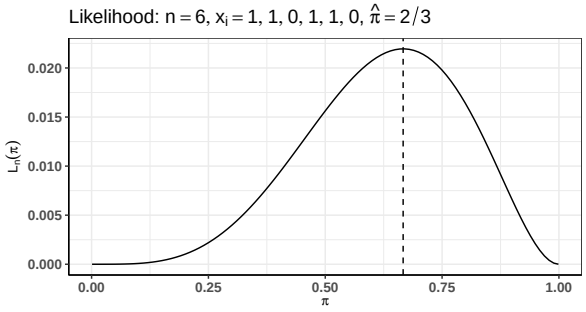
$$L_n(\theta|\boldsymbol{x}) = I(x_{(n)} - 1 \leq \theta \leq x_{(1)}).$$

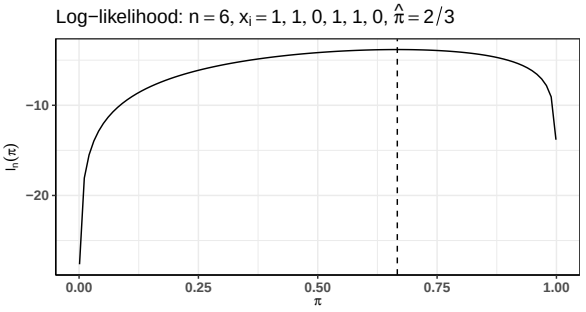
  So any $\hat{\theta}$ in $[X_{(n)} - 1, X_{(1)}]$ is a MLE estimator of θ.

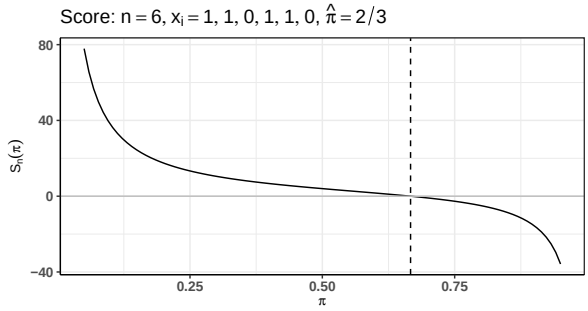
- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Ber(\pi)$, $\pi \in (0, 1)$.

$$\begin{aligned}
L_n(\pi|\boldsymbol{x}) &= \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i} \\
\ell_n(\pi|\boldsymbol{x}) &= \left(\sum_{i=1}^n x_i\right)\log(\pi) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\pi) \\
S_n(\pi, \boldsymbol{x}) &= \ell_n'(\pi|\boldsymbol{x}) = \frac{\sum_i x_i}{\pi} - \frac{n - \sum_i x_i}{1-\pi} \\
S_n'(\pi, \boldsymbol{x}) &= \ell_n''(\pi|\boldsymbol{x}) = -\left(\frac{\sum_i x_i}{\pi^2} + \frac{n - \sum_i x_i}{(1-\pi)^2}\right).
\end{aligned}$$

  $S_n = 0 \Leftrightarrow \hat{\pi} = \frac{1}{n}\sum_{i=1}^n x_i$, and ℓ_n is strictly concave ($\ell_n'' < 0$, $\forall \pi$), so $\hat{\pi} = \overline{X}_n$ is the MLE of π.







- Let X_i, $i = 1, \ldots, n$, be an iid sample from $N(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, and $\sigma^2 \in (0, \infty)$.

$$L_n(\mu, \sigma^2 | \boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$
$$\ell_n(\mu, \sigma^2 | \boldsymbol{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

  - Let's first consider the case of an unknown μ and a known σ^2. The score for μ and its derivative (with respect to μ) are

$$S_{1n} = \partial_\mu \ell_n(\mu, \sigma^2 | \boldsymbol{x}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$
$$\partial_\mu S_{1n} = \partial_\mu^2 \ell_n(\mu, \sigma^2 | \boldsymbol{x}) = -n/\sigma^2.$$

    As a function of μ, ℓ_n is strictly concave. Hence, setting the first derivative equal to zero (i.e. $S_{1n} = 0$) and solving for μ we get $\overline{X}_n$ as the MLE of μ.

  - Now, let's consider the case of a known μ and an unknown σ^2. The score for σ^2 and its derivative (with respect to σ^2) are

$$S_{2n} = \partial_{\sigma^2} \ell_n(\mu, \sigma^2 | \boldsymbol{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$$
$$\partial_{\sigma^2} S_{2n} = \partial_{\sigma^2}^2 \ell_n(\mu, \sigma^2 | \boldsymbol{x}) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2.$$

    As a function of σ^2, ℓ_n is not strictly concave, because $\partial_{\sigma^2}^2 \ell_n(\mu, \sigma^2 | \boldsymbol{x})$ is not negative for all possible parameter values. Now, by setting the first derivative equal to zero (i.e. $S_{2n} = 0$) and solving for σ^2 we obtain as the *unique* solution $\tilde{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} (x_i - \mu)^2$, and it is easy to see that

$$\partial_{\sigma^2}^2 \ell_n(\mu, \tilde{\sigma}_n^2 | \boldsymbol{x}) = -\frac{n}{2\tilde{\sigma}_n^4} < 0.$$

    We conclude that $\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is the MLE of σ^2.

  - Finally, let's consider the case of unknown $\boldsymbol{\theta} = (\mu, \sigma^2)$. The score (vector) and the

Hessian (matrix) are

$$S_n(\boldsymbol{\theta}, \boldsymbol{x}) = (S_{1n},\, S_{2n})$$
$$\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}|\boldsymbol{x}) = \begin{pmatrix} \partial_\mu S_{1n} & \partial_{\sigma^2} S_{1n} \\ \partial_\mu S_{2n} & \partial_{\sigma^2} S_{2n} \end{pmatrix} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_i (x_i - \mu) \\ \frac{1}{\sigma^4}\sum_i (x_i - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_i (x_i - \mu)^2 \end{pmatrix}.$$

We can see that $S_n(\hat{\boldsymbol{\theta}}, \boldsymbol{x}) = (0,\, 0) \Leftrightarrow \hat{\mu} = \overline{x}_n$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. So, $\hat{\boldsymbol{\theta}} = (\overline{x}_n,\, \hat{\sigma}^2_n)$ is the unique candidate for the MLE. Now,

$$\nabla^2_{\boldsymbol{\theta}} \ell_n(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) = -\begin{pmatrix} \frac{n}{\hat{\sigma}^2_n} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4_n} \end{pmatrix} \prec 0 \quad \text{(i.e. negative definite).}$$

We conclude that $\hat{\boldsymbol{\theta}} = \left(\overline{X}_n,\, n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right)$ is the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

- Let $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, be iid rve from the trinomial distribution with joint pd

$$f(\boldsymbol{x}; \boldsymbol{\pi}) = \frac{m!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

where $m \geq 1$, $\boldsymbol{x} = (x_1, x_2)$, and $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is unknown; with $x_1 = 0, \ldots, m$, $x_2 = 0, \ldots, m$, $x_3 = m - x_1 - x_2$, $\pi_1 \in (0,1)$, $\pi_2 \in (0,1)$, and $\pi_3 = 1 - \pi_1 - \pi_2$. We have that

$$L_n(\pi_1, \pi_2|\boldsymbol{x}) = \frac{(m!)^n}{\prod_{i=1}^n (x_{1i}! x_{2i}! x_{3i}!)} \pi_1^{\sum_{i=1}^n x_{1i}} \pi_2^{\sum_{i=1}^n x_{2i}} \pi_3^{\sum_{i=1}^n x_{3i}},$$
$$\ell_n(\pi_1, \pi_2|\boldsymbol{x}) = \sum_{i=1}^n x_{1i} \log(\pi_1) + \sum_{i=1}^n x_{2i} \log(\pi_2) + \sum_{i=1}^n x_{3i} \log(\pi_3) + const,$$
$$S_{1n} := \partial_{\pi_1} \ell_n(\pi_1, \pi_2|\boldsymbol{x}) = \frac{\sum_i x_{1i}}{\pi_1} - \frac{\sum_i x_{3i}}{\pi_3} \text{ and}$$
$$S_{2n} := \partial_{\pi_2} \ell_n(\pi_1, \pi_2|\boldsymbol{x}) = \frac{\sum_i x_{2i}}{\pi_2} - \frac{\sum_i x_{3i}}{\pi_3}.$$

The score equation is equivalent to $\sum_i x_{1i}\pi_3 = \sum_i x_{3i}\pi_1$ and $\sum_i x_{2i}\pi_3 = \sum_i x_{3i}\pi_2$. Summing these two equations yields to $nm\hat{\pi}_3 = \sum_i x_{3i}$. So the unique candidate for the MLE of (π_1, π_2) is $(\hat{\pi}_1, \hat{\pi}_2)$, with $\hat{\pi}_k = \sum_i x_{ki}/nm$. The Hessian of the log-likelihood is

$$\nabla^2_{\boldsymbol{\pi}} \ell_n(\boldsymbol{\pi}|\boldsymbol{x}) = -\begin{pmatrix} \frac{\sum_i x_{1i}}{\pi_1^2} + \frac{\sum_i x_{3i}}{\pi_3^2} & \frac{\sum_i x_{3i}}{\pi_3^2} \\ \frac{\sum_i x_{3i}}{\pi_3^2} & \frac{\sum_i x_{2i}}{\pi_2^2} + \frac{\sum_i x_{3i}}{\pi_3^2} \end{pmatrix}$$

This matrix is negative-definite, hence ℓ_n is strictly concave. We conclude that $(\hat{\pi}_1, \hat{\pi}_2) = \left(\frac{\overline{X}_{1n}}{m}, \frac{\overline{X}_{2n}}{m}\right)$ is the MLE of (π_1, π_2). □

Chapter 6

Finite and large sample properties of MLE

We now turn our attention to some appealing mathematical properties of the maximum likelihood estimators. We will start with key finite sample properties before discussing, in the next section, some asymptotic features.

6.1 Finite sample properties

6.1.1 Efficiency

Without loss of generality, we consider here the one parameter case. Suppose that an efficient estimator $\hat{\delta} \equiv \hat{\delta}(\boldsymbol{X})$ of θ exists. By the CRLB attainment theorem (assuming the required assumptions are met), $\hat{\delta}$ must satisfy the equation

$$S_n(\theta, \boldsymbol{X}) = I_n(\theta)\left(\hat{\delta} - \theta\right), \tag{6.1}$$

where $I_n(\theta) = Var\left(S_n^2(\theta, \boldsymbol{X})\right) = -E\left(\partial_\theta^2 \ell_n(\theta|\boldsymbol{x})\right)$ is the Fisher information, contained in the sample, about θ. Now, let $\hat{\theta}$ be the MLE of θ, then $\hat{\theta}$ satisfies $S_n(\hat{\theta}, \boldsymbol{X}) = 0$. Thus, provided that $I_n(\hat{\theta}) > 0$, the MLE $\hat{\theta}$ coincides with the efficient estimator $\hat{\delta}$.

Theorem 6.1 (MLE and efficiency)**.** *If an efficient estimator exists, then the maximum likelihood method of estimation will produce it.*

In other words, $\hat{\theta}$ is efficient $\implies \hat{\theta}$ is the MLE. The opposite of this statement is not true; as a counter-example, see the example above with $Unif[0, \theta]$.

6.1.2 Invariance to re-paramerization

Theorem 6.2. *A MLE is invariant with respect to any bijective transformation. That is, if $g : \Theta \longrightarrow \Lambda$ is bijective, then*

$$\hat{\theta} \text{ is a MLE of } \theta \Leftrightarrow g(\hat{\theta}) \text{ is a MLE of } g(\theta).$$

To see why this is the case, let $L_n(\theta)$ be the likelihood with the parametrization θ, i.e. $L_n(\theta) = \prod_i f(x_i, \theta)$, and $L_n^*(\eta)$ be the likelihood with the parametrization $\eta = g(\theta)$, i.e. $L_n^*(\eta) = \prod_i f^*(x_i, \eta)$, where $f^*(x, \eta) := f(x, g^{-1}(\eta))$. Put $\hat{\eta} = g(\hat{\theta})$, and let η be any point in Λ and $\theta = g^{-1}(\eta)$. Since $\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta)$, we have that,

$$L_n^*(\hat{\eta}) = L_n(g^{-1}(\hat{\eta})) = L_n(\hat{\theta}) \geq L_n(\theta) = L_n(g^{-1}(\eta)) = L_n^*(\eta).$$

We conclude that $L_n^*(\hat{\eta}) \geq L_n^*(\eta)$, $\forall \eta \in \Lambda$, and so $\hat{\eta}$ is the MLE of η.

Example 6.1. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Ber(\pi)$, $\pi \in (0, 1)$. We have seen that the likelihood function is given by $L_n(\pi|\boldsymbol{x}) = \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i}$, and the MLE of π is $\hat{\pi} = \overline{X}_n$.
Now, put $\eta = \log(\pi/(1-\pi))$; with this parametrization, the likelihood function becomes

$$L_n^*(\eta|\boldsymbol{x}) = \left(\frac{1}{1+e^{-\eta}}\right)^{\sum_{i=1}^n x_i} \left(\frac{1}{1+e^{\eta}}\right)^{n-\sum_{i=1}^n x_i}.$$

One can check, by taking the logarithm and then differentiate, that the maximizer of this likelihood is $\hat{\eta} = \log(\sum_i X_i/(n - \sum_i X_i))$. Hence, the MLE of $\eta = \log(\pi/(1-\pi))$ is $\hat{\eta} = \log(\hat{\pi}/(1-\hat{\pi}))$. The same result can be obtained directly (without any calculation) by simply applying the above theorem. □

This simple version of the invariance of MLE is not always useful because many of the functions we are interested in are not bijective. For example, in the case of $Ber(\pi)$, we can't apply the result above to get the MLE of $\pi(1-\pi)$.
Now, whether g is bijective or not, if $\hat{\theta}$ is the MLE of θ then it is reasonable to use $g(\hat{\theta})$ as an estimator for $g(\theta)$. *Even if g is not bejective, some authors refer to $g(\hat{\theta})$ as the MLE of $g(\theta)$*, although, according to the (classical) definition of MLE, we can't really call it so because we can't necessarily express the pd as a genuine function of $g(\theta)$.

Exercise 6.1. Let X_i, $i = 1, \ldots, n$, an iid sample from $N(\mu, \sigma^2)$. Propose a "good" estimator for the coefficient of variation $cv = \sigma/\mu$ ($\mu \neq 0$ and σ are unknown). Justify your choice.

6.2 Large sample properties

We will show that the MLE is ***often***: (1) consistent, (2) asymptotically normal, and (3) asymptotically efficient.

For (1) to hold, we need some regularity conditions:

- We observe X_i, $i = 1, \ldots, n$, an iid sample from a pd $f(x; \theta)$, $\theta \in \Theta$.
- The model is identifiable; that is, if $f(x; \theta) = f(x; \tilde{\theta})$ $\forall x$, then $\theta = \tilde{\theta}$.
- The model is correctly specified. *We denote by θ_0 is the true parameter value.*
- Θ is a compact set (closed and bounded) and $\theta \mapsto f(x; \theta)$ is continuous on Θ.
- $|f(X; \theta)| \leq d(X)$, $\forall \theta \in \Theta$, and $E_{\theta_0}(d(X)) < \infty$.

For (2)-(3) to hold, we need in addition to the above conditions the next assumptions:

- The support of f is independent of θ.
- θ_0 is in the interior of Θ.
- f is twice continuously differentiable in θ and $\int f(x,\theta)dx$ can be differentiated two times under the integral sign.
- The Fisher information satisfies $0 < I(\theta_0) < \infty$.
- In a neighborhood of θ_0, $\left|\partial^3_\theta \log f(x;\theta)\right| \leq M(x)\ \forall\theta$ and $E_{\theta_0}(M(X)) < \infty$.

