Politecnico di Milano

Lecture Notes for

STATISTICS

(FOR STUDENTS IN INFORMATION ENGINEERING)

code 079086

17th June 2008

These notes have been adapted by A. Barchielli and S. Salvati from the original "Appunti del corso di Statistica per Ingegneria Informatica" by Ilenia Epifani, Copyright 2005.

The English translation is by Simonetta Salvati.

Contents

Chapter 1

Introduction

These Lecture Notes are not entirely selfconsistent and do not cover the whole program of the course, but are to be intended as complements to our textbook [11]:
W. R. Pestman, *Mathematical Statistics, an Introduction* (1998, De Gruyter), ISBN 3-11-015356-4

First of all we need notions from Probability and from the theory of distributions. One can find such notions in Chapter I of the textbook [11], or in the lecture notes used in the course "Calcolo delle Probabilità" [5], or in [14, 15].

1.1 Some notations

i.i.d. = independent and identically distributed
c.d.f. = cumulative distribution function
iff = if and only if
$X \sim F$: the random variable X has c.d.f. F.
$X \sim f$: the random variable X has density f.
$X \sim$ "name": the random variable X follows the law (distribution) "name".

Indicator function: $$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \in A^c \end{cases}$$

1.2 Some parametric families of distributions

The tables of distributions from page 6 to page 9 can be extracted and used during the written examinations.

These tables may contain distribution which you never saw, but which could be useful, and the expressions for the "characteristic function", which is not in the program of the course.

1.2.1 Some discrete distributions

Name, symbol, parameters	Discrete density $p_X(k) = \mathrm{P}[X = k]$	Mean	Variance
Bernoulli $X \sim \mathrm{Be}(p) = \mathcal{B}(1, p)$ $0 \leq p = 1 - q \leq 1$	$p_X(1) = p$ $p_X(0) = q$	p	pq
binomial $X \sim \mathcal{B}(n, p)$ $0 \leq p = 1 - q \leq 1,$ $n = 1, 2, \ldots$	$\binom{n}{k} p^k q^{n-k}$ $k = 0, 1, \ldots, n$	np	npq
hypergeometric $X \sim \mathcal{H}(N, K, n)$ $K \geq 0$, $N, n \geq 1$ integer $n \leq N$, $K \leq N$	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}$ $k \geq 0$ integer, $k \leq n$, $k \leq K$, $n - k \leq N - K$	$n\,\frac{K}{N}$	$n\,\frac{K}{N}\left(1 - \frac{K}{N}\right)$ $\times\,\frac{N-n}{N-1}$
geometric $X \sim \mathcal{G}(p) = \mathcal{B}(-1, p)$ $0 < p = 1 - q \leq 1$	$p(1-p)^k$ $k = 0, 1, \ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$
geometric (from 1) $X - 1 \sim \mathcal{G}(p)$ $0 < p = 1 - q \leq 1$	$p_X(k) = pq^{k-1}$ $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{q}{p^2}$
negative binomial $X \sim \mathcal{B}(-n, p)$ $0 < p = 1 - q \leq 1,$ $n = 1, 2, \ldots$	$\binom{n+k-1}{k} p^n q^k$ $k = 0, 1, 2, \ldots$	$n\,\frac{q}{p}$	$n\,\frac{q}{p^2}$
Poisson $X \sim \mathcal{P}(\lambda)$ $\lambda > 0$	$\frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}$ $k = 0, 1, 2, \ldots$	λ	λ
uniform on $1, 2, \ldots, n$ $n \geq 1$ integer	$p_X(k) = 1/n$ $k = 1, 2, \ldots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

Name, symbol, parameters	Moment generating function $M\left(t\right)=\mathbb{E}\left[\mathrm{e}^{tX}\right]$	Characteristic function $H\left(t\right)=\mathbb{E}\left[\mathrm{e}^{\mathrm{i}tX}\right]$
Bernoulli $X\sim\mathrm{Be}(p)=\mathcal{B}(1,p)$ $0\leq p=1-q\leq 1$	$q+p\mathrm{e}^{t}$	$q+p\mathrm{e}^{\mathrm{i}t}$
binomial $X\sim\mathcal{B}(n,p)$ $0\leq p=1-q\leq 1$, $n=1,2,\ldots$	$\left(q+p\mathrm{e}^{t}\right)^{n}$	$\left(q+p\mathrm{e}^{\mathrm{i}t}\right)^{n}$
geometric $X\sim\mathcal{G}(p)=\mathcal{B}(-1,p)$ $0<p=1-q\leq 1$	$\dfrac{p}{1-q\mathrm{e}^{t}}$	$\dfrac{p}{1-q\mathrm{e}^{\mathrm{i}t}}$
geometric (from 1) $X-1\sim\mathcal{G}(p)$ $0<p=1-q\leq 1$	$\dfrac{p\mathrm{e}^{t}}{1-q\mathrm{e}^{t}}$	$\dfrac{p\mathrm{e}^{\mathrm{i}t}}{1-q\mathrm{e}^{\mathrm{i}t}}$
negative binomial $X\sim\mathcal{B}(-n,p)$ $0<p=1-q\leq 1$, $n=1,2,\ldots$	$\left(\dfrac{p}{1-q\mathrm{e}^{t}}\right)^{n}$	$\left(\dfrac{p}{1-q\mathrm{e}^{\mathrm{i}t}}\right)^{n}$
Poisson $X\sim\mathcal{P}(\lambda)$ $\lambda>0$	$\exp\left[\lambda\left(\mathrm{e}^{t}-1\right)\right]$	$\exp\left[\lambda\left(\mathrm{e}^{\mathrm{i}t}-1\right)\right]$
uniform on $1,2,\ldots,n$ $n\geq 1$ integer	$\dfrac{\mathrm{e}^{t}\left(1-\mathrm{e}^{nt}\right)}{n\left(1-\mathrm{e}^{t}\right)}$	$\dfrac{\mathrm{e}^{\mathrm{i}t}\left(1-\mathrm{e}^{\mathrm{i}nt}\right)}{n\left(1-\mathrm{e}^{\mathrm{i}t}\right)}$

1.2.2 A useful formula for the mean value

If $X\sim F(x)$, then

$$\mathbb{E}[X]=\int_{0}^{+\infty}\left[1-F(x)\right]\mathrm{d}x-\int_{-\infty}^{0}F(x)\mathrm{d}x\,. \tag{1.1}$$

The mean value exists iff both integrals converge.

1.2.3 Some continuous distributions

Name, symbol, parameters	Probability density $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{P}\left[X \le x\right]$	Mean	Variance
uniform $X \sim \mathcal{U}(a,b)$ $a < b$	$\frac{1}{b-a}$ $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
beta $X \sim \beta(a,b)$ $a > 0$, $b > 0$	$\frac{1}{B(a,b)}\, x^{a-1}\,(1-x)^{b-1}$ $0 < x < 1$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)} \times$ $\times\,(a+b)^{-2}$
normal or Gaussian $X \sim \mathcal{N}(\mu;\sigma^2)$ $\mu \in \mathbb{R}$, $\sigma > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2
Student t $X \sim t(k)$ $k > 0$	$\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \times$ $\times \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}$	0 for $k > 1$	$\frac{k}{k-2}$ for $k > 2$
log-normal $\mu \in \mathbb{R}$, $\sigma > 0$	$\left(x\sqrt{2\pi\sigma^2}\right)^{-1} \times$ $\times \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$ $x > 0$	$\mathrm{e}^{\mu+\sigma^2/2}$	$\mathrm{e}^{2\mu+\sigma^2} \times$ $\times \left(\mathrm{e}^{\sigma^2} - 1\right)$
exponential $X \sim \mathcal{E}(\theta)$ $= \Gamma(1,\theta)$ $\theta > 0$	$\frac{1}{\theta}\,\mathrm{e}^{-x/\theta}$ $x > 0$	θ	θ^2
Gamma $X \sim \Gamma(r,\theta)$ $r > 0$, $\theta > 0$	$\frac{x^{r-1}}{\theta^r\Gamma(r)}\,\mathrm{e}^{-x/\theta}$ $x > 0$	$r\theta$	$r\theta^2$
chi-square $X \sim \chi^2(k)$ $k \ge 1$, integer	$\chi^2(k) = \Gamma\left(\frac{k}{2}, 2\right)$	k	$2k$
Fisher F $X \sim F(m,n)$ $m, n \ge 1$, integer	$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{m/2} \times$ $\times\,\frac{x^{(m-2)/2}}{\left(1+\frac{m}{n}x\right)^{(m+n)/2}}$ $x > 0$	$\frac{n}{n-2}$ for $n > 2$	$\frac{2n^2}{m(n-2)^2} \times$ $\times\,\frac{(m+n-2)}{(n-4)}$ for $n > 4$
Weibull $a > 0$, $b > 0$	$abx^{b-1}\,\exp\left(-ax^b\right)$ $x > 0$	$a^{-1/b} \times$ $\Gamma(1+b^{-1})$	$a^{-2/b} \times$ $\left[\Gamma(1+2b^{-1}) -\right.$ $\left.\Gamma(1+b^{-1})^2\right]$