These assumptions are sufficient (to prove consistency and asymptotic normality) but not necessary. More general and weaker conditions can be found in the literature.

6.2.1 Consistency

Theorem 6.3 (Consistency of MLEs)**.** *Under the regularity assumptions stated above,* $\hat{\theta}_n \xrightarrow{p} \theta_0$.

We will not prove this result here, but will only sketch out why this is happening.

First, by definition, the MLE $\hat{\theta}_n$ is the maximizer of $\bar{\ell}_n(\theta) = n^{-1}\sum_{i=1}^n \log f(X_i;\theta)$ which is the log-likelihood function normalized by $1/n$ (of course, this does not affect maximization). Second, by the law of large numbers (WLLN),

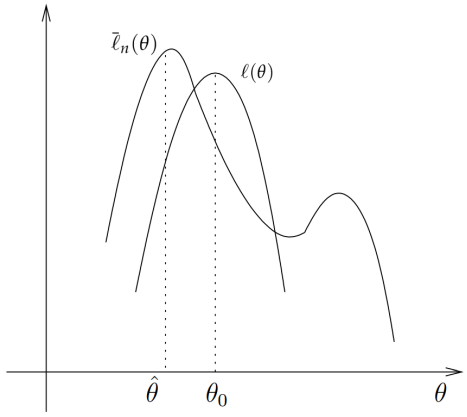
$$\bar{\ell}_n(\theta) \xrightarrow{p} \ell(\theta) := E_{\theta_0}(\log f(X;\theta)),\ \forall\theta.$$

The expectation operator E is indexed by θ_0 to explicitly point out that the expectation is evaluated using the true parameter θ_0, i.e. acknowledging that the pd of X is $f(x;\theta_0)$. Third, by Jensen's inequality, for any $\theta \neq \theta_0$,

$$\ell(\theta) - \ell(\theta_0) = E_{\theta_0}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) < \log E_{\theta_0}\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) = 0.$$

So, θ_0 is the maximizer of $\ell(\theta)$.

To sum up, we know that $\theta_0 = \arg\max \ell(\theta)$, $\hat{\theta}_n = \arg\max \bar{\ell}_n(\theta)$, and $\bar{\ell}_n(\theta) \xrightarrow{p} \ell(\theta)$, $\forall\theta$. So, we (intuitively) expect $\hat{\theta}_n$ to approach θ_0 as the sample size increases.



In fcat, the assumed hypothesis guarantees that this is indeed the case. Thus, $\hat{\theta}_n \xrightarrow{p} \theta_0$.

6.2.2 Asymptotic normality and asymptotic efficiency

Theorem 6.4 (Asymptotic normality of MLEs)**.** *Under the regularity assumptions stated above,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right),$$

where $I(\theta_0)$ is the Fisher information evaluated at θ_0.

The above result can be expressed as $\hat{\theta} \sim_a N(\theta_0, I_n^{-1}(\theta_0))$, where $I_n(\theta) = nI(\theta)$. And thus, *the MLE $\hat{\theta}$ is asymptotically efficient* for θ_0 (according to the definition given previously).

The proof uses Taylor's theorem, CLT, and Slutsky's theorem. In fact, Taylor expansion of $\theta \mapsto \bar{\ell}_n'(\theta)$ around θ_0 yields to $0 = \bar{\ell}_n'(\hat{\theta}) \approx \bar{\ell}_n'(\theta_0) + (\hat{\theta} - \theta_0)\bar{\ell}_n''(\theta_0)$. So,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\frac{1}{\bar{\ell}_n''(\theta_0)} \sqrt{n}\bar{\ell}_n'(\theta_0).$$

$\sqrt{n}\bar{\ell}_n'(\theta_0) = \sqrt{n}\big(n^{-1}\sum_i \partial_\theta \log f(X_i;\theta_0) - 0\big) \xrightarrow{d} N(0, I(\theta_0))$ by the CLT and the fact that $\partial_\theta \log f(X_i;\theta_0)$ has mean zero and variance $I(\theta_0)$. For the denominator, by WLLN, we have that $-\bar{\ell}_n''(\theta_0) = -n^{-1}\sum_i \partial^2_{\theta^2} \log f(X_i;\theta_0) \xrightarrow{p} I(\theta_0)$. The proof is completed by applying Slutsky's Theorem.

Attention: To simplify the notations, henceforth, we suppress the subscript 0 in θ_0, and write $\hat{\theta} \xrightarrow{p} \theta$, which must be understood as $\hat{\theta}$ converges to the true value of θ whatever this one is. In the same way, we write $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$.

The results stated above (consistency, asymptotic normality and asymptotic efficiency) can be extended to the multiparameter case.

If $\hat{\boldsymbol{\theta}}$ is $d-$dimensional MLE of $\boldsymbol{\theta}$, then, under some regularity assumptions (similar to those stated above),

- $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$,
- $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$, where $\boldsymbol{I}^{-1}$ is the inverse of the Fisher information matrix, and
- $\hat{\boldsymbol{\theta}}$ is asymptotically efficient for $\boldsymbol{\theta}$.

Example 6.2.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Bin(m, \pi)$, $\pi \in (0, 1)$. We have that

$$
\begin{aligned}
L_n(\pi|\boldsymbol{x}) &= \left(\prod_{i=1}^{n} C_m^{x_i}\right) \pi^{\sum_{i=1}^n x_i} (1-\pi)^{\sum_{i=1}^n (m-x_i)} \\
\ell_n(\pi|\boldsymbol{x}) &= \left(\sum_{i=1}^{n} x_i\right) \log(\pi) + \left(nm - \sum_{i=1}^{n} x_i\right) \log(1-\pi) + const, \\
S_n(\pi, \boldsymbol{x}) &= \partial_\pi \ell_n(\pi|\boldsymbol{x}) = \frac{\sum_i x_i}{\pi} - \frac{nm - \sum_i x_i}{1-\pi}, \\
\partial^2_\pi \ell_n(\pi|\boldsymbol{x}) &= \partial_\pi S_n(\pi, \boldsymbol{x}) = -\frac{\sum_i x_i}{\pi^2} - \frac{nm - \sum_i x_i}{(1-\pi)^2}.
\end{aligned}
$$

  Hence, the MLE of π is $\hat{\pi} = \dfrac{\sum_i X_i}{nm} = \dfrac{\overline{X}_n}{m}$. This later is consistent and asymptotically normal. More precisely

$$\hat{\pi} \sim_a N(\pi, I_n^{-1}(\pi)),$$

  where $I_n(\pi) := -E(\partial_{\pi^2} \ell_n(\pi|\boldsymbol{x})) = \dfrac{nm}{\pi(1-\pi)}$.

  Note that the asymptotic distribution of $\hat{\pi}$, as given above, can also be obtained by applying the CLT directly to $\overline{X}_n$.

- Let $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, be iid rve from the trinomial distribution with joint pd

$$f(\boldsymbol{x}; \boldsymbol{\pi}) = \frac{m!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

  We have seen the the MLE of (π_1, π_2) is $(\overline{X}_{1n}/m, \overline{X}_{2n}/m)$. We have also seen that

$$I^{-1}(\pi_1, \pi_2) = m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}.$$

  Hence,

$$\sqrt{n} \begin{pmatrix} \hat{\pi}_1 - \pi_1 \\ \hat{\pi}_2 - \pi_2 \end{pmatrix} \xrightarrow{d} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix} \right).$$

  Again, the same result can be obtained by applying the CLT directly on $\overline{\boldsymbol{X}}_{\boldsymbol{n}} = (\overline{X}_{1n}, \overline{X}_{2n})$.
  □

The theoretical results presented above can be extended to the situation where the parameter of interest is $\boldsymbol{g}(\boldsymbol{\theta})$ rather than $\boldsymbol{\theta}$. In fact, let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$. Then, *whether* $\boldsymbol{g} : \mathbb{R}^d \mapsto \mathbb{R}^p$ *is bijective or not*,

- By the continuous mapping theorem, $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{\theta})$;
- By the Delta method, $\sqrt{n}(\boldsymbol{g}(\hat{\boldsymbol{\theta}}) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{d} N_p\big(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{g}(\boldsymbol{\theta}))\big)$, where $\boldsymbol{I}^{-1}(\boldsymbol{g}(\boldsymbol{\theta})) = \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \dot{\boldsymbol{g}}^t(\boldsymbol{\theta})$;

- $g(\hat{\boldsymbol{\theta}})$ is asymptotically efficient for $g(\boldsymbol{\theta})$.

6.2.3 Observed Fisher information

We have seen that the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is asymptotically normal, i.e. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$. Equivalently, we can write that

$$\sqrt{\boldsymbol{I}_n(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}) \ ,$$

where $\mathbb{1}$ is the identity matrix and $\sqrt{\boldsymbol{I}_n(\boldsymbol{\theta})}$ is the *square-root matrix* of the FI

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -E\left\{\left[\partial_{\theta_j\theta_k}\ell_n(\boldsymbol{\theta})\right]_{j,k}\right\} = n\boldsymbol{I}(\boldsymbol{\theta}) \ , \text{with } \boldsymbol{I}(\boldsymbol{\theta}) = -E\left\{\left[\partial_{\theta_j\theta_k}\log f(X;\boldsymbol{\theta})\right]_{j,k}\right\}.$$

To use this asymptotic normality in practical inference, $\boldsymbol{I}(\boldsymbol{\theta})$ must be estimated. The most obvious estimator of $\boldsymbol{I}(\boldsymbol{\theta})$ is $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$. Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \boldsymbol{I}(\boldsymbol{\theta})$, provided that $\boldsymbol{\theta} \mapsto \boldsymbol{I}(\boldsymbol{\theta})$ is a continuous function. In this case, Slutsky's theorem ensures that

$$\sqrt{\boldsymbol{I}_n(\hat{\boldsymbol{\theta}})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}).$$

The disadvantage of this approach is that, in practice, it is not always easy to calculate $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ because of the difficulties of working out the expectations.

Let's define the matrix

$$\boldsymbol{J}_n(\boldsymbol{\theta}) := -\boldsymbol{\nabla}^2\ell_n(\boldsymbol{\theta}) = -\left[\partial_{\theta_j\theta_k}\ell_n(\boldsymbol{\theta})\right]_{j,k} = -\left[\sum_{i=1}^{n}\partial_{\theta_j\theta_k}\log f(X_i;\boldsymbol{\theta})\right]_{j,k}.$$

$\boldsymbol{J}_n(\boldsymbol{\theta})$ is the Hessian of the negative log-likelihood. This matrix is known as *the sample Fisher information* or *the observed Fisher information*. It can always be calculated as long as the second partial derivatives can be calculated. Observe that $\boldsymbol{I}_n(\boldsymbol{\theta}) = E\left(\boldsymbol{J}_n(\boldsymbol{\theta})\right)$. The law of large numbers guarantees that $n^{-1}\boldsymbol{J}_n(\boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{I}(\boldsymbol{\theta})$. So,

$$\sqrt{\boldsymbol{J}_n(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}).$$

Again Slutsky's theorem can be applied to show that

$$\sqrt{\boldsymbol{J}_n(\hat{\boldsymbol{\theta}})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}).$$

To sum up, ***for large sample sizes***, one can use $\boldsymbol{I}_n(\boldsymbol{\theta})$, $\boldsymbol{I}_n(\hat{\boldsymbol{\theta}})$, $\boldsymbol{J}_n(\boldsymbol{\theta})$ and $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}})$ interchangeably.

Example 6.3. Let X_i, $i = 1, \ldots, n$, be an iid sample from the pd

$$f(x;\theta) = \frac{1+\theta x}{2} I(-1 \le x \le 1);\ -1 < \theta < 1.$$

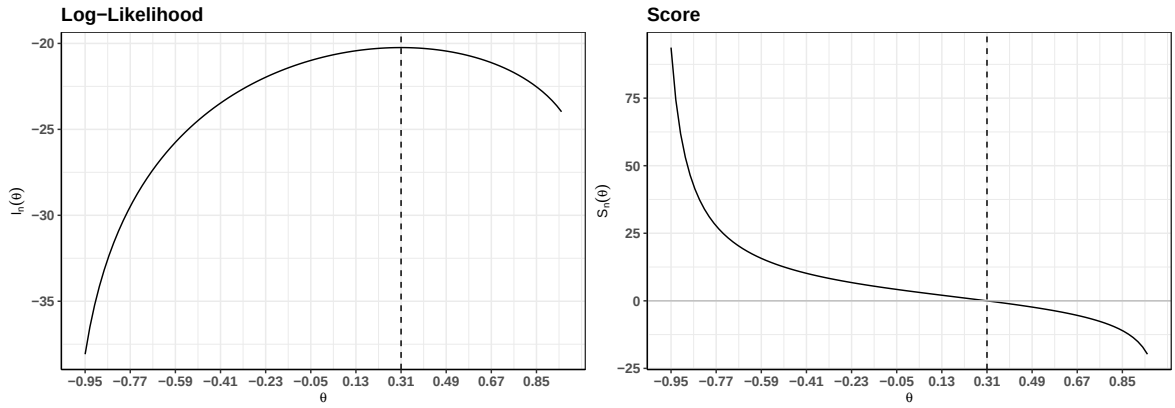
The log-likelihood, the Score and the observed FI are given by

$$\ell_n(\theta|x) = \sum_{i=1}^{n} \log(1+\theta x_i) - n\log(2),$$
$$S_n(\theta, x) = \partial_\theta \ell_n(\theta|x) = \sum_{i=1}^{n} \frac{x_i}{1+\theta x_i},$$
$$J_n(\theta) = -\partial_\theta S_n(\theta, x) = -\partial^2_\theta \ell_n(\theta|x) = \sum_{i=1}^{n} \frac{x_i^2}{(1+\theta x_i)^2}.$$

As a function of θ, the log-likelihood is *continuous and strictly concave* ($J_n > 0$), so there is a unique MLE, say $\hat{\theta}$, of θ. Now, although the likelihood equation of this model *cannot be solved explicitly* to get the analytic expression of $\hat{\theta}$, the theory tells us that $\sqrt{I_n(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$, where $I_n(\theta) = E(J_n(\theta))$.

The following data are simulated from the f above with $\theta = 0.5$ (hereafter, we'll pretend that we don't know the θ that generated the data and see how maximum likelihood behaves).

```
x <- c(0.9852, 0.0450, -0.6123, -0.7518, -0.2824, 0.7085, -0.0711, 0.9625,
       -0.4746, 0.1617, -0.4592, -0.3113, 0.6800, -0.6694, 0.1512, -0.7048,
       0.3421, -0.9658, 0.9809, -0.1205, 0.4730, -0.1665, 0.9956, 0.8720,
       0.9849, -0.7650, 0.4528, 0.2190, 0.9611, -0.0257)
```



By inspecting the graphs above, we can see that $\hat{\theta} = 0.31$ (we will see later how to obtain this result numerically).