Name, symbol, parameters	Moment generating function $M(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right]$	Characteristic function $H(t) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}tX}\right]$
uniform $X \sim \mathcal{U}(a,b)$ $a < b$	$\dfrac{\mathrm{e}^{bt} - \mathrm{e}^{at}}{(b-a)\,t}$	$\dfrac{\mathrm{e}^{\mathrm{i}bt} - \mathrm{e}^{\mathrm{i}at}}{(b-a)\,\mathrm{i}t}$
normal or Gaussian $X \sim \mathcal{N}(\mu;\sigma^2)$ $\mu \in \mathbb{R}$, $\sigma > 0$	$\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$	$\exp\left(\mathrm{i}\mu t - \frac{1}{2}\sigma^2 t^2\right)$
exponential $X \sim \mathcal{E}(\theta)$ $= \Gamma(1,\theta)$ $\theta > 0$	$\dfrac{1}{1-\theta t}$ $\theta t < 1$	$\dfrac{1}{1-\mathrm{i}\theta t}$
Gamma $X \sim \Gamma(r,\theta)$ $r > 0$, $\theta > 0$	$\left(\dfrac{1}{1-\theta t}\right)^r$ $\theta t < 1$	$\left(\dfrac{1}{1-\mathrm{i}\theta t}\right)^r$
chi-square $X \sim \chi^2(k)$ $k \geq 1$, integer	$\left(\dfrac{1}{1-2t}\right)^{k/2}$ $t < 1/2$	$\left(\dfrac{1}{1-2\mathrm{i}t}\right)^{k/2}$

1.2.4 The function Gamma

$$\Gamma(r) = \int_0^{+\infty} x^{r-1}\,\mathrm{e}^{-x}\,\mathrm{d}x\,, \qquad r > 0\,,$$

$$\Gamma(r+1) = r\Gamma(r)\,, \qquad \Gamma(n+1) = n! \quad \text{for } n \text{ integer},$$
$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\,, \qquad \Gamma\left(\frac{2n+1}{2}\right) = \frac{(2n-1)!!}{2^n}\sqrt{\pi}\,, \quad n \geq 1\,.$$

1.2.5 The function beta

$$B(a,b) = \int_0^1 x^{a-1}\,(1-x)^{b-1}\,\mathrm{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\,, \qquad a > 0\,, \quad b > 0\,.$$

Pay attention, not all the authors use the same parameterizations. For exponential and Gamma laws we have used the same parameterization as Pestman [11].

Chapter 2

Point estimation

We take estimation theory from our textbook [11]; in this chapter we only give some complements. In particular, we discuss "optimality" criteria for choosing among different estimators of the same quantity and we introduce some general methods to construct estimators.

2.1 Definitions and notation

Let us start by the basic definitions of random sample, statistics, sample mean, sample variance already given in [11] at pages 63–65, 103. Let us consider a random variable X with c.d.f. $F(\cdot;\theta)$, depending on an unknown parameter θ. The set of possible values of θ is denoted by Θ and it is called the *parameter space*. One speaks of *parametric model* when Θ is finite-dimensional. In this case Θ is a subset of $\mathbb{R}^m$ for some m, so that θ is a one-dimensional or multi-dimensional real vector; some examples are $\Theta = \mathbb{R}$, $\Theta = [0,1]$, $\Theta = (0,+\infty)$, $\Theta = \mathbb{R}\times(0,+\infty)$. One speaks instead of *nonparametric model* when Θ is infinite-dimensional; an example is "all absolutely continuous distributions on $\mathbb{R}$".

We want to make inferences on the unknown parameter θ and, to this end, we make n independent observations (random sample) of $X \sim F(\cdot;\theta)$.

Definition 2.1. Let $X_1,\dots,X_n$ be n random variables, i.i.d. and with distribution $X_i \sim F(\cdot;\theta)$, $\theta\in\Theta$. We say that $X_1,\dots,X_n$ is a (simple) *random sample* (or a sample of size n) extracted from the population with distribution $F(\cdot;\theta)$.

When a density exists, it can be given in place of the c.d.f. $F(\cdot;\theta)$. We will use the same symbol $f(x,\theta)$ for both the discrete and the continuous case.

Let us stress that the random vector of the i.i.d. observations $(X_1,\dots,X_n)$ has joint distribution $F(x_1,\theta)\times\cdots\times F(x_n,\theta)$, or joint density $f(x_1,\theta)\times\cdots\times f(x_n,\theta)$.

In order to make inferences on θ, we are free to process the data $X_1,\dots,X_n$, but in this operation we cannot use θ itself because it is unknown. This idea of processing the data is catched in the notion of statistics.

Definition 2.2. A *statistics* is a random variable $T = g(X_1, \dots, X_n)$ which depends only on the "observations" $X_1, \dots, X_n$. The distribution of a statistics T is called *sampling distribution.*

The important point in the definition of statistics is that T does not depend on the unknown parameter θ: given the function g and the values of the observations, a computer can calculate the value of T. Moreover, the sampling distribution of T can be obtained from the distribution of $(X_1, \dots, X_n)$; since $F(x, \theta)$ depends on θ, the distribution of T depends on θ too.

Finally let us recall the definitions of two basic statistics: the sample mean $\overline{X}$ and the sample variance S^2:

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j, \tag{2.1}$$

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2 = \frac{n}{n-1} \left[\frac{1}{n} \sum_{j=1}^{n} X_j^2 - \overline{X}^2 \right]. \tag{2.2}$$

2.1.1 Estimators

A first type of inference concerns the value of θ or of some function of it. We have to give rules to extract from the data sensible approximations of such unknown values. This is the subject of "point estimation".

Definition 2.3. A *population characteristic* is a function $\kappa : \Theta \to \mathbb{R}$. If κ is constant on Θ, then κ is said to be a *trivial characteristic.*

Definition 2.4. Let $X_1, \dots, X_n$ be a random sample from $F(x, \theta)$, $\theta \in \Theta$, and $\kappa(\theta)$ a population characteristic. An *estimator* of $\kappa(\theta)$, based on the random sample $X_1, \dots, X_n$, is a statistics $T = g(X_1, \dots, X_n)$ considered with the aim of estimating $\kappa(\theta)$. The value taken by an estimator T of $\kappa(\theta)$ is called an *estimate* of $\kappa(\theta)$.

We now clarify the definitions given above by showing some examples of characteristics and estimators in very well known models.

Example 2.1 (Gaussian model). Suppose that $X_1, \dots, X_n$ is a simple random sample from a Gaussian distribution $\mathcal{N}(\mu; \sigma^2)$ with unknown parameters μ, σ. Then, $\theta = (\mu, \sigma^2)$ is a two dimensional parameter taking values in the parameter space $\Theta = \mathbb{R} \times (0, \infty)$.

Some examples of characteristics of a Gaussian population are the following: $\kappa_1(\mu, \sigma^2) =$ "population mean" $= \mu$, or $\kappa_2(\mu, \sigma^2) =$ "population variance" $= \sigma^2$, or $\kappa_3(\mu, \sigma^2) =$ "population second moment" $= \sigma^2 + \mu^2$, or $\kappa_4(\mu, \sigma^2) = P_{(\mu, \sigma^2)}(X_1 \leq 2) = \Phi\left(\frac{2-\mu}{\sigma}\right)$.

Natural estimators of μ and σ^2 are the sample mean $\overline{X}$ and sample variance S^2, respectively. Finally, an estimator of $\kappa_4(\mu, \sigma^2)$ is given by the statistics

$$U := \Phi\left(\frac{2-\overline{X}}{\sqrt{(n-1)S^2/n}}\right).$$

Example 2.2 (Bernoulli model). Let $X_1, \dots, X_n$ be i.i.d. random variables with discrete density $f(x,\theta)$ given by

$$f(x,\theta) = \theta\mathbf{1}_{\{1\}}(x) + (1-\theta)\mathbf{1}_{\{0\}}(x) = \theta^x(1-\theta)^{1-x}\mathbf{1}_{\{0,1\}}(x), \quad \theta \in \Theta = [0,1].$$

With this parametrization of the the Bernoulli model, the characteristic "mean" is the parameter θ itself, while the characteristic "variance" is the function $\theta(1-\theta)$. A possible estimator of θ is $\overline{X}$, while the variance can be estimated by $\overline{X}(1-\overline{X})$ or by the sample variance S^2.

Note that, in the Bernoulli model, we have $S^2 = \frac{n}{n-1}\overline{X}(1-\overline{X})$. Indeed, since, for any j, the only possible values of X_j are 0 and 1, it follows that $X_j^2 = X_j$ for each j; then, from the last expression in (2.2), we get

$$S^2 = \frac{n}{n-1}\Big[\frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}^2\Big] = \frac{n}{n-1}\Big[\overline{X} - \overline{X}^2\Big] = \frac{n}{n-1}\overline{X}(1-\overline{X}).$$

Example 2.3 (Poisson model). Let $X_1, \dots, X_n$ be i.i.d. random variables with discrete density $f(x,\theta)$ given by

$$f(x,\theta) = \frac{\mathrm{e}^{-\theta}\theta^x}{x!}\mathbf{1}_{\{0,1,2,\dots\}}(x), \quad \theta \in \Theta = (0,\infty). \tag{2.3}$$

Both the characteristics "mean" and "variance" are the parameter θ. It follows that the statistics $\overline{X}$ e S^2 are two different estimators of θ. Another interesting characteristic in this model is $\kappa(\theta) = P_\theta(X_1 > 0)$, i.e. $\kappa(\theta) = 1 - \mathrm{e}^{-\theta}$ which can be estimated by $1 - \mathrm{e}^{-\overline{X}}$ or $1 - \mathrm{e}^{-S^2}$.

Example 2.4 (Exponential model). Let $X_1, \dots, X_n$ be i.i.d. random variables with probability density $f(x,\theta)$ given by

$$f(x,\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x), \quad \theta \in \Theta = (0,\infty). \tag{2.4}$$

In the parametrization (2.4) of the exponential model, the characteristic "mean" is θ, while the characteristic "variance" is given by the function θ^2.

Example 2.5 (Uniform model). Let $X_1, \dots, X_n$ be i.i.d. random variables with probability density $f(x,\theta)$ given by

$$f(x,\theta) = \frac{1}{\theta}\mathbf{1}_{(0,\theta)}(x), \quad \theta \in \Theta = (0,\infty).$$

Then, the mean is $\theta/2$ and two possible estimators are $\overline{X}$ or, as we shall see later, $\max\{X_1, \dots, X_n\}/2$.

We conclude this introduction with some comments on the notation. Let $T = g(X_1, \dots, X_n)$ be an estimator of $\kappa(\theta)$ with mean $\mathbb{E}[T]$. In general $\mathbb{E}[T]$ will depend on θ; for example, in the case of an absolutely continuous density $f(x, \theta)$,

$$\mathbb{E}[T] = \mathbb{E}[g(X_1, \dots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f(x_1, \theta) \cdots f(x_n, \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

In order to recall this dependence, we will write $\mathbb{E}_\theta$ instead of simply $\mathbb{E}$. Analogously, the mean of any function $h(T)$ of the estimator T will be denoted by $\mathbb{E}_\theta[h(T)]$.

2.1.2 Mean square error

It is clear from the examples in the preceding subsection that, when we face the problem of estimating a population characteristic $\kappa(\theta)$, we can find ourselves in the problem of choosing among various estimators. Obviously, we would like the estimator to be "as close as possible" to $\kappa(\theta)$: a good estimator should be concentrated around $\kappa(\theta)$. There are many possible ways to measure the proximity of the estimator to $\kappa(\theta)$, for example $P(|T - \kappa(\theta)| < \epsilon)$, for some $\epsilon > 0$. Alternatively, $\mathbb{E}_\theta[(T - \kappa(\theta))^2]$ gives a "mean measure" of the proximity.

Definition 2.5. If T is an estimator of $\kappa(\theta)$ and $\mathbb{E}_\theta[(T - \kappa(\theta))^2] < \infty$, then the quantity

$$\mathrm{MSE}_T(\theta) = \mathbb{E}_\theta\left[\left(T - \kappa(\theta)\right)^2\right]$$

is called the *mean square error*[1] of the estimator T.

The MSE of an estimator T exists if and only if T has a finite second moment, or, equivalently, finite (mean and) variance. Indeed, by using the inequality $(a+b)^2 \leq 2(a^2 + b^2)$, which holds[2] for any two real numbers a, b, we get

$$\mathbb{E}_\theta\left[\left(T - \kappa(\theta)\right)^2\right] \leq 2\,\mathbb{E}_\theta[T^2 + \kappa(\theta)^2] = 2\,\mathbb{E}_\theta[T^2] + 2\kappa(\theta)^2,$$

hence $\mathbb{E}_\theta[T^2] < \infty$ implies $\mathbb{E}_\theta\left[\left(T - \kappa(\theta)\right)^2\right] < \infty$. Conversely,

$$\begin{aligned}\mathbb{E}_\theta[T^2] = \mathbb{E}_\theta\left[\left(T - \kappa(\theta) + \kappa(\theta)\right)^2\right] &\leq 2\,\mathbb{E}_\theta\left[\left(T - \kappa(\theta)\right)^2 + \kappa(\theta)^2\right] \\ &= 2\,\mathbb{E}_\theta\left[\left(T - \kappa(\theta)\right)^2\right] + 2\kappa(\theta)^2,\end{aligned}$$

hence $\mathbb{E}_\theta\left[\left(T - \kappa(\theta)\right)^2\right] < \infty$ implies $\mathbb{E}_\theta[T^2] < \infty$.

We will therefore restrict our attention to the class of estimators with finite variance for all θ. The MSE of T is a natural measure for judging the goodness of the estimator T and the criterion we shall adopt for choosing an estimator will be based on minimizing the MSE.

[1] Also the expression *mean squared error* is in use

[2] $(|a| - |b|)^2 \geq 0 \Rightarrow a^2 - 2|a||b| + b^2 \geq 0 \Rightarrow 2|a||b| \leq a^2 + b^2$, which implies $(a+b)^2 \leq (|a| + |b|)^2 = a^2 + 2|a||b| + b^2 \leq 2(a^2 + b^2)$.

Remark 2.6. There is an alternative expression for the mean square error which is important both for computational and for theoretical reasons. By adding and subtracting the mean value of T inside the square we get

$$\begin{aligned}\mathbb{E}_\theta\left[(T-\kappa(\theta))^2\right] &= \mathbb{E}_\theta\left[\left((T-\mathbb{E}_\theta[T])+(\mathbb{E}_\theta[T]-\kappa(\theta))\right)^2\right] \\ &= \mathbb{E}_\theta\left[(T-\mathbb{E}_\theta[T])^2+(\mathbb{E}_\theta[T]-\kappa(\theta))^2+2\,(T-\mathbb{E}_\theta[T])\,(\mathbb{E}_\theta[T]-\kappa(\theta))\right].\end{aligned}$$

But $\mathbb{E}_\theta\left[(T-\mathbb{E}_\theta[T])\,(\mathbb{E}_\theta[T]-\kappa(\theta))\right]=(\mathbb{E}_\theta[T]-\kappa(\theta))\,\mathbb{E}_\theta\left[T-\mathbb{E}_\theta[T]\right]=0$ and we obtain the decomposition of the MSE:

$$\mathrm{MSE}_T(\theta)=\mathbb{E}_\theta\left[\left(T-\kappa(\theta)\right)^2\right]=\mathrm{Var}_\theta[T]+\left(\mathbb{E}_\theta[T]-\kappa(\theta)\right)^2. \tag{2.5}$$

Observe that all the passages above are allowed since we assumed the mean and the variance of T to be finite. The quantity $\mathbb{E}_\theta[T]-\kappa(\theta)$ in the decomposition formula (2.5) is called the *bias* of T.

Using the MSE as a measure of the closeness of the estimator to the characteristic to be estimated, given two estimators T_1 and T_2, we would prefer the estimator T_1 to T_2 if

- $\mathbb{E}_\theta[(T_1-\kappa(\theta))^2]\le\mathbb{E}_\theta[(T_2-\kappa(\theta))^2]$ for all $\theta\in\Theta$, and,
- $\mathbb{E}_\theta[(T_1-\kappa(\theta))^2]<\mathbb{E}_\theta[(T_2-\kappa(\theta))^2]$ for some $\theta\in\Theta$.