As we discussed above, we can estimate $I_n(\theta)$ using $I_n(\hat{\theta})$ or $J_n(\hat{\theta})$. To use the former, we must first derive the expression of I_n, which can be done by using the polynomial division. In fact,

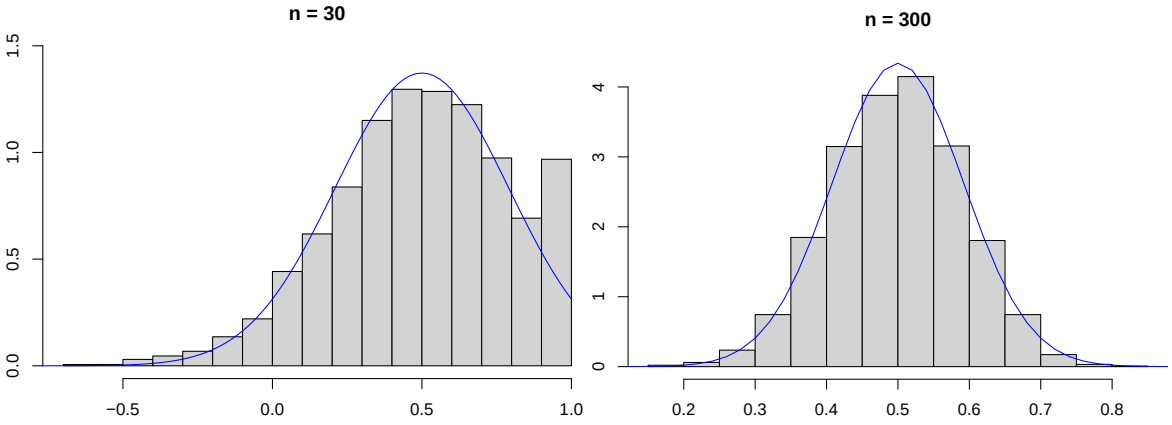
$$I_n(\theta) = nE\left(\frac{X^2}{(1+\theta X)^2}\right) = \frac{n}{2}\int_{-1}^{1}\left(\frac{1}{\theta}x - \frac{1}{\theta^2}\right)dx + \frac{n}{2\theta^2}\int_{-1}^{1}\frac{1}{1+\theta x}dx$$
$$= -\frac{n}{\theta^2} + \frac{n}{2\theta^3}\log\left(\frac{1+\theta}{1-\theta}\right).$$

In our case, $n = 30$, $\hat{\theta} = 0.31$, $I_{30}(\hat{\theta}) = 10.619$ and $1/I_{30}(\hat{\theta}) = 0.094$. Hence, based on the observed data, we conclude that $\hat{\theta}_{30} \sim_a N(\theta, 0.094)$.

The second method for estimating $I_n(\theta)$ is much simpler and consists in using the observed FI which is given here by $J_{30}(\hat{\theta}) = 11.846$. On this basis, we can write that $\hat{\theta}_{30} \sim_a N(\theta, 0.084)$.

Note that the true (but *unknown*) asymptotic variance of $\hat{\theta}_{30}$ is actually $1/I_{30}(0.5) = 0.085$.

To see how the asymptotic normal approximation works in practice, we repeat the data generation procedure 5000 times and calculate $\hat{\theta}$ each time. We did this for $n = 30$ and $n = 300$, respectively. The graph below shows the histogram of the simulated $\hat{\theta}$ and the curve of the true asymptotic normal density, i.e. $N(0.5, I_n^{-1}(0.5))$, in blue.



□

Chapter 7

More about likelihood

7.1 Numerical maximization of the likelihood

In several interesting cases, the maximization problem has no analytical solution. In other words, it is not possible to write $\hat{\theta}$ explicitly as a function of the data. In these cases, it is necessary to resort to numerical algorithms for the maximization of the likelihood.

7.1.1 The Newton-Raphson (NR) method

One of the most used method for optimization in statistics is the Newton-Raphson method. It is based on approximating the log-likelihood $\ell_n(\theta)$ by a quadratic function. For a given ***starting point*** $\theta_0 \in \Theta$, such that $\ell_n''(\theta_0) < 0$, define

$$\tilde{\ell}_n(\theta) = \ell_n(\theta_0) + (\theta - \theta_0)\ell_n'(\theta_0) + (\theta - \theta_0)^2\ell_n''(\theta_0)/2.$$

This is the second order Taylor serie approximation of $\ell_n(\theta)$ around θ_0.

The solution of the first-order condition for maximizing $\tilde{\ell}_n(\theta)$ is

$$\theta_1 = \theta_0 - \frac{\ell_n'(\theta_0)}{\ell_n''(\theta_0)}.$$

Since $\tilde{\ell}_n$ is an approximation of ℓ_n, θ_1, defined above, provides a guess value for the MLE.

We can try to improve the approximation by taking θ_1 as the new starting point and *keep repeating the process until convergence*. This suggests the following iterative procedure:

$$\theta_{k+1} = \theta_k - \frac{\ell_n'(\theta_k)}{\ell_n''(\theta_k)},\ k = 0, 1, \ldots$$

Equivalently, this algorithm can also be written as

$$\theta_{k+1} = \theta_k + \frac{S_n(\theta_k)}{J_n(\theta_k)},\ k = 0, 1, \ldots,$$

where $S_n(\theta) = \ell_n'(\theta)$ and $J_n(\theta) = -\ell_n''(\theta)$ are the score and the observed FI.

As we said, the procedure should be run until convergence, i.e. until there is no "significant" difference between θ_k and θ_{k+1}. No "significant" difference means that changes between consecutive iterations are less than a user-defined tolerance. For example, we may stop the algorithm whenever the difference $|\theta_{k+1} - \theta_k|$, or the relative difference $|\theta_{k+1} - \theta_k|/|\theta_k|$, is smaller than 10^{-8}.

Note that, $\theta_{k+1} = \theta_k$ is equivalent to $\ell_n'(\theta_k) = 0$. So, at the end of the process (i.e. when the iterations stop), the algorithm converges to (a neighborhood of) a stationary point of ℓ_n. This point could be a maximum point, a minimum point or even a inflection point. However, if $\ell_n''(\theta_k) < 0$, then the convergence point is a **local maximizer**. Moreover, if ℓ_n is strictly concave, then the convergence point is the (unique) global maximizer.

In all cases, when it converges, the NR algorithm reaches the stationary point closest to its starting point θ_0. *If several stationary points are present, the choice of the starting point becomes critical*. In many situations, the MoM can be used to obtain a reasonable starting point.

In the case where the likelihood is not strictly concave, it is recommended to:

- If possible, start by visually checking the graph of the (log-)likelihood function. Sometimes it's easy to see where an extremum occurs. But sometimes, local fluctuations of a relatively small scale can hide such points.
- Rerun the algorithm from different starting points, and then choose, among all the convergence points, the one that globally maximizes the likelihood.
- Perturb the convergence point by a "small" amount, and use this as a starting point for a new run of the algorithm. Then, see if it converges to a "better point", or "always" to the same one.

The arguments for deriving the NR algorithm for optimization in one dimension can be directly extended to multi-dimensional problems giving the multi-parameter NR method:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + J_n^{-1}(\boldsymbol{\theta}_k) S_n(\boldsymbol{\theta}_k), \ k = 0, 1, \ldots$$

where $S_n(\boldsymbol{\theta}) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta})$ is the Score vector and $J_n(\boldsymbol{\theta}) = -\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) = -\sum_{i=1}^n \boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta})$ is the observed FI matrix (i.e. Hessian of negative log-likelihood).

Let's consider our previous example with $f(x;\theta) = \frac{1+\theta x}{2} I(-1 \leq x \leq 1)$; $-1 < \theta < 1$, and the observed data

```
x <- c(0.9852, 0.0450, -0.6123, -0.7518, -0.2824, 0.7085, -0.0711, 0.9625,
       -0.4746, 0.1617, -0.4592, -0.3113, 0.6800, -0.6694, 0.1512, -0.7048,
       0.3421, -0.9658, 0.9809, -0.1205, 0.4730, -0.1665, 0.9956, 0.8720,
       0.9849, -0.7650, 0.4528, 0.2190, 0.9611, -0.0257)
```

The following code gives the R functions needed for running the NR algorithm.

```
LogLik <- function(theta, x) {sum(log((1 + theta * x) / 2))}
LogLikGrad <- function(theta, x) {sum(x / (1 + theta * x))}
LogLikHess <- function(theta, x) {-sum((x / (1 + theta * x))^2)}

NRoptim <- function(theta0, x, eps = 1e-06, trace = FALSE) {
  diff <- eps + 1
  theta <- theta0
  while (diff > eps) {
    theta.old <- theta
    Grad <- LogLikGrad(theta, x)
    Hess <- LogLikHess(theta, x)
    if(trace == TRUE) print(c(theta, LogLik(theta, x), Grad, Hess))
    theta <- theta.old - Grad / Hess
    diff <- abs(theta - theta.old)
  }
  c(Estimate = theta, Std.Error = sqrt(-1/Hess))
}
```

Let's run this function with starting point -0.3.

```
NRoptim(-0.3, x, trace = T)
```

```
[1]  -0.30 -22.48   7.93 -18.05
[1]   0.14 -20.41   1.96 -11.46
[1]   0.3103 -20.2413  -0.0124 -11.8473
[1]  3.09e-01 -2.02e+01 -3.66e-06 -1.18e+01

 Estimate Std.Error
    0.309     0.291
```

We see that the NR algorithm reaches its target after only 2-3 iterations. Here, since the log-likelihood function is strictly concave, there is no need for further investigation, and the point of convergence (0.309) is certainly the MLE.

The two figures below show how ℓ_n is approximated by $\tilde{\ell}_n$ and how the algorithm moves to the maximum.

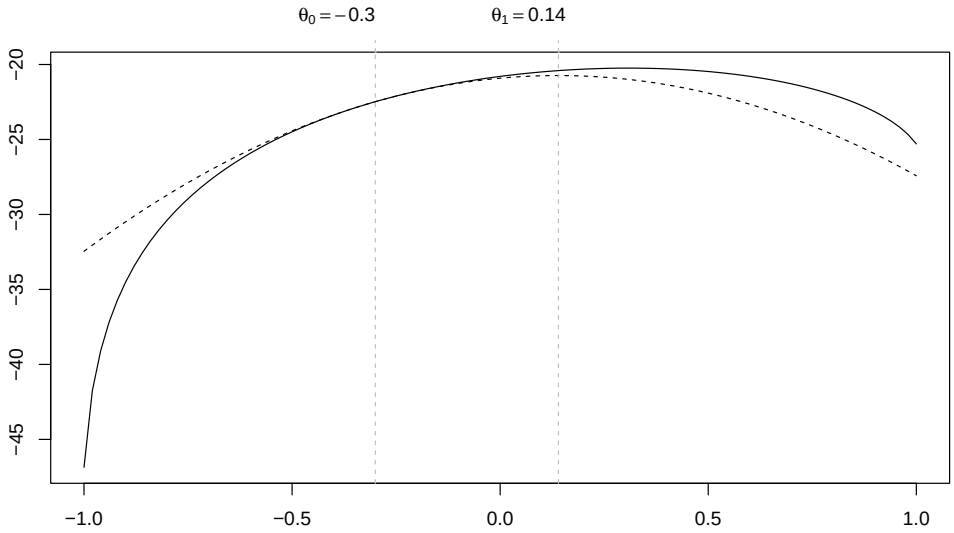


Figure 7.1: ℓ_n (solid line) and its quadratic approximation $\tilde{\ell}_n$ (dashed line) at $\theta_0 = -0.3$

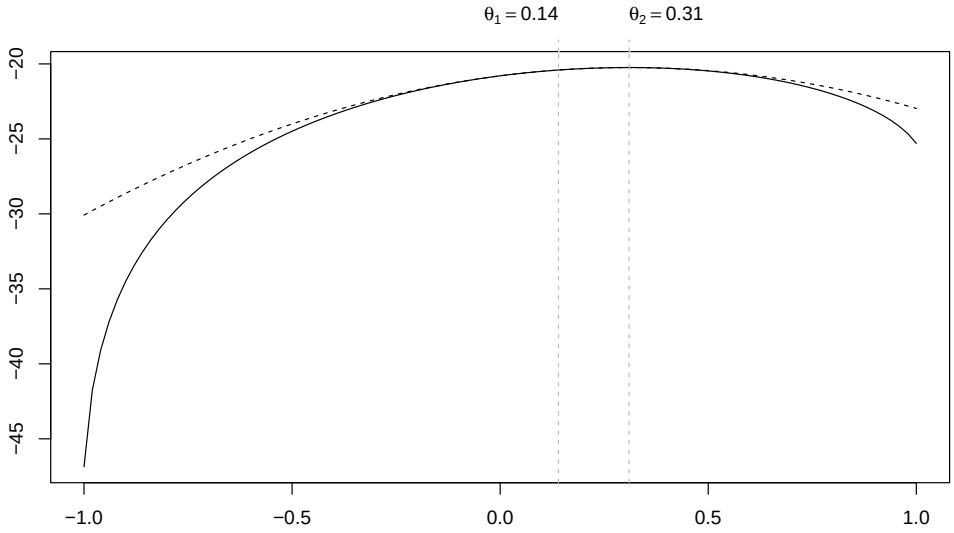


Figure 7.2: ℓ_n (solid line) and its quadratic approximation $\tilde{\ell}_n$ (dashed line) at $\theta_1 = 0.14$

As starting point, we could used the MoM estimator of θ. In fact, it is easy to check that $E(X) = \theta/3$, so the MoM of θ is $\hat{\theta}_0 = 3\bar{x}_n = 0.36$.

```
NRoptim(0.36, x, trace = T)
```

```
[1]   0.360 -20.257  -0.611 -12.255
[1]   0.3101 -20.2413  -0.0108 -11.8464
[1]  3.09e-01 -2.02e+01 -2.78e-06 -1.18e+01

 Estimate Std.Error
    0.309     0.291
```

When it works, the NR method converges very quickly to the maximum, especially when the starting point is not very far from the maximizer. However, this method has some serious

problems, especially if the likelihood is non-concave:

- NR is sensitive to the starting point.
- A NR step may jump far away from the target.
- As the number of parameters increases, the NR method becomes very computationally expensive.

A large number of enhanced alternative methods (as for example the Quasi-Newton methods) can be found in the literature, but this is beyond the scope of this course.

7.1.2 Maximum Likelihood in R

R provides a function called `optim()` which, by default, ***performs minimization***. To maximize the likelihood, provide `optim()` with *the negative of the log-likelihood* (for any function f, minimizing $(-f)$ maximizes f).

As main arguments, `optim()` takes: `par`, the starting vector point, i.e. the initial values for the parameters to be optimized over; `fn`: the function to be minimized, with first argument the vector of parameters over which minimization is to take place; and some other optional arguments.

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent"),
      lower = -Inf, upper = Inf, control = list(), hessian = FALSE)
```

The default algorithm (`method`) for `optim()` is a derivative-free optimization routine called the "Nelder-Mead" simplex algorithm. The other optimization methods are: BFGS, CG, L-BFGS-B, ... Here we will use the L-BFGS-B method. L-BFGS-B is a variant of BFGS method, an optimization algorithm in the family of quasi-Newton methods. L-BFGS-B is more memory-efficient than BFGS and allows the incorporation of "box" constraints, i.e. constraints of the form $a < \theta < b$; see the the Help of `optim()` for more details.

`optim()` returns a list of values of which `par`: the local minimizer, `value`: the target function evaluated at the solution found, `convergence`: an integer code indicating successful convergence (code 0) or a warning or an error code (see the Help), and `hessian`: the Hessian at the solution found (if `hessian = TRUE`).

Let's consider again our previous example with $f(x;\theta) = \frac{1+\theta x}{2} I(-1 \leq x \leq 1)$; $-1 < \theta < 1$, and the observed data

```
x <- c(0.9852, 0.0450, -0.6123, -0.7518, -0.2824, 0.7085, -0.0711, 0.9625,
       -0.4746, 0.1617, -0.4592, -0.3113, 0.6800, -0.6694, 0.1512, -0.7048,
       0.3421, -0.9658, 0.9809, -0.1205, 0.4730, -0.1665, 0.9956, 0.8720,
       0.9849, -0.7650, 0.4528, 0.2190, 0.9611, -0.0257)
```

```r
negLogLik <- function(theta, x) {-sum(log((1 + theta * x) / 2))}
optim(-0.3, fn = negLogLik, x = x, method = "L-BFGS-B", lower = -1, upper = 1, hessian = TRUE)
```

```
$par
[1] 0.309

$value
[1] 20.2

$counts
function gradient
       6        6

$convergence
[1] 0

$message
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"

$hessian
     [,1]
[1,] 11.8
```

Another option to do optimization in R is the function `mle()` from `stats4` base package. Internally, `mle()` calls `optim()` to perform the computations. But `mle()` does extra work to make the output more familiar/useful, and the `stats4` package comes with some very useful functions that facilitate the use of the maximum likelihood method in both estimation and inference.