It would follow that, given the class of all estimators of $\kappa(\theta)$ with finite variance, we would choose the estimator T_0 which minimizes the MSE for all θ, i.e. such that

$$\mathbb{E}_\theta[(T_0-\kappa(\theta))^2]\le\mathbb{E}_\theta[(T-\kappa(\theta))^2],\qquad\forall\theta\in\Theta,\qquad\forall T.$$

Unfortunately, such a T_0 does not exist. Suppose that $\kappa(\theta)=\theta\in\mathbb{R}$. Nobody forbids us to estimate $\kappa(\theta)$ by the constant 5, $\tilde{T}=5$. This choice is obviously unreasonable, but the MSE of $\tilde{T}$ in $\theta=5$ is $\mathbb{E}_5[(\tilde{T}-5)^2]=\mathbb{E}_5[(5-5)^2]=0$. In spite of the absurdity of the choice of the estimator $\tilde{T}=5$, this estimator behaves better than any other estimator in $\theta=5$.

The problem above arises because the class of all estimators with finite MSE is too large: it contains unreasonable estimators which however are good in a single point. The way out is to choose some reasonable subclass of estimators. The MSE decomposition formula (2.5), which says "MSE = variance + squared bias", suggests immediately a possible choice: the class of the estimators with null bias, which are called unbiased estimators. Then, the problem reduces to find the unbiased estimator with minimum variance for all $\theta\in\Theta$.

2.2 Unbiased estimators

In this section we present the class of unbiased estimators with some of their properties and some examples; we will come back to the problem of the best estimator in the next section.

Definition 2.6. Let T be a statistics such that $\mathbb{E}_\theta[T]$ exists for any θ in Θ. Then, T is said to be an *unbiased estimator* of the characteristic $\kappa(\theta)$ if

$$\mathbb{E}_\theta[T] = \kappa(\theta), \qquad \forall \theta \in \Theta. \tag{2.6}$$

Example 2.7. If $X_1, \dots, X_n$ is a random sample with common c.d.f. $F(x, \theta)$, $\theta \in \Theta$, and $\mathbb{E}_\theta[X_1]$ exists for any θ, then $\mathbb{E}_\theta[\overline{X}] = \mathbb{E}_\theta[X_1]$, $\forall \theta$. This shows that

The sample mean $\overline{X}$ is an unbiased estimator of the population mean $\mathbb{E}_\theta[X_1]$.

Moreover, if $\mathrm{Var}_\theta[X_1]$ exists $\forall \theta \in \Theta$, then $\mathbb{E}_\theta[S^2] = \mathrm{Var}_\theta[X_1]$, $\forall \theta \in \Theta$, i.e.

The sample variance S^2 is an unbiased estimator of the population variance $\mathrm{Var}_\theta[X_1]$.

Example 2.8. If $X_1, \dots, X_n$ are i.i.d. random variables with c.d.f. F, then $\mathbf{1}_{(-\infty,x]}(X_1)$ is a statistics with Bernoulli density with parameter $F(x)$, since $P[\mathbf{1}_{(-\infty,x]}(X_1) = 1] = P[X_1 \leq x] = F(x)$. Therefore, $\mathbf{1}_{(-\infty,x]}(X_1)$ is an unbiased estimator of $F(x)$. Another unbiased estimator of $F(x)$ is

$$F_n(x) := \frac{\sum_{j=1}^n \mathbf{1}_{(-\infty,x]}(X_j)}{n}.$$

The function defined by $x \mapsto F_n(x)$ is called *empirical cumulative function* and we can use it to estimate the c.d.f. when F is completely unknown. We will come back to this function in the last part of the course, which is dedicated to nonparametric statistics.

The unbiasedness property expresses the fact that the estimator does not systematically underestimate or overestimate the population characteristic under investigation. In other words, this property translates the request that the sample should "represent" the population. If this is the case, the estimator will, in the mean, restitute the unknown characteristic.

Even if we restrict ourselves to the class of unbiased estimators, we can be in troubles. Indeed, it can happen that an unbiased estimator does not exists (see Example 2.9), or, that it exists but it is not unique (see Example 2.10). Finally, it can happen that the unbiased estimator of $\kappa(\theta)$ exists and is unique, but it does not make sense (see Example 2.12).

Example 2.9. Let $X_1 \sim \mathbf{Bi}(n, \theta)$, $\theta \in (0,1)$, with known n and $\kappa(\theta) = 1/\theta$. Then $T = g(X_1)$ is an unbiased estimator of $1/\theta$ if and only if

$$\mathbb{E}_\theta[T] = \sum_{k=0}^n g(k) \binom{n}{k} \theta^k (1-\theta)^{n-k} = \frac{1}{\theta}, \qquad \forall \theta \in (0,1).$$

But, for any function g, this equality cannot hold for all θ because $\mathbb{E}_\theta[T]$ is a polynomial of degree $n \geq 1$ in θ.

Example 2.10. Let $X_1, \dots, X_n$ be a random sample from a Poisson distribution with parameter θ. Then $T_1 = \overline{X}$ and $T_2 = S^2$ are two different unbiased estimators of θ since both the population mean and variance are equal to θ.

Remark 2.11. If there exist two different unbiased estimators, then we can construct infinitely many unbiased estimators. Indeed, let T_1 and T_2 be two different unbiased estimators of $\kappa(\theta)$. Then, for all $a \in \mathbb{R}$, $T_a := aT_1 + (1-a)T_2$ has finite mean and we have, $\forall \theta \in \Theta$,

$$\mathbb{E}_\theta[T_a] = a\,\mathbb{E}_\theta[T_1] + (1-a)\,\mathbb{E}_\theta[T_2] = a\kappa(\theta) + (1-a)\kappa(\theta) = \kappa(\theta)\,.$$

Example 2.12. Suppose that the rush hour number of calls to a freephone number at i-th day can be modelled by a discrete random variable having a Poisson distribution with parameter θ. Let us assume that the numbers of calls in different days are independent. We know the number X_1 of calls during the first day and we want to estimate the probability that no calls would arrive during the successive two days. In other words, we are looking for an unbiased estimator $T = g(X_1)$ of the characteristic $\kappa(\theta) = P_\theta[X_2 + X_3 = 0]$. On one side we have

$$\mathbb{E}_\theta[T] = \sum_{k=0}^{\infty} g(k)\mathrm{e}^{-\theta}\frac{\theta^k}{k!}.$$

On the other side, since X_2, X_3 are i.i.d. $\sim \mathcal{P}(\theta)$, then $X_2 + X_3 \sim \mathcal{P}(2\theta)$ and

$$\kappa(\theta) = P_\theta[X_2 + X_3 = 0] = \mathrm{e}^{-2\theta} = \mathrm{e}^{-\theta}\sum_{k=0}^{\infty}(-1)^k\frac{\theta^k}{k!}.$$

Therefore, the unbiasedness condition $\mathbb{E}_\theta[T] = \kappa(\theta)$ gives

$$\sum_{k=0}^{\infty} g(k)\frac{\theta^k}{k!} = \sum_{k=0}^{\infty}(-1)^k\frac{\theta^k}{k!}\,, \qquad \forall\theta > 0.$$

The two power series in θ in the equation above are equal if and only if the coefficients coincide, i.e. if and only if $g(k) = (-1)^k$, $\forall k = 0, 1, 2, \ldots$. It follows that the only unbiased estimator of the characteristic $\mathrm{e}^{-2\theta}$ is $T = (-1)^{X_1}$. But, do you find it reasonable to estimate a quantity belonging to $(0,1)$ by using a statistics whose possible values are -1 and 1?

2.2.1 UMVU Estimators

We now come back to the problem of finding the optimal estimator: we start with a simple random sample $X_1, \ldots, X_n$ from a c.d.f. $\{F(\cdot, \theta), \theta \in \Theta\}$ and we look for "the" optimal estimator T^* of $\kappa(\theta)$, that is for the estimator with the following properties:

1. T^* is an unbiased estimator of $\kappa(\theta)$;
2. $\mathrm{Var}_\theta[T^*] \le \mathrm{Var}_\theta[T]$ for any θ and for any unbiased estimator T with finite variance.

Definition 2.7. An estimator T^* satisfying points 1. and 2. above is called *uniformly minimum variance unbiased estimator.*

In the statistical literature, if such an estimator T^* exists, it is shortly denoted by UMVUE. We now introduce some properties of UMVUE's, precisely *uniqueness* and *symmetry.*

Proposition 2.1 (Uniqueness of the UMVUE). *If it exists, the UMVUE is essentially unique, that is, if T_1, T_2 are both UMVUE, then $P_\theta[T_1 = T_2] = 1$, $\forall\theta \in \Theta$.*

Proof. See Pestman [11], page 110 (Theorem II.9.4). □

The uniqueness of the UMVUE allows us to speak of "the" UMVU estimator.

Proposition 2.2 (Symmetry of the UMVUE). *Let $T^* = g(X_1, \ldots, X_n)$ be an UMVUE. Then*

$$P_\theta[g(X_1, \ldots, X_n) = g(X_{\pi(1)}, \ldots, X_{\pi(n)})] = 1\,, \qquad \forall\theta \in \Theta \tag{2.7}$$

for any permutation π of $\{1, \ldots, n\}$.

Proof. See Pestman [11], page 113 (Theorem II.9.6). □

In brief, Proposition 2.2 says that the value of the UMVU estimator does not depend on the order of the data.

In general, any estimator satisfying Eq. (2.7) is said to be *essentially symmetric.* Roughly speaking, an estimator is symmetric when the observations have all the same role in its expression. For example, the sample mean $\overline{X}$ and the sample variance S^2 are symmetric statistics.

Nonsense. Finally we show an example where the UMVU estimator exists but it doesn't make sense. We refer again to Example 2.12 concerning the Poisson model where we obtained $T = (-1)^{X_1}$ as the unique unbiased estimator of $\mathrm{e}^{-2\theta}$. Obviously $T = (-1)^{X_1}$ has finite variance and therefore it is the unique UMVUE of $\mathrm{e}^{-2\theta}$. We already noticed that this estimator does not make sense.

In order to complete the subject, we should show when an UMVUE does exist and how it can be found. This would request some knowledge about the concept of "sufficient and complete" statistics and the proofs of remarkable results known as "Rao-Blackwell Lemma" and "Lehmann-Scheffé Lemma". But all this would in turn request the concept of conditional mean. We decided not to treat this part of the subject and we refer the interested reader to Rohatgi and Saleh [13]. Sufficiency is also treated in Pestman [11].

However, we will be able to establish in some particular but relevant cases if a given estimator is UMVU, by giving a lower bound for the variance of an estimator in statistical models satisfying some regularity properties (Section 2.5).

2.3 Some asymptotic properties of estimators

Suppose we can repeat an experiment in the same conditions infinitely many times. We then have a sequence of independent random variables $X_1, \dots, X_n, \dots$ where X_n describes the result of the experiment at the n-th repetition.

Intuitively we expect that a large sample is more "representative" of the underlying population than a small sample. Therefore we expect that a good estimator $T_n = g(X_1, \dots, X_n)$ of a characteristic $\kappa(\theta)$ approaches in some sense $\kappa(\theta)$ as n increases, that is, some convergence relation of the type $T_n \underset{n\to\infty}{\longrightarrow} \kappa(\theta)$ holds. Depending on the type of limit considered, we can define different properties translating this intuition, technically indicated as "consistency". We only mention here the "mean square error consistency" (*MSE-consistency*) or L^2-consistency.

Definition 2.8. Let $X_1, \dots, X_n, \dots$ be a sequence of i.i.d. random variables with c.d.f. $F(x, \theta)$, $\theta \in \Theta$, and, for any $n \in \mathbb{N}$, let T_n be an estimator of $\kappa(\theta)$ depending on the first n observations. The sequence $\{T_n\}_{n\in\mathbb{N}}$ of estimators of $\kappa(\theta)$ is said to be *MSE-consistent* if

$$\lim_{n\to\infty} \mathbb{E}_\theta[(T_n - \kappa(\theta))^2] = 0, \qquad \forall \theta \in \Theta.$$

Another important and natural definition for a sequence of estimators is the following one, which generalizes to a sequence the concept of unbiasedness.

Definition 2.9. Let $X_1, \dots, X_n, \dots$ be a sequence of i.i.d. random variables with c.d.f. $F(x, \theta)$, $\theta \in \Theta$, and, for any $n \in \mathbb{N}$, let T_n be an estimator of $\kappa(\theta)$ depending on the first n observations. The sequence $\{T_n\}_{n\in\mathbb{N}}$ of estimators of $\kappa(\theta)$ is said to be *asymptotically unbiased* if

$$\lim_{n\to\infty} \mathbb{E}_\theta[T_n] = \kappa(\theta), \qquad \forall \theta \in \Theta.$$

Remark 2.13. The decomposition formula (2.5) says that the MSE is the sum of two non negative terms,

$$\mathbb{E}_\theta[(T_n - \kappa(\theta))^2] = \mathrm{Var}_\theta[T_n] + [\mathbb{E}_\theta(T_n) - \kappa(\theta)]^2. \tag{2.8}$$

Therefore, the MSE-consistency is equivalent to the asymptotical unbiasedness plus the asymptotical vanishing of the variance: $\lim\limits_{n\to\infty} \mathrm{Var}_\theta[T_n] = 0$.

The following notion is useful when one needs an approximate distribution of an estimator in the case of a large sample.

Definition 2.10. Let $X_1, \dots, X_n, \dots$ be a sequence of i.i.d. random variables with c.d.f. $F(x, \theta)$, $\theta \in \Theta$, and, for any $n \in \mathbb{N}$, let T_n be a statistics depending on the first n observations. The sequence $\{T_n\}_{n\in\mathbb{N}}$ is said to be *asymptotically normal* (or Gaussian) with *asymptotic mean* $\mu_n(\theta)$ and *asymptotic variance* $\sigma_n^2(\theta)$ if

$$\lim_{n\to\infty} P\left[\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \le z\right] = \Phi(z), \qquad \forall z \in \mathbb{R}.$$

When the statistics T_n is used as an estimator of the characteristic $\kappa(\theta)$, we require the asymptotic mean to be $\kappa(\theta)$ itself and $P\left[\frac{T_n - \kappa(\theta)}{\sigma_n(\theta)} \le z\right] = \Phi(z)$, $\forall z \in \mathbb{R}$.

2.4 Likelihood function

Let us consider now random samples extracted from distributions admitting a density $f(\cdot, \theta)$. If $f(\cdot, \theta)$ is the density function of the i.i.d. variables X_i, for $i = 1, \dots, n$, then the density function $f(x_1, \dots, x_n; \theta)$ of the vector $(X_1, \dots, X_n)$ is given by the product $f(x_1, \dots, x_n; \theta) = f(x_1, \theta) \cdots f(x_n, \theta)$.

Definition 2.11. The *likelihood function* of n random variables $X_1, \dots, X_n$ is defined to be the joint density of the n random variables, say $f_{X_1,\dots,X_n}(x_1, \dots, x_n; \theta)$, which is considered as a function of θ. In particular, if $X_1, \dots, X_n$ is a random sample from the density $f(\cdot, \theta)$, then the likelihood function is

$$\theta \mapsto L_\theta(x_1, \dots, x_n) = \prod_{j=1}^{n} f(x_j, \theta).$$

The likelihood function is simply the joint density of the sample seen as a function of θ. The values $x_1, \dots, x_n$ are considered as fixed; they are the actually observed values, the *sample realization.*

The likelihood function summarizes the whole information about the statistical model which generated the data. In the "frequency" approach, it is supposed that, to do any inference, the statistician has at his disposal only the data and the likelihood function. The situation changes in the "Bayesian" point of view, where the parameter θ becomes a random variable. While important, we do not touch this approach. A brief presentation is given in our textbook [11] on pages 97-103.

We now determine the likelihood function for some known statistical models.

Example 2.14 (Gaussian model). Let us consider a Gaussian random sample with unknown mean and variance: $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu; \sigma^2)$. Then, $\theta = (\mu, \sigma^2)$ is a two dimensional parameter with values in the parameter space $\Theta = \mathbb{R} \times (0, \infty)$. The likelihood function $L_{\mu,\sigma^2}(x_1, \dots, x_n)$ is given by

$$\begin{aligned} L_{\mu,\sigma^2}(x_1, \dots, x_n) &= \prod_{j=1}^{n} f(x_j, \mu, \sigma^2) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{\sum_{j=1}^n (x_j - \mu)^2}{2\sigma^2}\Big\} \\ &= \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{\sum_{j=1}^n (x_j - \overline{x})^2}{2\sigma^2} - \frac{n(\overline{x} - \mu)^2}{2\sigma^2}\Big\} \\ &= \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\overline{x} - \mu)^2}{2\sigma^2}\Big\}, \end{aligned} \tag{2.9}$$

where $s^2 = \sum_{j=1}^n (x_j - \overline{x})^2/(n-1)$. Note that the likelihood function L_{μ,σ^2} depends on the observations only through their sample mean and variance.