```r
library(stats4)
mle.negLogLik <- mle(\(theta) negLogLik(theta, x = x), start = -0.3,
                     lower = -1, upper = 1, method = "L-BFGS-B")
mle.negLogLik
```

```
Call:
mle(minuslogl = function(theta) negLogLik(theta, x = x), start = -0.3,
    method = "L-BFGS-B", lower = -1, upper = 1)

Coefficients:
theta
0.309
```

The objective returned by the `mle()` function can be used by some useful well-known

generic `R` functions like `summary()`, `logLik()`, `vcov()`. This latter gives the asymptotic variance-covariance matrix of the estimated vector. The following example provides an illustration.

Example 7.1 (Two dimensional case)**.** Here we consider the problem of estimating $\boldsymbol{\theta} = (l, s)$ the location-scale parameters of the Cauchy distribution:

$$f(x; l, s) = \frac{1}{\pi s \left[1 + \left(\frac{x-l}{s}\right)^2\right]}, \; l \in \mathbb{R}, s > 0.$$

We start by generating 30 observations with true values $l = 0$ and $s = 1$.

```
set.seed(1)
y <- rcauchy(30)
y
```

```
 [1]   1.1025   2.3538  -4.2926  -0.2966   0.7346  -0.3305  -0.1756  -1.8082
 [9]  -2.3286   0.1966   0.7556   0.6195  -1.5015   2.6241  -0.8825 138.3476
[17]  -1.2274  -0.0254   2.5265  -0.8409  -0.2081   0.7865  -1.9374   0.4163
[25]   1.1145   2.6747   0.0421   2.5821  -0.4339   1.8237
```

```
neglogLi2 <- function(l, s, y) {-sum(dcauchy(y, location = l, scale = s, log = TRUE))}
mle.negLogLik2 <- mle(\(l, s) neglogLi2(l, s, y = y),
                      start = list(l = median(y), s = IQR(y)),
                      lower = c(-Inf, 0), upper = c(Inf, Inf),
                      method = "L-BFGS-B")
summary(mle.negLogLik2)
```

```
Maximum likelihood estimation

Call:
mle(minuslogl = function(l, s) neglogLi2(l, s, y = y), start = list(l = median(y),
    s = IQR(y)), method = "L-BFGS-B", lower = c(-Inf, 0), upper = c(Inf,
    Inf))

Coefficients:
  Estimate Std. Error
l    0.129      0.266
s    0.962      0.235

-2 log L: 141
```

```
# the maximum log-likelihood value,
# i.e. the log-likelihood function evaluated at the MLE
```

```
logLik(mle.negLogLik2)
```

```
'log Lik.' -70.6 (df=2)
```

```
# asymptotic variance-covariance matrix
vcov(mle.negLogLik2)
```

```
        l       s
l 0.07082 0.00425
s 0.00425 0.05520
```

This latter is nothing but the inverse of the negative of the Hessian matrix (of the log-likelihood) evaluated at the maximum likelihood.

```
solve(mle.negLogLik2@details$hessian)
```

```
        l       s
l 0.07082 0.00425
s 0.00425 0.05520
```

Remarks

- In the example above, the same solution can be found directly via `optim()` as follows

  ```
  neglogLi2 <- function(ls, y) {
    -sum(dcauchy(y, location = ls[1], scale = ls[2], log = TRUE))
  }
  optim(c(l = median(y), s = IQR(y)), fn = neglogLi2, y = y, method = "L-BFGS-B",
    lower = c(-Inf, 0), upper = c(Inf, Inf), hessian = TRUE)
  ```

- There are many packages and functions in `R` designed to facilitate MLE calculation, most of which call the `optim()` function, in the background, to perform the optimization. For example, the function `fitdistr()` of the package `MASS` can be used to fit a classical distribution to a given data. This function uses the following basic syntax:

  ```
  fitdistr(x, densefun, ...)
  ```

  where `x` is the observed data and `densefun` is the name of the distribution to be adjusted; see the Help of `fitdistr()` for the list of accepted distributions. Here is an example.

  ```
  Loading required package: MASS
  ```

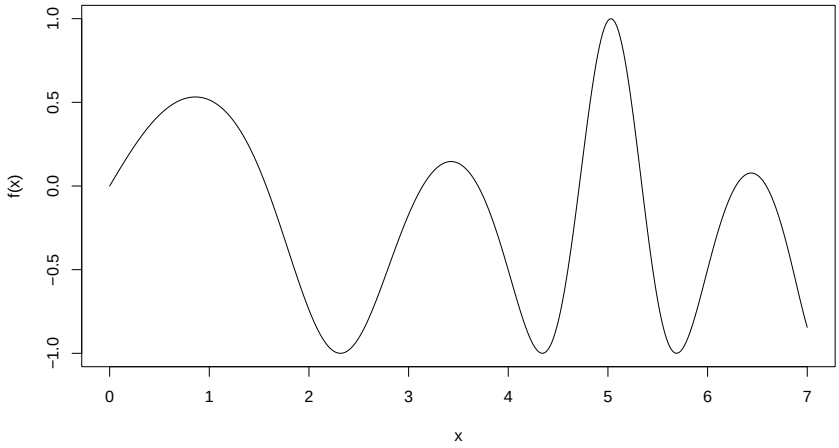
  ```
  fitdistr(y, densfun = "cauchy")
  ```

  ```
    location     scale
    0.129        0.962
   (0.266)     (0.235)
  ```

A similar but much more general package is `fitdistrplus`. See its online documentation for more information.

- Be careful when performing numerical optimization, as there is no guarantee that the resulting convergence point is actually the global maximizer. Here is an illustration.

```
f <- function(x) sin(x * cos(x))
curve(f, 0, 7, n = 400)
```




```
mle(\(x) -f(x), start = 2)@coef
[1] 0.86
mle(\(x) -f(x), start = 4)@coef
[1] 3.43
mle(\(x) -f(x), start = 5)@coef
[1] 5.03
mle(\(x) -f(x), start = 6)@coef
[1] 8.45
```

The problem becomes more and more difficult as the number of parameters increases.

7.2 Profile likelihood

Although the definition of likelihood covers the *multiparameter case*, the resulting multidimensional likelihood function can be difficult to deal with. Furthermore, in many practical multi-parameter problems, only a subset of the parameters is of interest; in the normal model, we might be interested only in the mean μ, while σ^2 is a "nuisance", which is there only to make the model correct. And even if we are interested in several parameters, it is always easier to describe one parameter at a time.

For a given model, let (θ, η) be the full parameter, where both θ and η are unknown. Let's say that θ is our primary parameter of interest and η is a nuisance parameter (θ and η may be vectors).

Definition 7.1 (Profile likelihood)**.** Given a model with (full) likelihood $L(\theta, \eta)$, *the profile*

likelihood for θ is

$$L_p(\theta) = max_\eta L(\theta, \eta) = L(\theta, \hat{\eta}(\theta)),$$

where $\hat{\eta}(\theta) = \arg\max_\eta L(\theta, \eta)$, i.e. $\hat{\eta}(\theta)$ is the maximizer of the function $\eta \mapsto L(\theta, \eta)$ when θ *is regarded as known*.

Let $\hat{\theta} = \arg\max_\theta L_p(\theta)$, a bit of logical deduction shows that $(\hat{\theta}, \hat{\eta}(\hat{\theta}))$ is the MLE of (θ, η). In fact,

$$L(\hat{\theta}, \hat{\eta}(\hat{\theta})) = L_p(\hat{\theta}) \geq L_p(\theta) = max_\eta L(\theta, \eta) \geq L(\theta, \eta).$$

This procedure is illustrated in the following example.

Example 7.2. Let X_i, $i = 1, \ldots, n$, be an iid sample from $N(\mu, \sigma^2)$, where $\mu \in (-\infty, \infty)$, and $\sigma^2 \in (0, \infty)$ are unknown. The log-likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

We know that if we consider μ as known and maximize this function with respect to σ^2 we get the following explicit expression for the MLE of σ^2

$$\hat{\sigma}^2(\mu) = n^{-1}\sum_{i=1}^{n}(x_i - \mu)^2.$$

So, the profile log-likelihood for μ is

$$\ell_p(\mu) = \ell(\mu, \hat{\sigma}^2(\mu)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right) - \frac{n}{2}.$$

Clearly, the $\arg\max_\mu \ell_p(\mu)$ is $\bar{x}_n$, and so the MLE of (μ, σ^2) is $\left(\overline{X}_n, n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)$. □

In practice, it is rarely possible to calculate a profile likelihood explicitly, but its numerical evaluation is often possible. In addition to reducing the size of the likelihood, profile likelihood functions can be used in the same way as (ordinary) likelihood functions not only to estimate the parameter(s) of interest, but also to obtain their asymptotic variance, which is equal to the inverse of the negative of the second derivative (the Hessian matrix) of the profile log-likelihood.

Exercise 7.1. Let X_i, $i = 1, \ldots, n$, be an iid sample from the gamma distribution with pd

$$f(x; \alpha, \sigma) = \frac{1}{\sigma^\alpha \Gamma(\alpha)} x^{\alpha - 1}\exp(-x/\sigma) I(x > 0),$$

where $\alpha > 0$ (shape), $\sigma > 0$ (scale) are unknown, and $\Gamma(\cdot)$ is the gamma function.

Show that the maximum likelihood estimator of (α, σ) is $(\hat{\alpha}, \hat{\alpha}^{-1}\overline{X}_n)$ where $\hat{\alpha}$ is the arg max

of

$$\ell_p(\alpha) = -n \log \Gamma(\alpha) - n\alpha \log\left(\alpha^{-1}\bar{X}_n\right) + (\alpha - 1)\sum_{i=1}^{n} \log(X_i) - n\alpha.$$

Use the above result to calculate the MLE of (α, σ) from the following simulated data. Estimate the asymptotic standard deviation of $\hat{\alpha}$.

```
set.seed(1)
x <- rgamma(n = 35, shape = 5, scale = 1/3)
x
```

```
 [1] 1.090 2.588 2.535 1.808 0.609 1.864 2.068 1.935 1.292 0.988 1.093 1.296
[13] 1.468 0.936 2.137 1.949 2.220 2.104 1.553 0.423 1.877 2.366 1.051 1.181
[25] 1.338 1.428 1.076 0.684 2.284 1.212 2.379 2.088 1.386 1.197 1.919
```

7.3 Likelihood for regression models

Consider the simple linear regression model with Gaussian noise, defined as $Y = \beta_0 + \beta_1 x + \epsilon$. ϵ is an unobservable noise/error variable with $N(0, \sigma^2)$ distribution. The aim is to estimate and make inference about $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ on the basis of an iid sample (Y_i, x_i) of (Y, x). For now, we assume that x_i are *known non-random* quantities (fixed design). According to this model $Y_i \equiv Y|X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, so the log-likelihood function is given by

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2.$$

It's easy to see that maximizing this function is equivalent to resolve the following optimization problems

$$(1)\ (\hat{\beta}_0, \hat{\beta}_0) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n}(x_i - (\beta_0 + \beta_1 x_i))^2,$$
$$(2)\ \hat{\sigma}^2 = \arg\min_{\sigma^2} \ell_n(\hat{\beta}_0, \hat{\beta}_1, \sigma^2).$$

Equation (1) above is identical to the optimization equation that defines linear least-squares (LS) regression. So the MLE of (β_0, β_1) is the same the least squares estimator. In fact, setting the derivatives to zero and solving produces

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \bar{x},\ \hat{\beta}_1 = \frac{\overline{xY} - \overline{x}\,\overline{Y}}{\overline{x^2} - \overline{x}^2},\ \text{and}\ \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

Maximum likelihood theory can of course be applied in this situation (Fisher information, asymptotic normality/efficiency, etc.).

For a random sampling design in which X_i are also sampled, we have that $Y_i|X_i \sim N(\beta_0 +$

$\beta_1 X_i, \sigma^2)$ and the (full) likelihood is

$$\prod_i f_{Y|X}(y_i|x_i;\boldsymbol{\theta}) f_X(x_i),$$

where $f_{Y|X}$ is the pd of $Y|X$, and f_X is the pd of X. If the latter does not depend on $\boldsymbol{\theta}$, as is commonly supposed, then f_X plays no role in the likelihood function. In this case, it is equivalent to work with the "full" log-likelihood or with the "conditional" log-likelihood as given by $\sum_i \log f_{Y|X}(y_i|x_i;\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta})$, i.e. the same as for fixed design. Consequently, the estimators and information matrix are the same as before, and the asymptotic results are the same also.