Example 2.15 (Bernoulli model). Let $X_1, \dots, X_n$ be i.i.d. random variables with discrete density

$$f(x, \theta) = \theta \mathbf{1}_{\{1\}}(x) + (1 - \theta)\mathbf{1}_{\{0\}}(x) = \theta^x (1 - \theta)^{1-x} \mathbf{1}_{\{0,1\}}(x), \qquad \theta \in \Theta = [0, 1].$$

The likelihood function has the following expression

$$L_\theta(x_1,\dots,x_n) = \prod_{j=1}^n f(x_j,\theta) = \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}$$
$$= \theta^{n\overline{x}}(1-\theta)^{n-n\overline{x}}, \qquad x_1,\dots,x_n = 0,1.$$

Example 2.16 (Exponential model). Let $X_1,\dots,X_n$ be i.i.d. random variables with exponential distribution $\mathcal{E}(\theta)$, $\theta > 0$, parameterized as in (2.4). The likelihood function has the following expression

$$L_\theta(x_1,\dots,x_n) = \prod_{j=1}^n f(x_j,\theta) = \frac{1}{\theta^n}\,\mathrm{e}^{-(1/\theta)\sum_{j=1}^n x_j} = \frac{1}{\theta^n}\,\mathrm{e}^{-\frac{n\overline{x}}{\theta}}, \qquad x_1,\dots,x_n > 0.$$

Note that, in both the Bernoulli and exponential models, L_θ depends on the observations only through the sample realization $\overline{x}$ of the sample mean $\overline{X}$.

Exercise 2.17 (Gamma model). Let $X_1,\dots,X_n$ be a random sample from a population with $\Gamma(a,\beta)$ density, i.e.

$$f(x,a,\beta) = \frac{1}{\beta^a\Gamma(a)}\,x^{a-1}\mathrm{e}^{-x/\beta}\mathbf{1}_{(0,+\infty)}(x), \qquad (a,\beta)\in\Theta = (0,+\infty)^2.$$

Determine the likelihood function.

Example 2.18 (Poisson model). Let $X_1,\dots,X_n$ be a random sample from a population with Poisson density with unknown parameter θ, i.e.,

$$f(x,\theta) = \frac{\mathrm{e}^{-\theta}\theta^x}{x!}\mathbf{1}_{\{0,1,2,\dots\}}(x), \qquad \theta\in\Theta = (0,\infty).$$

Then, for $x_1,\dots,x_n \in \{0,1,2,\dots\}$, the likelihood function is given by

$$L_\theta(x_1,\dots,x_n) = \prod_{j=1}^n f(x_j,\theta) = \frac{\mathrm{e}^{-n\theta}\theta^{\sum_{j=1}^n x_j}}{\prod_{j=1}^n x_j!} = \frac{\mathrm{e}^{-n\theta}\theta^{n\overline{x}}}{\prod_{j=1}^n x_j!}. \tag{2.10}$$

Example 2.19 (Uniform model). Let $X_1,\dots,X_n$ be i.i.d. $\sim f(x,\theta)$ with

$$f(x,\theta) = \frac{1}{\theta}\mathbf{1}_{(0,\theta)}(x), \qquad \theta\in\Theta = (0,\infty).$$

The likelihood function is

$$L_\theta(x_1,\dots,x_n) = \prod_{j=1}^n f(x_j,\theta) = \frac{1}{\theta^n}\mathbf{1}_{(0,\theta)}(x_1)\cdots\mathbf{1}_{(0,\theta)}(x_n), \qquad x_1,\dots,x_n > 0\,.$$

For $x_1,\dots,x_n > 0$, by using the notation $x_{(n)} = \max\{x_1,\dots,x_n\}$, we can write

$$\mathbf{1}_{(0,\theta)}(x_1)\cdots\mathbf{1}_{(0,\theta)}(x_n) = \mathbf{1}_{(0,\theta)}(x_{(n)}) = \mathbf{1}_{(x_{(n)},+\infty)}(\theta).$$

Therefore, we have

$$L_\theta(x_1,\dots,x_n) = \frac{1}{\theta^n}\mathbf{1}_{(x_{(n)},+\infty)}(\theta), \qquad x_1,\dots,x_n > 0\,. \tag{2.11}$$

Note that in this case the likelihood function depends on the observations only through the maximum $x_{(n)}$ of the observations.

Exercise 2.20 (Other uniform models).

(a) Let $X_1,\dots,X_n$ be i.i.d. $\sim f(x,\theta)$ with

$$f(x,\theta) = \frac{1}{|\theta|}\mathbf{1}_{(\theta,0)}(x), \qquad \theta \in \Theta = (-\infty,0).$$

Determine the likelihood function.

(b) Let $X_1,\dots,X_n$ be i.i.d. $\sim f(x,\theta_1,\theta_2)$ with

$$f(x,\theta_1,\theta_2) = \frac{1}{\theta_2-\theta_1}\mathbf{1}_{(\theta_1,\theta_2)}(x), \qquad (\theta_1,\theta_2) \text{ such that } -\infty < \theta_1 < \theta_2 < +\infty.$$

Determine the likelihood function.

2.5 Fréchet-Cramer-Rao inequality

In this section we come back to the problem of finding the optimal estimator. In Section 2.1.2 the mean square error was introduced as a measure of the goodness of an estimator: the smaller the MSE, the better the estimator. In Section 2.2 we noticed that, in the class of unbiased estimators, the MSE reduces to the variance and therefore the problem of minimizing the MSE becomes the problem of minimizing the variance. Therefore, it is natural to ask how small can the variance be, or, more precisely, if there exists a (non trivial) lower bound for the variance of an unbiased estimator, depending only on the characteristic $\kappa(\theta)$ to be estimated and on the statistical model through the likelihood function. If we exhibit such a lower bound and we find an estimator whose variance reaches this bound, then this estimator is the UMVU estimator. This section is devoted to show that, under some regularity conditions, the variance of an unbiased estimator is actually limited from below and that, in some cases, it is possible to find an estimator whose variance reaches the lower bound. For simplicity, we consider the case where the parameter space Θ is one-dimensional.

Theorem 2.3. *Let $X_1,\dots,X_n$ be i.i.d. random variables with density $f(x,\theta)$, $\theta\in\Theta\subset\mathbb{R}$ and let $T = g(X_1,..,X_n)$ be an unbiased estimator of $\kappa(\theta)$ with finite variance. Let us assume that the following regularity conditions are satisfied:*

(i) Θ is an open interval of $\mathbb{R}$;

(ii) $S = \{x:\ f(x,\theta) > 0\}$ is independent of θ;

(iii) $\theta \mapsto f(x,\theta)$ *is differentiable on* Θ, $\forall x \in S$*;*

(iv) $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right] = 0, \qquad \forall\theta\in\Theta$*;*

(v) $0 < \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\Big)^2\Big] < \infty, \qquad \forall\theta\in\Theta$*;*

(vi) $\kappa : \Theta \to \mathbb{R}$ *is differentiable on* Θ*;*

(vii)

$$\frac{\partial}{\partial\theta}\,\mathbb{E}_\theta[T] = \mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n)\right], \qquad \forall\theta\in\Theta. \tag{2.12}$$

Then, we have

$$\mathrm{Var}_\theta[T] \geq \frac{(\kappa'(\theta))^2}{nI(\theta)}, \qquad \forall\theta\in\Theta, \tag{2.13}$$

where

$$I(\theta) := \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\Big)^2\Big]. \tag{2.14}$$

Moreover, equality in (2.13) *holds if and only if there exists a function* $a(n,\theta)$ *such that*

$$P_\theta\left[\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n) = a(n,\theta)(T-\kappa(\theta))\right] = 1, \qquad \forall\theta\in\Theta. \tag{2.15}$$

Hypotheses (i), (ii), (iii), (v) and (vi) are apparently regularity assumptions. The other hypotheses seem to be more demanding, but this is not the case. Let us explain this. We elaborate only the continuous case, but the discrete case is similar.