```
set.seed(5)
x <- 1:10
y <- 10 + 20 * x + rnorm(10, sd = 10)

neglogLiReg <- function(a, b, s, y) {-sum(dnorm(y, mean = a + b * x, sd = s, log = TRUE))}
mle(\(a, b, s) neglogLiReg(a, b, s, y = y),
    start = list(a = mean(y), b = 0, s = sd(y))) |> coef()
     a      b      s
10.445 19.775  9.011

lm(y ~ x) |> coef()
(Intercept)           x
      10.45       19.77
```

The same theory applies to multiple linear regression. Without loss of generality, let's consider the case of 2 covariates X_1 and X_2. The linear equation is $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$. The log-likelihood (under fixed or random design) is

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2,$$

where $\mu_i = \boldsymbol{\beta}^t \boldsymbol{x}_i$, $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \beta_2)$ and $\boldsymbol{x}_i^t = (x_{i0}, x_{i1}, x_{i2})$, with $x_{i0} = 1$, $\forall i$. Here again, the maximum likelihood method and the LS method lead to the same estimators. $\hat{\boldsymbol{\beta}}$ is the root of the system of equations

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \mu_i)^2 := \left(\partial_{\beta_1}\sum_{i=1}^{n}(y_i - \mu_i)^2, \partial_{\beta_2}\sum_{i=1}^{n}(y_i - \mu_i)^2, \partial_{\beta_3}\sum_{i=1}^{n}(y_i - \mu_i)^2\right) = \mathbf{0}.$$

Since, $\partial_{\beta_k}\sum_{i=1}^{n}(y_i - \mu_i)^2 = -2\sum_i (y_i - \mu_i)x_{ik}$,

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \mu_i)^2 = -2(\boldsymbol{X}^t\boldsymbol{y} - \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^t \\ \vdots \\ \boldsymbol{x}_n^t \end{bmatrix}$ and $\boldsymbol{y}^t = (y_1, \ldots, y_n)$. It fallows that $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^t\boldsymbol{X}^t\boldsymbol{Y}$. The maximum likelihood estimator of σ^2 has the same form as in the one-covariate case, i.e., the average squared residual $n^{-1}\sum_i(Y_i - \hat{\mu}_i)^2$, with $\hat{\mu}_i = \hat{\boldsymbol{\beta}}^t\boldsymbol{x}_i$.

Linear regression is the simplest regression model that can be estimated and inferred using likelihood theory. This theory can be applied to a very wide range of other regression setting such as, for example, non-linear regression models, generalized linear models (based on the exponential family), generalized linear mixed models, censored regression models, to name only a few.

Chapter 8

Hypothesis testing

There are some problems we meet in statistical practice in which estimation of a parameter is not the primary goal of the analysis; rather, we wish to use our data to decide to trust or not a particular claim about a given population. For example, a drug company may claim that its new drug reduces "bad" cholesterol (or LDL cholesterol) level by more than 20 points (after a certain period of treatment). To try to prove this, the drug company may design a clinical trial and collect data on the level of LDL reduction in selected subjects. The observed data can be analyzed to confirm (or refute) the manufacturer's claim. Properly formulated, such inference problems are called *testing of hypothesis* problems or simply test problems.

8.1 Basic concepts

Before attempting any inferential procedure, the model and all its underlying assumptions must clearly be defined. This is referred to the ***maintained hypothesis***. These hypothesis define the framework within which the inferential procedure can be applied and interpreted correctly. In the context of the parametric hypothesis testing problem, the maintained hypothesis are simply the assumption made on the distribution of the observed data. In simple situations, this reduces to assuming that we observe an iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a pd $f(x;\theta)$, for some $\theta \in \Theta$ (real or a vector parameter space).

Based on the observed data, we have to decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, where Θ_0, Θ_1 are a partition of the parameter space Θ, i.e. $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. We formulate this in the following manner:

$$H_0 : \ \theta \in \Theta_0 \ \ vs \ \ H_1 : \ \theta \in \Theta_1.$$

The hypothesis H_0 is called the ***null hypothesis*** and H_1 is called the ***alternative hypothesis***. In a statistical test procedure, these two hypothesis play an asymmetric role:

- H_0 *represents the* ***current theory***: "null" means statu quo, no change or no effect. Typically, H_0 is also simpler (lower-dimensional) than the alternative hypothesis.

- H_1 *is referred as the* ***researcher's hypothesis***: it is the new claim that one would really like to validate or the question to be answered.

Example 8.1. Take the cholesterol example. Suppose that, in a clinical trial, the observed LDL *decrease* of n subjects are $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 25)$, for some $\mu \in (-\infty, \infty)$. Here, iid and normality are the maintained hypothesis. The drug company's claim is $\mu > 20$. This claim can be formulated as a test problem with two hypothesis: the null hypothesis $H_0 : \mu \leq 20$, and the alternative hypothesis $H_1 : \mu > 20$. □

A hypothesis is called ***simple*** if it completely specifies the underlying distribution; otherwise it is called ***composite***. In our example above, we are testing a composite null against a composite alternative.

In addition, a test is called ***one-tailed*** if the alternative hypothesis is articulated directionally ($H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$, for some given θ_0); otherwise it is called ***two-tailed*** ($H_1 : \theta \neq \theta_0$). In our example above, we have a one-tailed test.

We will use our data to choose between the two hypothesis H_0 and H_1. For that, the basic idea is to try to *"compare" the observed sample with the model under* H_0 *and under* H_1. The comparison is typically based on a statistic $T \equiv T_n(\boldsymbol{X})$, called the ***test statistic***, which is designed in order to measure the discrepancy between the data and the models under H_0 and under H_1. Most of the time, a test statistic is taken to be a *sufficient statistic* for the parameter of interest θ, whose distribution (under H_0) is known exactly or asymptotically.

Assume that an appropriate test statistic T has been chosen and, without loss of generality, *suppose that small values of* T *support* H_0*, while large values support* H_1. The next step is to select a number, called the ***critical value***, k and apply the following rule:

> if $T_n(\boldsymbol{x}) \geq k$, reject H_0 in favor of H_1; otherwise, do not reject H_0.

In this way, our *test is nothing but the statistic/function* $\varphi_n(\boldsymbol{X}) = I(T_n(\boldsymbol{X}) \geq k) : \mathbb{R}^n \mapsto \{0, 1\}$, called the ***test function*** or simply test, that takes only values 0 and 1:

- $\varphi_n(\boldsymbol{x}) = 1 \Leftrightarrow$ reject H_0 in favor of H_1.
- $\varphi_n(\boldsymbol{x}) = 0 \Leftrightarrow$ do not reject H_0.

The subset $\mathcal{R}_n = \{\boldsymbol{x} : T_n(\boldsymbol{x}) \geq k\}$ of the sample space for which H_0 will be rejected is called the ***rejection region or critical region***. $\mathcal{R}_n^c$, the complement of the rejection region, is called the ***non rejection region***.

Remark The tests used in most practical situations can be expressed as $I(T_n \leq k)$ or $I(T_n \geq k)$. We use $I(T_n \leq k)$ in situations where T_n tends to be small under H_1 and $I(T_n \geq k)$ in situations where T_n tends to be large under H_1. In these notes, we will focus on the last case, but all the methodology can be applied directly to the first situation by using $-T_n$ as the actual test statistic.

Example 8.2. In the cholesterol example, our hypothesis are $H_0 : \mu \leq 20$ vs $H_1 : \mu > 20$.

As a consistent estimator of μ, $\overline{X}_n$ tends to be larger under H_1 than under H_0. Hence, it is natural to reject H_0 for large values of $\overline{X}_n$ and so to consider, for example, the test function $\varphi_1 = I(\overline{X}_n \geq 21)$. The same reasoning can be applied to the sample median, say M_n, which leads to the test function $\varphi_2 = I(M_n \geq 21)$. The first test φ_1 uses $\overline{X}_n$ as test statistic, while the second test φ_2 uses M_n as test statistic. The rejection region of φ_1 is $\{(x_1 \ldots, x_n) : \bar{x}_n \geq 21\}$ and that of φ_2 is $\{(x_1 \ldots, x_n) : m_n \geq 21\}$. □

This example raises a couple of questions:

- Why 21 is used as the critical value? How to choose a suitable critical value?
- How to choose between φ_1 and φ_2 ? Is there an "optimal" test? If so, how to find it?

8.2 Evaluating a test

In a situation where the analyst has to decide between H_0 and H_1, and given that in reality one of these hypothesis is true and the other is false, there are four possible states summarized in the following table:

Decision/Reality	H_0 is true	H_1 is true
Do not reject H_0	Correct	Error of type II
Reject H_0	Error of type I	Correct

For a given test φ_n, the function defined on Θ by

$$\pi_n(\theta) = E_\theta(\varphi_n(\boldsymbol{X})) = P_\theta(\varphi_n(\boldsymbol{X}) = 1) = P_\theta(\text{Reject } H_0)$$

is called the ***power function*** of the test φ_n. This function indicates the probability of rejecting H_0 for every possible value of θ. It plays an important role in evaluating the quality of a test and in comparing different test procedures.

Related to the power function we define:

- ***Probability of type I error***. This is the probability that the test rejects a true null hypothesis; i.e. $P_\theta(\text{Reject } H_0|H_0)$ or, equivalently, $\pi_n(\theta), \theta \in \Theta_0$.
- The ***size*** is the largest probability of committing a type I error, i.e. $\max_{\theta\in\Theta_0} \pi_n(\theta)$. If the size of a test is known not to exceed a given $\alpha \in (0,1)$, i.e. $\max_{\theta\in\Theta_0} \pi_n(\theta) \leq \alpha$, then we say that the test is of (significance) ***level*** α.
- ***Power***. This is the probability that the test rejects a false null hypothesis; i.e. $P_\theta(\text{Reject } H_0|H_1)$ or, equivalently, $\pi_n(\theta), \theta \in \Theta_1$.
- ***Probability of type II error***. This is the probability that the test does not reject a false null hypothesis; i.e. $P_\theta(\text{Do not reject } H_0|H_1) = 1 - \text{Power}$ or, equivalently, $1 - \pi_n(\theta), \theta \in \Theta_1$.

Example 8.3. Recall our cholesterol example, with $H_0 : \mu \leq 20$ vs $H_1 : \mu > 20$. Assuming that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 25)$, the power function of the test $\varphi_n = I(\overline{X}_n \geq 21)$ is

$$\pi_n(\mu) = P_\mu(\overline{X}_n \geq 21) = 1 - \Phi\left(\sqrt{n}\frac{21-\mu}{5}\right),$$

where Φ is the cdf of $N(0,1)$.

As this is an increasing function in μ, the size of our test φ_n is $\max_{\mu \leq 20} \pi_n(\mu) = \pi_n(20) = 1 - \Phi(\sqrt{n}/5)$. For example, with $n = 10$, the size is 0.2635 and with $n = 100$ the size is 0.0228.

Here is the plot of $\pi_n(\mu)$, for different values of the sample size n.

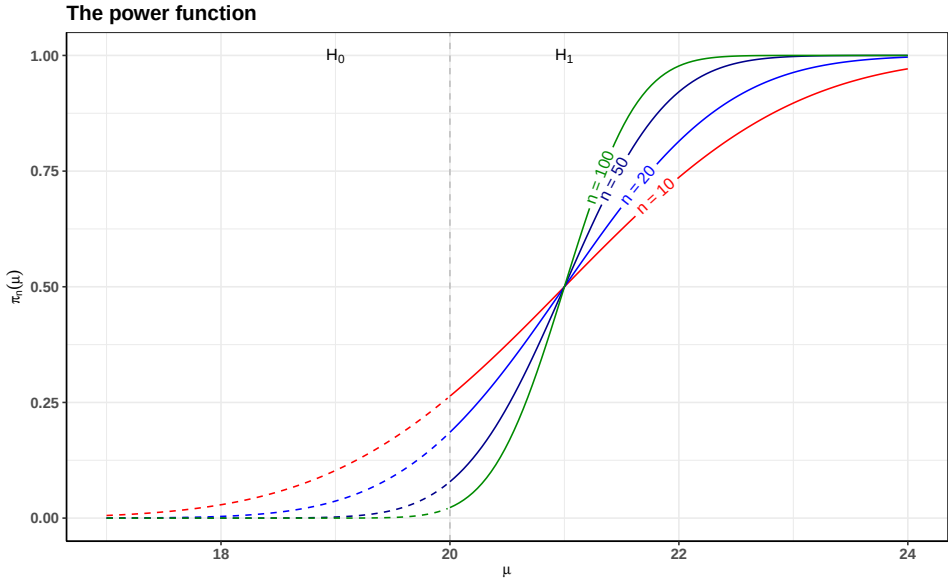


Figure 8.1: Power (solid line) and Type-I-error probability (dashed line)

For n fixed, as the true μ moves away from 20, the power increases to 1 and the probability of type I error decreases to 0. On the other hand, for fixed μ, as n increases, the power increases (for μ larger than 21) and the probability of type I error decreases. □

Ideally, we would like to have a statistical test with the lowest probability of Type I error (0) and the highest power (1). For a fixed sample size, *it is usually impossible to control both Type I error and power*. In fact, to minimize the probability of type I error, one must not reject H_0 more often and to maximize the power, one must reject H_0 more often: these two goals work against each other. Typically, even in very simple problems, larger power comes at the expense of a larger probability of Type I error. Vice versa, when one tries to reduce the probability of type I error, the power also gets reduced. All that can be done is to control one of these two quantities by adjusting the critical value of the test. Here is an example.

Example 8.4. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\pi)$, where $\pi \in \{0.3, 0.5\}$. Consider testing

$$H_0: \ \pi = 0.5 \ \ vs \ \ H_1: \ \pi = 0.3.$$

Put $T_n = \sum_{i=1}^n X_i$. It is reasonable to reject the null hypothesis if the total number of successes T_n is "too small". Let's define the test $\varphi_n = I(T_n \leq k)$, where k is the critical value that needs to be fixed. The power function of φ_n is given by

$$\pi_n(\pi) = P(T_n \leq k) = \sum_{\ell=0}^{k} C_n^\ell \pi^\ell (1-\pi)^{(n-\ell)}.$$

For example, with $n = 15$ we get (using the `R` function `pbinom(k, size = n, prob =`π`)`)

	Type-I-Error	Power
k	$P(T_n \leq k \mid \pi = 0.5)$	$P(T_n \leq k \mid \pi = 0.3)$
1	4.88×10^{-4}	0.03
4	0.06	0.52
7	0.50	0.95
10	0.94	≈ 1

For example, we can see that the test $\varphi_n = I(T_n \leq 4)$ gives a low type-I-error probability (0.06), but gives also a low power (0.52) and therefore does not provide adequate protection against a type-II-error (probability 0.48). We can increase the power to 0.95 by taking $\varphi_n = I(T_n \leq 7)$. However, by doing this, we increase the type-I-error probability to 0.5.

This demonstrates the conjunction of type-I-error probability and power of a test. The only way to decrease the probability of type I error and to increase the power simultaneously is to increase the sample size and to choose an appropriate critical value. For example, with $n = 150$ and $k = 60$, we get

k	Type-I-Error	Power
60	0.009	0.996

□

8.3 Testing strategy: the Neyman-Pearson approach