Remark 2.21. Let us start from the normalization condition

$$1 = \int_{\mathbb{R}} f(x,\theta)\,\mathrm{d}x = \int_S f(x,\theta)\,\mathrm{d}x\,.$$

By differentiation with respect to θ of both sides in this equation we get

$$0 = \frac{\partial}{\partial\theta}\int_S f(x,\theta)\,\mathrm{d}x.$$

If S does not depend on θ and f is "regular enough" to allow for the interchange of derivation and integration we have

$$\frac{\partial}{\partial\theta}\int_S f(x,\theta)\,\mathrm{d}x = \int_S \frac{\partial}{\partial\theta}f(x,\theta)\,\mathrm{d}x. \tag{2.16}$$

By using $f(x,\theta)\frac{\partial}{\partial\theta}\log f(x,\theta) = \frac{\partial f(x,\theta)}{\partial\theta}$, and the fact that, for any function h,

$$\mathbb{E}_\theta[h(X_1)] = \int_{\mathbb{R}} h(x)f(x,\theta)\,\mathrm{d}x = \int_S h(x)f(x,\theta)\,\mathrm{d}x\,,$$

we get

$$0=\int_S \frac{\partial f(x,\theta)}{\partial\theta}\,\mathrm{d}x=\int_S\Big(\frac{\partial}{\partial\theta}\log f(x,\theta)\Big)f(x,\theta)\,\mathrm{d}x=\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta}\log f(X_1,\theta)\Big].$$

Therefore, hypothesis (*iv*) is nothing but the regularity assumption (2.16).

Remark 2.22. Also the hypothesis (*vii*) is nothing but a request of interchange of derivative and integral. Let us consider the unbiasedness condition

$$\kappa(\theta)=\mathbb{E}_\theta\Big[g(X_1,\dots,X_n)\Big]=\int_{S^n} g(x_1,\dots,x_n)\Big(\prod_{j=1}^n f(x_j,\theta)\Big)\,\mathrm{d}x_1\cdots\mathrm{d}x_n\,.$$

The characteristic $\kappa(\theta)$ is differentiable by assumption and we get

$$\frac{\partial}{\partial\theta}\kappa(\theta)=\frac{\partial}{\partial\theta}\int_{S^n} g(x_1,\dots,x_n)\Big(\prod_{j=1}^n f(x_j,\theta)\Big)\,\mathrm{d}x_1\cdots\mathrm{d}x_n\,.$$

Assuming that g and f are "regular enough" to allow for the interchange of derivation and integration, it follows that

$$\begin{aligned}\frac{\partial}{\partial\theta}\kappa(\theta)&=\int_{S^n} g(x_1,\dots,x_n)\Big[\frac{\partial}{\partial\theta}\prod_{j=1}^n f(x_j,\theta)\Big]\,\mathrm{d}x_1\cdots\mathrm{d}x_n\\&=\int_{S^n} g(x_1,\dots,x_n)\Big[\frac{\partial}{\partial\theta}\log\prod_{j=1}^n f(x_j,\theta)\Big]\Big(\prod_{i=1}^n f(x_i,\theta)\Big)\mathrm{d}x_1\cdots\mathrm{d}x_n\\&=\int_{S^n} g(x_1,\dots,x_n)\Big[\frac{\partial}{\partial\theta}\log L_\theta(x_1,\dots,x_n)\Big]\Big(\prod_{i=1}^n f(x_i,\theta)\Big)\mathrm{d}x_1\cdots\mathrm{d}x_n\\&=\int_{\mathbb{R}^n} g(x_1,\dots,x_n)\Big[\frac{\partial}{\partial\theta}\log L_\theta(x_1,\dots,x_n)\Big]\Big(\prod_{i=1}^n f(x_i,\theta)\Big)\mathrm{d}x_1\cdots\mathrm{d}x_n\\&=\mathbb{E}_\theta\Big[g(X_1,\dots,X_n)\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n)\Big].\end{aligned}$$

Remark 2.23. The proof of the theorem needs some properties of covariance, which we recall here. The *covariance* of two random variables X,Y, both with finite variance, is defined as

$$\operatorname{Cov}[X,Y]=\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]=\mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y].$$

Note that $\mathbb{E}[XY]=\operatorname{Cov}(X,Y)$ when $\mathbb{E}[Y]=0$.

The following properties of covariance hold:

(*j*) $|\operatorname{Cov}[X,Y]|\le\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}$,

(*jj*) $|\operatorname{Cov}[X,Y]|=\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}$ if and only if there exist $a,b\in\mathbb{R}$ such that $P[Y=aX+b]=1$. Moreover, if $\operatorname{Var}[X]>0$ then $a=\operatorname{Cov}[X,Y]/\operatorname{Var}[X]$ and $b=\mathbb{E}[Y]-a\,\mathbb{E}(X)$, *(jjj)* If $\mathbb{E}[Y]=0$ then $b=-a\,\mathbb{E}[X]$ and $P[Y=a(X-\mathbb{E}[X])]=1$.

Proof of Theorem 2.3. For the sake of simplicity, we introduce the variables $Y_1,\ldots,Y_n$ defined as

$$Y_j=\frac{\partial}{\partial\theta}\log f(X_j,\theta),\quad \forall j=1,\ldots,n.$$

The random variables $Y_1,\ldots,Y_n$ are i.i.d., with mean zero and finite variance $I(\theta)$, $\forall\theta$. In fact,

$$\mathbb{E}_\theta[Y_j]=\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_j,\theta)\right]=0 \qquad \text{[by hypothesis } (iv)]$$

$$\mathrm{Var}_\theta[Y_j]=\mathbb{E}_\theta[{Y_j}^2]=\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X_j,\theta)\right)^2\right]=I(\theta)\in(0,\infty) \quad \text{[by hypothesis } (v)].$$

Moreover, we have

$$\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)=\frac{\partial}{\partial\theta}\log\prod_{j=1}^n f(X_j,\theta)=\sum_{j=1}^n\frac{\partial}{\partial\theta}\log f(X_j,\theta)=\sum_{j=1}^n Y_j\,,$$

which implies

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right]=\sum_{j=1}^n\mathbb{E}_\theta[Y_j]=0 \tag{2.17}$$

and

$$\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right]=\sum_{j=1}^n\mathrm{Var}_\theta[Y_j]=nI(\theta). \tag{2.18}$$

Hypotheses (vi), (vii) and Eq. (2.17) together imply

$$\kappa'(\theta)=\mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right]=\mathrm{Cov}\left[T,\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right].$$

Hence

$$\begin{aligned}(\kappa'(\theta))^2&=\left(\mathrm{Cov}\left[T,\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right]\right)^2\\&\le\mathrm{Var}_\theta[T]\ \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right]\\&=\mathrm{Var}_\theta[T]\,nI(\theta)\qquad\text{[by property } (j) \text{ of covariance and Eq. (2.18)]}\end{aligned}$$

Inequality (2.13) then follows by hypothesis (v).

Finally, in order to verify that condition (2.15) is necessary and sufficient for equality (2.13), apply property (jjj) of the covariance to the random variables $X=T$ and $Y=\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)$ for any θ and for any n. □

Definition 2.12. $I(\theta) = \mathbb{E}_\theta\Big[\big(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\big)^2\Big]$ is called the *Fisher information on* θ in a single observation X_1 and $I_n(\theta) = \mathbb{E}_\theta\Big[\big(\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n)\big)^2\Big]$ is called the *Fisher information on* θ in the whole sample.

In the proof of Theorem 2.3 we showed that $I_n(\theta) = nI(\theta)$, in other words, when we deal with a random sample, the Fisher information in the whole sample is n times the Fisher information in the single observation.

Theorem 2.3 shows that, under certain regularity conditions, the variance of any unbiased estimator of θ is not less than the reciprocal of the Fisher information in the sample. As the Fisher information gets bigger, we have a smaller bound on the variance of an unbiased estimator, and therefore a more accurate estimate for θ. This consideration motivates the name "information".

Another important point is that the square root of the lower bound goes as $1/\sqrt{n}$: **"in the usual conditions, the errors go as $1/\sqrt{n}$"**.