Given that it is not possible to reduce simultaneously the probabilities of both types of errors, Neyman and Pearson suggested to hold one of the two error probabilities at a pre-specified level that can be considered tolerable, and then minimize the other error probability. More specifically, Neyman-Pearson framework of statistical hypothesis is to

(1) choose a small number $\alpha \in (0, 1)$ ($\alpha = 0.01, 0.05$, and 0.1 are commonly used in practice). Restrict attention to tests with the pre-chosen level α only,

(2) *among these tests*, try to find a test that has the largest power. When it exists, such a test is called the ***uniformly most powerful test*** (UMP) of level α. More precisely, a test φ^*

of level α is UMP if $\forall \varphi$ of level α, $\pi_n^{\varphi^*}(\theta) \geq \pi_n^{\varphi}(\theta)$, $\forall \theta \in \Theta_1$, $\forall n$.

Since reducing type I error also reduces power, to obtain a UMP test (with a pre-chosen α level), one must necessarily stick to tests with the highest possible type I error (not exceeding α).

To apply the Neyman-Pearson approach, two questions must be answered:

(1) how to control the level of a given test and fix it at the pre-chosen α?
(2) how to find an UMP test (*when it exists*)?

We will defer this last question to Section 8.5 and discuss here the first one. Before, observe that the Neyman-Pearson approach makes H_0 **and** H_1 **asymmetric**: we control the type-I-error probability to be at most α; then, we try to make the type-II-error probability *as small as possible*. However, this latter can still be very big even for an UMP test. If we take this approach, the test must be set up so that H_1, the alternative hypothesis, is the one we seek to prove (this is why we refer to it as the *researcher's hypothesis*). By using an α level test, with small α, we guard against saying that the data support the research hypothesis (H_1) when it is false.

Now regarding our first question, without lost of generality and to make things clearer, let's consider the case where our test is given by $\varphi_n = I(T_n \geq k)$, i.e. large values of T_n give evidence against H_0. To fix the level of φ_n to α, we must choose an appropriate critical value k. More precisely, we must choose *the smallest* $k \equiv k(n, \alpha)$ such that

$$\max_{\theta \in \Theta_0} P_\theta(T_n \geq k) \leq \alpha.$$

Finding such a k is greatly simplified if ***the null distribution*** *of the test statistic T_n (i.e. the distribution of T_n under H_0) is known*; otherwise, k can be obtained by *Monte Carlo methods* (not discussed here).

Once the critical value k is chosen, the next step is to calculate $t_n = T_n(\mathbf{x})$, the observed value of T_n, and then to take one of the following decisions:

- If $t_n \geq k$, then, *at level* α, the data **reject** H_0 in favor of H_1 because the event $\{T_n \geq k\}$ is quite rare (it occurs with probability at most α) under H_0. We say that the ***test is significant***.
- If $t_n < k$, then, *at level* α, the data **do not reject** H_0. In this case, either
  - the power is known to be "very large" $\Longrightarrow$ **accept** H_0, i.e reject H_1 in favor of H_0
  - or the power is unknown or known to be "small". In this case, accepting H_0 is a risky decision $\Longrightarrow$ we don't have enough information (i.e. sufficient sample size) to accept or reject H_0. In this case, we say that the ***test is non significant***.

Example 8.5. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known. For some given μ_0, consider

testing

$$H_0 : \ \mu \leq \mu_0 \ \ vs \ H_1 : \ \mu > \mu_0.$$

We have seen that it is natural to reject H_0 when $\overline{X}_n$ is "large". For ease of calculation, it is convenient to replace $\overline{X}_n$ by the test statistic $Z_n = \sqrt{n}\dfrac{\overline{X}_n - \mu_0}{\sigma}$. So we consider the test $\varphi_n = I(Z_n \geq k)$. All we need now is to choose k so that the level of this test is α. We have that

$$\begin{aligned} & P_\mu(Z_n \geq k) = P\left(Z \geq k + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right), \text{ with } Z \sim N(0,1) \\ \Rightarrow & \max_{\mu \leq \mu_0} P_\mu(Z_n \geq k) = P(Z \geq k). \end{aligned}$$

So the smallest k for which $\max_{\mu \leq \mu_0} P_\mu(Z_n \geq k) \leq \alpha$ is obtained by setting $P(Z \geq k) = \alpha$, or $k = z_{1-\alpha}$, i.e. the $(1-\alpha)-$quantile of $N(0,1)$ (in `R` `qnorm(1-α)`). To conclude, the test

$$\varphi_n = I(Z_n \geq z_{1-\alpha}),$$

or equivalently $\varphi_n = I(\overline{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha})$, is of size α.
The power function of this test is given by

$$\pi_n(\mu) = P\left(Z \geq z_{1-\alpha} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right) = 1 - \Phi\left(z_{1-\alpha} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right).$$

Observe that $\pi_n(\mu) \xrightarrow[n\to\infty]{} 1$, $\forall \mu > \mu_0$, such a test is said to be ***consistent***. More precisely, *a test of level α is consistent if its power converges to* 1 *as n tends to infinity*. This property is often considered necessary for a test to be useful in practice.

Here is the plot of $\pi_n(\mu)$, for $\mu_0 = 20$, $\sigma = 5$ and $\alpha = 5\%$, for different values of the sample size n.

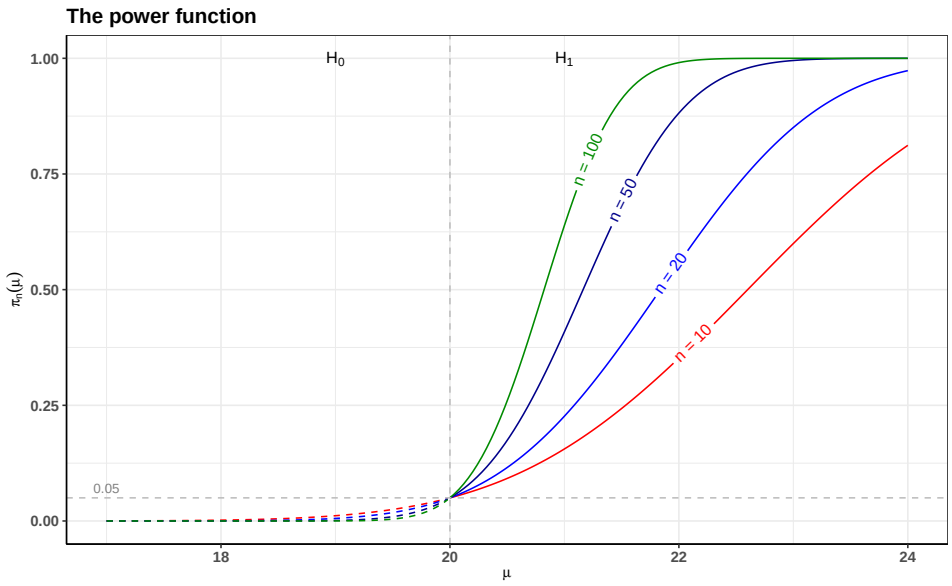


Figure 8.2: Power (solid line) and Type-I-error probability (dashed line)

□

8.4 The $p-$Value

The classic approach described above, based on critical values, can be summarized in two main steps:

1. specify the significance level α,
2. calculate the test statistic and compare it to the critical value, then reject/not reject H_0.

Simply reporting that a given hypothesis is rejected is not very informative. This says nothing about the fact that the computed value of the test statistic just barely fell into the rejection region or whether it exceeded the critical value by a large amount. What we need is a relevant way to ***measure the strength of evidence*** against a null hypothesis. And this is exactly what a $p-$value does.

To make things clearer, let's consider the test we studied in the Example 8.5 above. Recall that our decision is to reject H_0 whenever z_n, the observed value of the test statistic Z_n, is greater than $z_{1-\alpha}$. Since $z \mapsto P_{\mu_0}(Z_n \geq z)$ is a decreasing function, we have that

$$\begin{aligned} z_n \geq z_{1-\alpha} &\Leftrightarrow P_{\mu_0}(Z_n \geq z_n) \leq P_{\mu_0}(Z_n \geq z_{1-\alpha}) \\ &\Leftrightarrow P(Z \geq z_n) \leq \alpha, \ Z \sim N(0,1). \end{aligned}$$

The quantity $P(Z \geq z_n)$, that we can write as $P(Z_n \geq z_n | H_0)$, is called the $p-$value. From the above calculation, we can reformulate our test decision as follows:

- $p-$value $\leq \alpha \ \Rightarrow$ reject H_0 at level α.

- $p-$value $> \alpha \Rightarrow$ do not reject H_0 at level α.

A more general definition of the $p-$value can be stated as follows. *Let $\boldsymbol{T} \equiv T(\boldsymbol{X})$ be a test statistic such that large values of T give evidence against H_0.* Given $t = T(\boldsymbol{x})$, the observed sample value of T, the $p-$value of such a test is

$$\max_{\theta \in \Theta_0} P_\theta(T \geq t).$$

In the particular case where $\Theta_0 = \{\theta_0\}$, the above $p-$value reduces to $P_{\theta_0}(T \geq t) \equiv P(T \geq t|H_0)$ which can be seen as *the probability of observing a test statistic at least as "extreme" (i.e. at the opposite of H_0 but along the direction of H_1) as the one actually observed, assuming the null hypothesis H_0 is true*. This is a simple and an intuitive definition of the $p-$value which is used in many textbooks. *The smaller the $p-$value, the more evidence there is in the observed data against the null hypothesis and in favor of the alternative hypothesis*. Note that, unlike the critical value, the calculation of the $p-$value does not involve the level of the test α.

Example 8.6. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known. For some given μ_0, consider testing

$$H_0 : \ \mu = \mu_0 \ \ vs \ \ H_1 : \ \mu \neq \mu_0.$$

A natural test function is given by $\varphi_n = I(|Z_n| \geq k)$, where $Z_n = \sqrt{n}\dfrac{\overline{X}_n - \mu_0}{\sigma}$. The test statistic here is $|Z_n|$.
For this test to be of size α, we need to choose k so that $P_{\mu_0}(|Z_n| \geq k) = \alpha$. Thus, $k = z_{1-\alpha/2}$. So, based on this critical value, one should reject H_0, at level α, if $|z_n| \geq z_{1-\alpha/2}$.
An alternative and more convenient approach to do this test is to calculate its $p-$value which, according to the definition above, is given by

$$P_{\mu_0}(|Z_n| \geq |z_n|) = P(|Z| \geq |z_n|) = 2P(Z \geq |z_n|).$$

Based on this, one should reject H_0, at level α, if $2P(Z \geq |z_n|) \leq \alpha$. It is easy to see that

$$|z_n| \geq z_{1-\alpha/2} \Leftrightarrow 2P(Z \geq |z_n|) \leq \alpha. \square$$

Exercise 8.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known. Find a consistent test of size α for

$$H_0 : \ \mu \geq \mu_0 \ \ vs \ \ H_1 : \ \mu < \mu_0.$$

Calculate its $p-$value and its power function.

8.5 Likelihood Ratio Test (LRT)

The likelihood ratio test (LRT) is the most general and the most used test in practice. This method provides a unified approach for developing *optimal test procedures*. In fact, it have been shown that *whenever the uniformly most powerful test exists, the LRT procedure leads to it*. In addition to hypothesis testing, the LRT procedure can also be used to construct optimal confidence intervals.

Suppose we observe an iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a pd $f(x;\theta)$, for some $\theta \in \Theta$ (Θ can be a vector space). The ***likelihood ratio statistic*** for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$, with $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$, is defined by

$$\Lambda_n(\boldsymbol{x}) := \frac{\max_{\theta\in\Theta_0} L_n(\theta|\boldsymbol{x})}{\max_{\theta\in\Theta} L_n(\theta|\boldsymbol{x})} = \frac{\max_{H_0} L_n(\theta|\boldsymbol{x})}{\max_{H_0\cup H_1} L_n(\theta|\boldsymbol{x})} = \frac{L_n(\hat{\theta}_0)}{L_n(\hat{\theta})},$$

where $L_n(\theta|\boldsymbol{x}) = \prod_i f(x_i;\theta)$ is the likelihood function of θ based on of the observed sample $\boldsymbol{x}$,
$\hat{\theta} \equiv \hat{\theta}(\boldsymbol{x}) = \arg\max_{\theta\in\Theta} L_n(\theta|\boldsymbol{x})$ is the MLE of θ obtained by maximizing the likelihood over the whole parameter space Θ, and $\hat{\theta}_0 \equiv \hat{\theta}_0(\boldsymbol{x}) = \arg\max_{\theta\in\Theta_0} L_n(\theta|\boldsymbol{x})$ be the ***restricted*** MLE of θ obtained by maximizing the likelihood over the null parameter space $\Theta_0 \subset \Theta$. Notice that, by definition, $0 \leq \Lambda_n \leq 1$.

To understand the reasoning behind the LRT, consider the discrete case where $f(x;\theta)$ is a probability mass function. In this case, given an observed data $\boldsymbol{x}$, Λ_n can be expressed as

$$\frac{P_{\hat{\theta}_0}(\boldsymbol{X}=\boldsymbol{x})}{P_{\hat{\theta}}(\boldsymbol{X}=\boldsymbol{x})} = \frac{\max_{\theta\in\Theta_0} P_\theta(\boldsymbol{X}=\boldsymbol{x})}{\max_{\theta\in\Theta_0\cup\Theta_1} P_\theta(\boldsymbol{X}=\boldsymbol{x})} = \frac{\max Prob(\text{obs. data}|H_0)}{\max Prob(\text{obs. data}|H_0\cup H_1)}.$$

- A small value of this ratio means that the observed data are less likely to occur under the null hypothesis than in the case where no restrictions (as imposed by the null hypothesis) are applied. This means that there should be a parameter point in Θ_1 for which the observed sample is more likely than for any parameter point in Θ_0. Consequently, H_0 has to be rejected in favor of H_1.
- A value of this ratio close to 1 means that the observed data are nearly as likely to occur under the null hypothesis as without it. H_0 is therefore not binding and there is no reason to reject it.

This motivates the definition of LRT as $I(\Lambda_n \leq c)$, for some $c \in [0,1)$. In this way, the LRT rejects H_0 whenever $\Lambda_n \leq c$. c must be chosen, in $[0,1)$, so that the test level is α, i.e. $\max_{\theta\in\Theta_0} P_\theta(\Lambda_n \leq c) \leq \alpha$. According to the definition given above, the $p-$value of the LRT is $\max_{\theta\in\Theta_0} P_\theta(\Lambda_n \leq \lambda_n)$, where λ_n is the observed value of Λ_n, i.e. $\lambda_n = \Lambda_n(\boldsymbol{x})$.

To perform LRT, we need to (i) maximize the likelihood function L_n over the full and restricted parameter spaces and (ii) calculate the critical value or the $p-$value as defined

above. To do this, we need to know the "null distribution" of Λ_n (i.e. its distribution under H_0) or to express it as a monotonic function of some statistic T_n whose null distribution is known. The following examples illustrate this.

Example 8.7. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The likelihood function of (μ, σ^2) is given by

$$L_n(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

Assume that $\sigma^2 > 0$ *is known*.

- Let's find the LRT for testing $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$.
  Since $\max_{\mu=\mu_0} L_n(\mu, \sigma^2) = L_n(\mu_0, \sigma^2)$ and $\max_{\mu} L_n(\mu, \sigma^2) = L_n(\overline{X}_n, \sigma^2)$,

$$\begin{aligned}\Lambda_n = \frac{L_n(\mu_0, \sigma^2)}{L_n(\overline{X}_n, \sigma^2)} &= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((X_i - \mu_0)^2 - (X_i - \overline{X}_n)^2\right)\right) \\ &= \exp\left(-\frac{n}{2\sigma^2}(\overline{X}_n - \mu_0)^2\right) = \exp\left(-\frac{1}{2}Z_n^2\right),\end{aligned}$$

  where $Z_n = \sqrt{n}\dfrac{\overline{X}_n - \mu_0}{\sigma}$.
  The LRT rejects H_0 if $\Lambda_n \leq c$ (with $c \in [0,1)$) $\Leftrightarrow |Z_n| \geq k$, where $k = \sqrt{-2\log(c)} > 0$. This latter is determined so that the size of the test is α, that is $P_{\mu_0}(|Z_n| \geq k) = \alpha$. This is equivalent to say that $P(|Z| \geq k) = \alpha$, thus $k = z_{1-\alpha/2}$. We conclude that the LRT of level α for the above hypothesis is equivalent to $I(|Z_n| \geq z_{1-\alpha/2})$.

- Let's find the LRT for testing $H_0 : \mu \leq \mu_0$ *vs* $H_1 : \mu > \mu_0$.
  Since $\mu \mapsto \ell_n(\mu, \sigma^2) = \log L_n(\mu, \sigma^2)$ is strictly concave and has a (unique) maximum at $\overline{X}_n$, we have that

$$\arg\max_{\mu \leq \mu_0} L_n(\mu, \sigma^2) = \arg\max_{\mu \leq \mu_0} \ell_n(\mu, \sigma^2) = \begin{cases} \overline{X}_n & \text{if } \overline{X}_n \leq \mu_0 \\ \mu_0 & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

  So,

$$\Lambda_n = \begin{cases} 1 & \text{if } \overline{X}_n \leq \mu_0 \\ \frac{L_n(\mu_0, \sigma^2)}{L_n(\overline{X}_n, \sigma^2)} = \exp\left(-\frac{1}{2}Z_n^2\right) & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

  The LRT rejects H_0 if $\Lambda_n \leq c$ (with $c \in [0,1)$) $\Leftrightarrow |Z_n| \geq k$ **and** $\overline{X}_n > \mu_0 \Leftrightarrow Z_n \geq k$ (with $k > 0$). This latter is determined so that the size of the test is α, that is $\max_{\mu \leq \mu_0} P_\mu(Z_n \geq k) = \alpha$. This is equivalent to say that $P(Z \geq k) = \alpha$, thus $k = z_{1-\alpha}$. We conclude that the LRT of level α for the above hypothesis is equivalent to $I(Z_n \geq z_{1-\alpha})$.

Assume now that $\sigma^2 > 0$ *is unknown*.

- Let's find the LRT for testing $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$.

We know that

$$\max_{\mu=\mu_0,\sigma^2} L_n(\mu,\sigma^2) = L_n(\mu_0,\tilde{\sigma}_0^2) \text{ and } \max_{\mu,\sigma^2} L_n(\mu,\sigma^2) = L_n(\overline{X}_n,\hat{\sigma}_n^2),$$

where $\tilde{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (X_i-\mu_0)^2$ and $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2$. So,

$$\begin{aligned}
\Lambda_n &= \frac{L_n(\mu_0,\tilde{\sigma}_0^2)}{L_n(\overline{X}_n,\hat{\sigma}_n^2)} = \left(\frac{\tilde{\sigma}_0^2}{\hat{\sigma}_n^2}\right)^{-n/2}, \text{ with} \\
\frac{\tilde{\sigma}_0^2}{\hat{\sigma}_n^2} &= \frac{\sum_i (X_i-\mu_0)^2}{\sum_i (X_i-\overline{X}_n)^2} = \frac{\sum_i (X_i-\overline{X}_n)^2 + n(\overline{X}_n-\mu_0)^2}{\sum_i (X_i-\overline{X}_n)^2} = 1 + \frac{n(\overline{X}_n-\mu_0)^2}{\sum_i (X_i-\overline{X}_n)^2} \\
&= 1 + \frac{n(\overline{X}_n-\mu_0)^2}{(n-1)S_n^2} = 1 + \frac{T_n^2}{n-1},
\end{aligned}$$

where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2$ and $T_n = \sqrt{n}\frac{\overline{X}_n-\mu_0}{S_n}$. We see that $\Lambda_n \leq c \iff |T_n| \geq k$. And since under H_0, $T_n \sim t_{n-1}$, where t_{n-1} is the Student's t-distribution with $n-1$ degrees of freedom, the LRT of size α for the above hypothesis is equivalent to $I\left(|T_n| \geq t_{n-1;1-\frac{\alpha}{2}}\right)$.

- Let's find the LRT for testing $H_0: \mu \leq \mu_0 \ \ vs \ \ H_1: \mu > \mu_0$.
  Under H_0, the MLE of (μ,σ^2) is

$$\begin{cases} (\overline{X}_n,\hat{\sigma}_n^2) & \text{if } \overline{X}_n \leq \mu_0 \\ (\mu_0,\tilde{\sigma}_0^2) & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

  So,

$$\Lambda_n = \begin{cases} 1 & \text{if } \overline{X}_n \leq \mu_0 \\ \frac{L_n(\mu_0,\tilde{\sigma}_0^2)}{L_n(\overline{X}_n,\hat{\sigma}_n^2)} = \left(1+\frac{T_n^2}{n-1}\right)^{-n/2} & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

  The LRT rejects H_0 if $\Lambda_n \leq c$ (with $c \in [0,1)$) $\Leftrightarrow |T_n| \geq k$ **and** $\overline{X}_n > \mu_0 \Leftrightarrow T_n \geq k$. By the same reasoning as before (see the case of known σ), we conclude that the LRT of level α is equivalent to $I(T_n \geq t_{n-1;1-\alpha})$. □

Exercise 8.2. Let $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. Find the LRT of level α for testing $H_0: \mu \geq \mu_0 \ \ vs \ \ H_1: \mu < \mu_0$, assuming first that σ is known and then without this assumption.

8.6 Asymptotic Tests

LRT construction, as discussed in the previous section, is a complicated process that requires a case-by-case analysis. The main difficulty is to find the ***exact*** distribution, under H_0, of Λ_n (or some function of it), so that the appropriate critical value can be determined.

A way to overcome this difficulty is to rely on asymptotic theory, which, as we will see, allows the development of unified test procedures that can be applied to a wide range of problems. Unlike exact tests, asymptotic tests (those based on asymptotic theory) can, by

definition, only be applied in situations where ***the sample size is large enough***. We say that a test with power function $\pi_n(\theta)$ is of ***asymptotic level*** α if

$$\lim_{n\to\infty} \pi_n(\theta) \leq \alpha, \ \forall \theta \in \Theta_0.$$

There are three classical asymptotic test methods based on the likelihood function: the ***Wald*** tests, the ***Score*** (or ***Rao***) tests, and the (log-)***likelihood ratio*** tests. Under appropriate regularity conditions, these tests are asymptotically equivalent with the best possible power (asymptotically).

Let $\hat{\boldsymbol{\theta}}$ denote the MLE of $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. In what follows, the *regularity assumptions* required to ensure that $\sqrt{\boldsymbol{I}_n(\hat{\boldsymbol{\theta}})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1})$ are supposed to be fulfilled; see previous sections.

8.6.1 Simple null hypothesis

To aid intuition, we first consider the simplest case of one parameter ($d = 1$). Let

$$H_0 : \ \theta = \theta_0 \ \text{ vs } H_1 : \ \theta \neq \theta_0.$$

The three test statistics of interest are:

$$W_n \equiv W_n(\theta_0) := I_n(\hat{\theta})(\hat{\theta} - \theta_0)^2 \qquad \text{(Wald)}$$

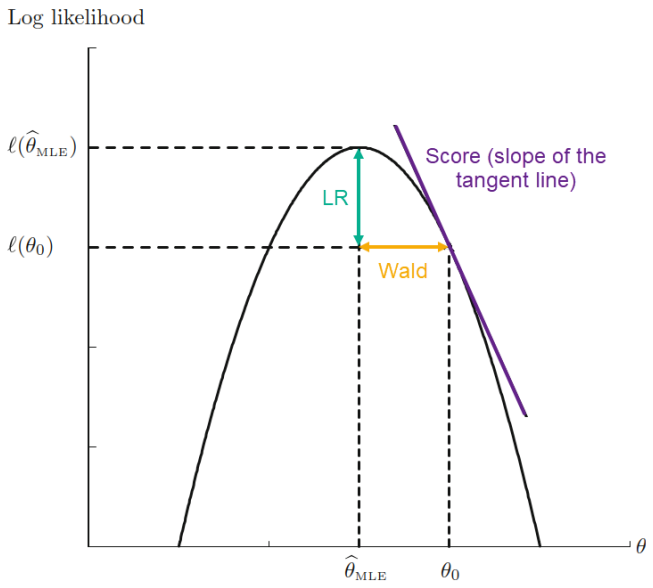
$$R_n \equiv R_n(\theta_0) := \frac{S_n^2(\theta_0)}{I_n(\theta_0)} \qquad \text{(Score)}$$

$$T_n \equiv T_n(\theta_0) := 2\left(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\right) \qquad \text{(Likelihood ratio)}$$

In the above formulas $\ell_n(\theta) = \sum_i \log f(X_i, \theta)$ is the log-likelihood and S_n is the corresponding Score function, i.e. $S_n(\theta) = \ell_n'(\theta|\boldsymbol{X}) = \sum_i \partial_\theta \log f(X_i, \theta)$.

Note that $T_n = -2\log \dfrac{L_n(\theta_0)}{L_n(\hat{\theta})} = -2\log \Lambda_n \in [0, \infty)$. This is the so-called *log-likelihood ratio statistic*, but for simplicity's sake we'll just call it the likelihood ratio (LR) statistic.

Although the above formulas defining the three statistics are very different, they are based on the same principle of assessing the distance between the MLE $\hat{\theta}$ and the null value θ_0: the greater the difference between these two, the stronger the evidence against the null hypothesis H_0. Wald uses $(\hat{\theta} - \theta_0)^2$, which we can describe as a "horizontal" difference (see the figure below). LRT uses $\left(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\right)$, i.e. the "vertical" or likelihood difference. Score uses $\left(S_n(\hat{\theta}) - S_n(\theta_0)\right)^2 = S_n^2(\theta_0)$, i.e. the "slope" (of the likelihood) difference. In any case, once the "distance" has been chosen, H_0 must be rejected if it is found to be "too far" away from the data (i.e. the test statistic is "too large").



In the Wald statistic, $(\hat{\theta} - \theta_0)^2$ is multiplied by the Fisher information I_n to obtain a "standardized distance" with a fixed/known distribution. In fact, we have seen that $\sqrt{I_n(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$, so, by Slutsky's theorem, $W_n(\theta) \xrightarrow{d} \chi^2_1$ (*here θ refers to the true parameter value*). Thus, under H_0, $W_n \xrightarrow{d} \chi^2_1$.
Note that the Wald statistic can be expressed as $W_n = Z_n^2$, with

$$Z_n \equiv Z_n(\theta_0) := \frac{\hat{\theta} - \theta_0}{\sqrt{\widehat{Avar}(\hat{\theta})}},$$

with $\widehat{Avar}(\hat{\theta}) = I_n^{-1}(\hat{\theta})$. $I_n(\hat{\theta})$ can be replaced by the observed Fisher information $J_n(\hat{\theta}) = -\ell_n''(\hat{\theta})$, or by any asymptotically equivalent quantity (like $I_n(\theta_0)$ or $J_n(\theta_0)$), without altering the asymptotic distribution of W_n.

As for the LRT, observe that, by second-order Taylor's expansion of $\theta \mapsto \ell_n(\theta)$ around $\hat{\theta}$,

$$\begin{aligned} \ell_n(\theta) - \ell_n(\hat{\theta}) &\approx \frac{1}{2}(\theta - \hat{\theta})^2 \ell_n''(\hat{\theta}) \\ \Rightarrow 2\big(\ell_n(\hat{\theta}) - \ell_n(\theta)\big) &\approx \frac{-n^{-1}\ell_n''(\hat{\theta})}{I(\theta)} \left(\sqrt{nI(\theta)}(\hat{\theta} - \theta)\right)^2 \end{aligned}$$

Since, $\sqrt{I_n(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$ and $-n^{-1}\ell_n''(\hat{\theta}) \xrightarrow{p} I(\theta)$, we conclude that $T_n(\theta) \xrightarrow{d} \chi^2_1$. Thus, under H_0, $T_n \xrightarrow{d} \chi^2_1$.

As for the Score test, using the first-order Taylor expansion of $\theta \mapsto S_n(\theta)$ around θ_0, and following the same line of reasoning as above shows that, under H_0, $S_n \xrightarrow{d} \chi^2_1$. Similarly to the Wald statistic, in the Score statistic, one could replace $I_n(\theta_0)$ by any consistent estimator of it like for example $J_n(\theta_0)$. Note also that the Score statistic can be expressed as $\left(\frac{S_n(\theta_0)}{\sqrt{Var(S_n(\theta_0))}}\right)^2$.

To conclude, we can say that, *under H_0, the three statistics (Wald, Score, LR) share the same asymptotic distribution, namely the χ^2_1*. Based on this result, the Wald test, the asymptotic LRT and the Score test reject H_0, in favor of H_1, if the observed value of their statistics (i.e. W_n for Wald, R_n for the Score and T_n for LR) is $\geq \chi^2_{1;1-\alpha}$, where $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)-$quantile of χ^2_1. These tests have an ***asymptotic size*** equal to α. We can check this easily for the LRT, for example, by observing that its type I error probability is $P_{\theta_0}(T_n \geq \chi^2_{1;1-\alpha}) \to P(\chi^2_1 \geq \chi^2_{1;1-\alpha}) = \alpha$.

In practice, it is common to use the ***asymptotic $p-$value*** $P(\chi^2_1 \geq t_n)$ (t_n is the observed value of T_n) and to reject H_0 whenever this quantity is less than α. The exact same considerations applies to the other two statistics (Wald and Score).

Wald, Score, and LR tests are asymptotically equivalent: they reach the same decision with probability approaching 1 as $n \to \infty$. However, their performance can be *quite different for a finite sample size*. Each method has its strengths and limitations:

- Based on the statistic $Z_n = \sqrt{I_n(\hat{\theta})}(\hat{\theta} - \theta_0)$, it is straightforward to create one-sided Wald tests (e.g. tests of $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$), but this is more difficult with Score and likelihood ratio statistics.
- The Wald statistic is by far the simplest and the easiest to interpret. It yields immediate confidence intervals and it is largely available in standard computing packages.
- The Wald test is not limited to MLE estimation, one just need to know the asymptotic distribution of the estimator under studied.
- The Score test does not require the MLE $\hat{\theta}$ whereas the other two tests do. The Score test also tends to give the best Type I error rates for small sample sizes.
- The Score test and likelihood ratio test are invariant under reparameterization, whereas the Wald test is not. For example, the Wald test about a scale parameter σ depends on whether the null hypothesis is expressed as $H_0 : \sigma = \sigma_0$ or $H_0 : \sigma^2 = \sigma_0^2$.
- In general, the likelihood ratio is more difficult to compute than Wald or Score tests, but it tends to have the greatest power.

Example 8.8. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Consider $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. Recall the log-likelihood function

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Assume that $\sigma^2 > 0$ *is known*. Also recall that

$$S_n(\mu) = \partial_\mu \ell_n(\mu, \sigma^2) = \frac{n}{\sigma^2}(\overline{X}_n - \mu),$$
$$I_n(\mu) = -E\left(S_n'(\mu)\right) = \frac{n}{\sigma^2},$$

and $\overline{X}_n$ is the MLE of μ. It follows that

$$W_n = \frac{n}{\sigma^2}(\overline{X}_n - \mu_0)^2 = \left(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}\right)^2,$$
$$R_n = \frac{\left(\frac{n}{\sigma^2}(\overline{X}_n - \mu_0)\right)^2}{\frac{n}{\sigma^2}} = \frac{n}{\sigma^2}(\overline{X}_n - \mu_0)^2,$$
$$T_n = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((X_i - \mu_0)^2 - (X_i - \overline{X}_n)^2\right) = \frac{n}{\sigma^2}(\overline{X}_n - \mu_0)^2.$$

The three test statistics are identical in this model. □

Example 8.9. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\pi)$. Consider $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$. Recall that

$$\ell_n(\pi) = n\hat{\pi}\log(\pi) + n(1-\hat{\pi})\log(1-\pi),$$
$$S_n(\pi) = n\frac{\hat{\pi} - \pi}{\pi(1-\pi)},$$
$$I_n(\pi) = \frac{n}{\pi(1-\pi)},$$

where $\hat{\pi} = n^{-1}\sum_i X_i$ is the MLE of π. It follows that

$$W_n = n\frac{(\hat{\pi} - \pi_0)^2}{\hat{\pi}(1-\hat{\pi})} = \left(\sqrt{n}\frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})}}\right)^2,$$
$$R_n = \frac{\left(n\frac{\hat{\pi}-\pi_0}{\pi_0(1-\pi_0)}\right)^2}{\frac{n}{\pi_0(1-\pi_0)}} = n\frac{(\hat{\pi} - \pi_0)^2}{\pi_0(1-\pi_0)},$$
$$T_n = 2\left\{n\hat{\pi}\log\frac{\hat{\pi}}{\pi_0} + n(1-\hat{\pi})\log\frac{1-\hat{\pi}}{1-\pi_0}\right\}.\square$$

Let's check the Type I error and the power of these tests using some simulations.

```
p.value <- function(pi0, pi, n) {
  x <- rbinom(n, size = 1, prob = pi)
  hat.pi <- mean(x)
  Wald <- n * ((hat.pi - pi0)^2 / (hat.pi * (1 - hat.pi)))
  Score <- n * ((hat.pi - pi0)^2 / (pi0 * (1 - pi0)))
  LR <- 2 * (n * hat.pi * ifelse(hat.pi == 0, 0, log(hat.pi / pi0)) +
               n * (1 - hat.pi) * ifelse(hat.pi == 1, 0, log((1 - hat.pi) / (1 - pi0))))
  p.valueW <- pchisq(Wald, df = 1, lower.tail = FALSE)
  p.valueS <- pchisq(Score, df = 1, lower.tail = FALSE)
  p.valueL <- pchisq(LR, df = 1, lower.tail = FALSE)
  return(c(wald = p.valueW, Score = p.valueS, LR = p.valueL))
}
# Type I error (pi.0 = true.pi = 0.5); n = 10, 100, 1000
```

```
rbind((replicate(5000, p.value(0.5, 0.5, 10)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.5, 0.5, 100)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.5, 0.5, 1000)) <= 0.05) |> rowMeans())
```

```
      wald  Score     LR
[1,] 0.1088 0.0222 0.1088
[2,] 0.0566 0.0566 0.0566
[3,] 0.0580 0.0580 0.0580
```

```
# Power (pi.0 = 0.4 and true.pi = 0.5); n = 10, 100, 1000
rbind((replicate(5000, p.value(0.4, 0.5, 10)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.4, 0.5, 100)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.4, 0.5, 1000)) <= 0.05) |> rowMeans())
```

```
      wald  Score     LR
[1,] 0.1968 0.0594 0.0700
[2,] 0.5478 0.5478 0.5478
[3,] 1.0000 1.0000 1.0000
```

The generalization of the three tests to the ***multiparameter case*** is not very difficult. Suppose we wish to test $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta} \in \mathbb{R}^d$. Then

$$W_n \equiv W_n(\boldsymbol{\theta}_0) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^t \boldsymbol{I}_n(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \quad \text{(Wald)}$$

$$R_n \equiv R_n(\boldsymbol{\theta}_0) = \boldsymbol{S}_n^t(\boldsymbol{\theta}_0)\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}_0)\boldsymbol{S}_n(\boldsymbol{\theta}_0) \quad \text{(Score)}$$

$$T_n \equiv T_n(\boldsymbol{\theta}_0) = 2\left(\ell_n(\hat{\boldsymbol{\theta}}) - \ell_n(\boldsymbol{\theta}_0)\right) \quad \text{(Likelihood ratio)}$$

Under H_0, these three statistics converge to χ^2_d, i.e. chi-squared distribution with d degrees of freedom. As consequence, the likelihood ratio test, for example, rejects H_0 when $T_n \geq \chi^2_{d;1-\alpha}$, or, equivalently, when $P(\chi^2_d \geq t_n) \leq \alpha$. The same applies to the other two statistics.

Example 8.10. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ and σ are unknown. Consider testing

$$H_0: \ \mu = \mu_0 \text{ and } \sigma^2 = \sigma_0^2 \ \ vs \ H_1: \ \mu \neq \mu_0 \text{ or } \sigma^2 \neq \sigma_0^2.$$

We know that

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2,$$

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \left(n\frac{\overline{X}_n - \mu}{\sigma^2}, -\frac{n}{2\sigma^2} + \frac{\sum_i (X_i - \mu)^2}{2\sigma^4}\right)^t,$$

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

and that $(\overline{X}_n, \hat{\sigma}_n^2)$, with $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, is the MLE of (μ, σ^2). Put $\tilde{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (X_i - \mu_0)^2$.
The Wald statistic is given by

$$\begin{pmatrix} \overline{X}_n - \mu_0 & \hat{\sigma}_n^2 - \sigma_0^2 \end{pmatrix} \begin{pmatrix} \frac{n}{\hat{\sigma}_n^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}_n^4} \end{pmatrix} \begin{pmatrix} \overline{X}_n - \mu_0 \\ \hat{\sigma}_n^2 - \sigma_0^2 \end{pmatrix} = n\frac{(\overline{X}_n - \mu_0)^2}{\hat{\sigma}_n^2} + \frac{n}{2}\left(1 - \frac{\sigma_0^2}{\hat{\sigma}_n^2}\right)^2 \xrightarrow{d} \chi_2^2, \text{ if } H_0 \text{ is true.}$$

The Score statistic is given by

$$\begin{aligned} &\begin{pmatrix} n\frac{\overline{X}_n - \mu_0}{\sigma_0^2} & -\frac{n}{2\sigma_0^2} + \frac{\sum_i (X_i - \mu_0)^2}{2\sigma_0^4} \end{pmatrix} \begin{pmatrix} \frac{\sigma_0^2}{n} & 0 \\ 0 & \frac{2\sigma_0^4}{n} \end{pmatrix} \begin{pmatrix} n\frac{\overline{X}_n - \mu_0}{\sigma_0^2} \\ -\frac{n}{2\sigma_0^2} + \frac{\sum_i (X_i - \mu_0)^2}{2\sigma_0^4} \end{pmatrix} \\ &= n\frac{(\overline{X}_n - \mu_0)^2}{\sigma_0^2} + \frac{n}{2}\left(1 - \frac{\tilde{\sigma}_0^2}{\sigma_0^2}\right)^2 \xrightarrow{d} \chi_2^2, \text{ if } H_0 \text{ is true.} \end{aligned}$$

The LRT statistic is given by

$$\begin{aligned} T_n &= 2(\ell_n(\bar{X}_n, \hat{\sigma}_n^2) - \ell_n(\mu_0, \sigma_0^2)) \\ &= n\frac{(\bar{X}_n - \mu_0)^2}{\sigma_0^2} - n\log\left(\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right) + n\left(\frac{\hat{\sigma}_n^2}{\sigma_0^2} - 1\right) \xrightarrow{d} \chi_2^2, \text{ if } H_0 \text{ is true.} \square \end{aligned}$$

8.6.2 Composite null hypothesis

In practice, we are rarely interested in a simple null hypothesis in which all the d components of the vector parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ are assigned, as done previously. Often, we're only interested in a subset of $\boldsymbol{\theta}$, or in some specific constraints on its components. This can be formulated, in most cases, as follows

$$H_0 : \boldsymbol{A\theta} = \boldsymbol{a} \text{ vs } H_1 : \boldsymbol{A\theta} \neq \boldsymbol{a},$$

where $\boldsymbol{A}$ is an $r \times d$ *full rank matrix* ($r \leq d$ and the r rows of $\boldsymbol{A}$ are linearly independent) and $\boldsymbol{a}$ is an r-vector.
For example, suppose that $d = 5$, then

- $\theta_1 = \ldots = \theta_5 = 0 \Leftrightarrow \boldsymbol{A\theta} = \boldsymbol{0}$, with $\boldsymbol{A} = \mathbb{1}$ (5×5 identity matrix).
- $\theta_1 = 1, \theta_2 = 2, \ldots, \theta_5 = 5 \Leftrightarrow \boldsymbol{A\theta} = \boldsymbol{a}$, with $\boldsymbol{A} = \mathbb{1}$ and $\boldsymbol{a} = (1, 2, \ldots, 5)^t$.
- $\theta_2 = \theta_4 = 0 \Leftrightarrow \boldsymbol{A\theta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, with $\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$
- $\theta_2 + \theta_4 = 1 \Leftrightarrow \boldsymbol{A\theta} = 1$, with $\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \end{pmatrix}$
- $\theta_1 = 2\theta_2$ and $\theta_3 = \theta_4 \Leftrightarrow \boldsymbol{A\theta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, with $\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix}$

The generalized Wald test statistic, the generalized Score test statistic and the generalized

LRT statistic are given, respectively, by

$$\begin{aligned} W_n &= (\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{a})^t (\boldsymbol{A}\boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{A}^t)^{-1}(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{a}), \\ R_n &= \boldsymbol{S}_n^t(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}})\boldsymbol{I}_n^{-1}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}})\boldsymbol{S}_n(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}), \\ T_n &= 2\left(\ell_n(\hat{\boldsymbol{\theta}}) - \ell_n(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}})\right), \end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}$ is the *restricted MLE of* $\boldsymbol{\theta}$ *obtained by maximizing the (log-)likelihood under* H_0*, i.e. under the constraint that* $\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{a}$.

Similar to the simple null hypothesis case, ***these three statistics converge, Under*** H_0***, to*** χ^2_r, i.e. chi-squared distribution with r degrees of freedom. This latter corresponds to the number of restrictions imposed by H_0.

Before moving on and looking at some examples, let's focus on the LRT and consider the case of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$, for some given $\boldsymbol{\theta}_{10}$.

In the remainder of this section, in order to simplify calculations and make it easier to follow-up, we'll concentrate solely on the Wald and the LR tests.

Example 8.11. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ and σ are unknown. Consider testing

$$H_0 : \ \mu = \mu_0 \ \ vs \ \ H_1 : \ \mu \neq \mu_0.$$

The Wald statistic is given by

$$\begin{aligned} W_n &= \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \overline{X}_n \\ \hat{\sigma}_n^2 \end{pmatrix} - \mu_0 \right)^t \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{\hat{\sigma}_n^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}_n^4}{n} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \overline{X}_n \\ \hat{\sigma}_n^2 \end{pmatrix} - \mu_0 \right) \\ &= n \frac{(\overline{X}_n - \mu_0)^2}{\hat{\sigma}_n^2} \xrightarrow{d} \chi^2_1, \text{ if } H_0 \text{ is true.} \end{aligned}$$

Since the MLE of (μ, σ^2) under H_0 is $(\mu_0, \tilde{\sigma}_0^2)$, where $\tilde{\sigma}_0^2 = n^{-1}\sum_i (X_i - \mu_0)^2$, the LRT statistic is given by

$$\begin{aligned} T_n &= 2(\ell_n(\bar{X}_n, \hat{\sigma}_n^2) - \ell_n(\mu_0, \tilde{\sigma}_0^2)) \\ &= n \log\left(\frac{\tilde{\sigma}_0^2}{\hat{\sigma}_n^2}\right) = n \log\left(1 + \frac{(\overline{X}_n - \mu_0)^2}{\hat{\sigma}_n^2}\right) \xrightarrow{d} \chi^2_1, \text{ if } H_0 \text{ is true.} \square \end{aligned}$$

Example 8.12. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\pi_1)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} Ber(\pi_2)$ be two independent samples. Consider testing

$$H_0 : \ \pi_1 = \pi_2 \ \ vs \ \ H_1 : \ \pi_1 \neq \pi_2.$$

Based on the sample (X_i, Y_i), $i = 1, \ldots, n$, it's easy to see that the Log-likelihood for (π_1, π_2)

is

$$\ell_n(\pi_1, \pi_2) = n\hat{\pi}_1 \log(\pi_1) + n(1-\hat{\pi}_1)\log(1-\pi_1) + n\hat{\pi}_2 \log(\pi_2) + n(1-\hat{\pi}_2)\log(1-\pi_2),$$

where $(\hat{\pi}_1, \hat{\pi}_2) := \left(n^{-1}\sum_i X_i, n^{-1}\sum_i Y_i\right)$ is the MLE of (π_1, π_2). The FI matrix is

$$\boldsymbol{I}_n(\pi_1, \pi_2) = \begin{pmatrix} \frac{n}{\pi_1(1-\pi_1)} & 0 \\ 0 & \frac{n}{\pi_2(1-\pi_2)} \end{pmatrix}.$$

The Wald statistic is given by

$$\begin{aligned} W_n &= \left(\begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - 0\right)^t \left(\begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n} & 0 \\ 0 & \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n} \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)^{-1} \left(\begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - 0\right) \\ &= n\frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_2(1-\hat{\pi}_2)} \xrightarrow{d} \chi^2_1, \text{ if } H_0 \text{ is true.} \end{aligned}$$

Under H_0, the log-likelihood reduces to

$$\ell_n(\pi_1, \pi_1) = n(\hat{\pi}_1 + \hat{\pi}_2)\log(\pi_1) + n(2 - \hat{\pi}_1 - \hat{\pi}_2)\log(1-\pi_1).$$

It's easy to see that, in this case, the MLE of π_1 $(= \pi_2)$ is $\tilde{\pi}_0 = \frac{\sum_{i=1}^n (X_i + Y_i)}{2n}$. As a result, the LRT statistic is given by

$$\begin{aligned} T_n &= 2\left(\ell_n(\hat{\pi}_1, \hat{\pi}_2) - \ell_n(\tilde{\pi}_0, \tilde{\pi}_0)\right) \\ &= 2n\sum_{j=1}^{2}\left(\hat{\pi}_j \log\left(\frac{\hat{\pi}_j}{\tilde{\pi}_0}\right) + (1-\hat{\pi}_j)\log\left(\frac{1-\hat{\pi}_j}{1-\tilde{\pi}_0}\right)\right) \xrightarrow{d} \chi^2_1, \text{ if } H_0 \text{ is true.} \square \end{aligned}$$

We now turn to a special, but very useful, case of the general formulation of the LR statistic presented above. Consider the case of a parametric model indexed by (θ, η), where θ and η are unknown. Let's say that θ is our parameter of interest and η is a nuisance parameter (θ and η can be vectors). In this context, our objective is to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, for some given θ_0. Let L denote the likelihood function of the model under study, $(\hat{\theta}, \hat{\eta}) = \arg\max_{\theta,\eta} L(\theta, \eta)$ be the MLE of (θ, η) and $\hat{\eta}_0 \equiv \hat{\eta}(\theta_0) := \arg\max_\eta L(\theta_0, \eta)$ be the restricted MLE of η under H_0. According to the general definition given above, the LRT statistic is

$$T_n \equiv T_n(\theta_0) := -2\log\frac{L(\theta_0, \hat{\eta}_0)}{L(\hat{\theta}, \hat{\eta})} = -2\log\frac{L_p(\theta_0)}{L_p(\hat{\theta})},$$

where $L_p(\theta) = \max_\eta L(\theta, \eta)$ is the profile likelihood for θ, i.e. $L_p(\theta) = L(\theta, \hat{\eta}(\theta))$, with $\hat{\eta}(\theta) = \arg\max_\eta L(\theta, \eta)$. In this case, $T_n \xrightarrow{d} \chi^2_r$, if H_0 is true, where r is the length of θ. T_n is sometimes referred to as the profile likelihood ratio statistic.

Example 8.13. Let's consider the case of the linear model $Y = \beta_0 + \beta_1 x + \epsilon$, where, for example, $\beta_0 = 10$, $\beta_1 = 20$, $x = 1, \ldots, 10$ and $\epsilon \sim N(0, 10^2)$. Let's say we observe $n = 10$

observations from (Y, x). In this context, we'd like to test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Under the linear model assumption, the LRT statistic is

$$2(\ell_n(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) - \ell_n(\hat{\beta}_{00}, 0, \hat{\sigma}_0^2)),$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ are the unrestricted MLEs (i.e. those we saw in Section 7.3) and $\hat{\beta}_{00}$ and $\hat{\sigma}_0^2$ are the restricted MLEs (i.e. those that maximize likelihood assuming $\beta_1 = 0$).

```
library(stats4)

set.seed(5)
x <- 1:10
y <- 10 + 20 * x + rnorm(10, sd = 10)

neglogLiReg <- function(a, b, s, y) {
  -sum(dnorm(y, mean = a + b * x, sd = s, log = TRUE))
}

loglik <- mle(\(a, b, s) neglogLiReg(a, b, s, y = y),
              start = list(a = mean(y), b = 0, s = sd(y))) |> logLik()
loglik0 <- mle(\(a, s) neglogLiReg(a, b = 0, s, y = y),
               start = list(a = mean(y), s = sd(y))) |> logLik()

LogLR <- (2 * (loglik - loglik0))[1] |> print()
pchisq(LogLR, df = 1, lower.tail = FALSE)
```

```
[1] 37.07
[1] 1.14e-09
```

A simple way to perform such a test in R is to use the `drop1()` function.

```
glm(y ~ x) |> drop1(test = "LRT")
```

```
Single term deletions

Model:
y ~ x
       Df Deviance   AIC scaled dev. Pr(>Chi)
<none>         812  78.3
x       1    33072 113.4        37.1  1.1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```