Definition 2.13. An unbiased estimator T^* whose variance reaches the Fréchet-Cramer-Rao lower bound is called *efficient.*

If there exists an efficient estimator T, obviously *it is an UMVUE.*

Example 2.24. Let $X_1,\dots,X_n$ be random variables i.i.d. $\sim \mathcal{E}(\theta)$, $\theta > 0$, where $\mathcal{E}(\theta)$ denotes the exponential distribution parameterized as in (2.4). Conditions (i)-(iii) of Theorem 2.3 are satisfied since $S = \{x : f(x,\theta) > 0\} = (0,\infty)$ and $f(x,\theta) = \mathrm{e}^{-x/\theta}/\theta$ on S. Since

$$\frac{\partial}{\partial\theta}\log f(X_1,\theta) = -\frac{1}{\theta} + \frac{X_1}{\theta^2}$$

then

$$\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta}\log f(X_1,\theta)\Big] = \mathbb{E}_\theta\Big[-\frac{1}{\theta} + \frac{X_1}{\theta^2}\Big] = -\frac{1}{\theta} + \frac{\theta}{\theta^2} = 0,$$

and, $\forall\theta > 0$,

$$I(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\Big)^2\Big] = \mathrm{Var}_\theta\Big[-\frac{1}{\theta} + \frac{X_1}{\theta^2}\Big] = \frac{\mathrm{Var}_\theta[X_1]}{\theta^4} = \frac{1}{\theta^2} \in (0,\infty),$$

therefore (iv) and (v) are also satisfied. The Fréchet-Cramer-Rao lower bound for $\kappa_1(\theta) = \theta$ is $\dfrac{(\kappa_1'(\theta))^2}{nI(\theta)} = \dfrac{\theta^2}{n}$ which coincides with the variance of $\overline{X}$. We already know that $\overline{X}$ is an unbiased estimator of θ. Moreover,

$$\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n) = \sum_{j=1}^{n}\Big(-\frac{1}{\theta} + \frac{X_j}{\theta^2}\Big) = -\frac{n}{\theta} + \frac{n\overline{X}}{\theta^2}$$

hence

$$\mathbb{E}_\theta\Big[\overline{X}\Big(\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n)\Big)\Big] = -n + \frac{n}{\theta^2}\Big(\frac{\theta^2}{n} + \theta^2\Big) = 1 = \kappa_1'(\theta),$$

which proves that the estimator $\overline{X}$ satisfies hypothesis (vi) and (vii). Therefore, we can conclude that the sample mean is an efficient estimator for the characteristic "mean" in the exponential model.

If we want instead to estimate the characteristic "variance" θ^2, then the Fréchet-Cramer-Rao lower bound for $\kappa_2(\theta) = \theta^2$ is $\dfrac{(2\theta)^2}{nI(\theta)} = \dfrac{4\theta^4}{n}$. We know from the second part of Theorem 2.3 that such a bound is reachable by an unbiased estimator T if and only if there exists $a(n, \theta)$ such that

$$P_\theta \left[-\frac{n}{\theta} + \frac{n\overline{X}}{\theta^2} = a(n,\theta)(T - \theta^2) \right] = 1.$$

But this equation is satisfied if and only if $a(n, \theta) = n/\theta^3$ and simultaneously $a(n, \theta) \propto b(n)/\theta^2$, which is impossible. Therefore we can conclude that efficient estimators of the variance θ^2 do not exist. We only mention here without proof that $\dfrac{(n+1)\overline{X}^2}{n}$ is the UMVUE of θ^2.

Exercise 2.25. Let $X_1, \ldots, X_n$ be a random sample from a population with Poisson distribution with unknown parameter θ.

1. Prove that $\overline{X}$ is an efficient estimator of the mean.

2. Prove that efficient estimators of $\kappa(\theta) = P_\theta[X_1 = 0]$ do not exist.

When the hypotheses of Theorem 2.3 hold, the variance of the unbiased estimator T is not less than something proportional to $1/n$. When those hypotheses do not hold, we cannot conclude anything. In some cases we can even do something better than a $1/n$ bound, as the following example shows.

Example 2.26. Let $X_1, \ldots, X_n$ be i.i.d. $\sim \mathcal{U}(0, \theta)$, $\theta > 0$. In this case the density is

$$f(x,\theta) = \frac{1}{\theta}\,\mathbf{1}_{[0,\theta]}(x) = \frac{1}{\theta}\,\mathbf{1}_{[0,+\infty)}(x)\,\mathbf{1}_{[x,+\infty)}(\theta).$$

We have $S = [0, \theta]$ and, so, hypothesis (ii) is violated. Moreover, $f(x, \theta)$ is not θ-differentiable in the point $\theta = x$ and hypothesis (iii) is violated, too.

Let $X_{(n)} = \max\{X_1, \ldots, X_n\}$ denote the maximum observation and set $T = (n + 1)X_{(n)}/n$; the c.d.f. of $X_{(n)}$ is

$$\begin{aligned} F_{X_{(n)}}(x,\theta) = P_\theta[X_{(n)} \le x] = P_\theta[X_1 \le x, X_2 \le x, \cdots, X_n \le x] = \prod_{j=1}^{n} P_\theta[X_j \le x] \\ = (P_\theta[X_1 \le x])^n = (F_{X_1}(x,\theta))^n = \begin{cases} 0 & \text{if } x \le 0 \\ \left(\dfrac{x}{\theta}\right)^n & \text{if } 0 < x < \theta \\ 1 & \text{if } x \ge \theta. \end{cases} \end{aligned}$$

Therefore,

$$\frac{\partial F_{X_{(n)}}(x,\theta)}{\partial x} = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{if } 0 < x < \theta \\ 0 & \text{if } x \le 0 \text{ or } x \ge \theta \end{cases}$$

and, hence, the density of $X_{(n)}$ is

$$f_{X_{(n)}}(x,\theta) = \frac{nx^{n-1}}{\theta^n}\mathbf{1}_{(0,\theta)}(x).$$

Therefore,

$$\mathbb{E}_\theta[X_{(n)}] = \int_0^\theta x\,\frac{nx^{n-1}}{\theta^n}\,\mathrm{d}x = \frac{n\theta}{n+1}$$

and

$$\begin{aligned}\mathrm{Var}_\theta[X_{(n)}] = \mathbb{E}_\theta[X_{(n)}^2] - \left(\frac{n\theta}{n+1}\right)^2 &= \int_0^\theta x^2\,\frac{nx^{n-1}}{\theta^n}\,\mathrm{d}x - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}\,.\end{aligned}$$

It follows that T is an unbiased estimator of θ and its variance is

$$\mathrm{Var}_\theta[T] = \frac{(n+1)^2}{n^2} \times \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}.$$

For large n the MSE of the unbiased estimator T behaves as $1/n^2$ and not as $1/n$, which is the “usual” behaviour.

2.5.1 Exponential family of distributions

Example 2.24 showed that in the exponential model the UMVU estimator of the parameter θ^2 exists but it is not efficient. Non efficiency of the estimator was proved by using the second part of Theorem 2.3. The fact that an UMVUE does not exist or that it exists but it is not efficient is not exceptional. In any case, in this section, we point our attention on the second part of Theorem 2.3 to find a characterization of those models for which an efficient estimator exists.

Let us consider Eq. (2.15) and write T_n instead of T in order to underline all n dependencies; recall that $T_n = g_n(X_1,\ldots,X_n)$. Then, Eq. (2.15) says that, with probability 1, for all θ in the interval Θ,

$$\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n) = a(n,\theta)\big(T_n - \kappa(\theta)\big).$$

By integrating this equation from $\theta_0 \in \Theta$ to $\theta \in \Theta$, we get

$$\log\frac{L_\theta(X_1,\ldots,X_n)}{L_{\theta_0}(X_1,\ldots,X_n)} = \int_{\theta_0}^\theta a(n,u)\,[g_n(X_1,\ldots,X_n) - \kappa(u)]\,\mathrm{d}u\,.$$

But the likelihood function is the joint density of the random sample, so we have

$$\sum_{j=1}^{n} \log \frac{f(x_j,\theta)}{f(x_j,\theta_0)} = g_n(x_1,\dots,x_n) \int_{\theta_0}^{\theta} a(n,u)\,\mathrm{d}u - \int_{\theta_0}^{\theta} a(n,u)\kappa(u)\,\mathrm{d}u\,. \tag{2.19}$$

This equations implies that g_n has to be the sum of n terms, each of them depending only on one of the x_j's. So, we set

$$g_n(x_1,\dots,x_n) = \frac{1}{n}\sum_{j=1}^{n} g(x_j)\,, \qquad a(n,u) = na(u)\,,$$

$$C(x) = f(x,\theta_0)\,, \qquad A(\theta) = \int_{\theta_0}^{\theta} a(u)\,\mathrm{d}u\,, \qquad B(\theta) = -\int_{\theta_0}^{\theta} a(u)\kappa(u)\,\mathrm{d}u\,,$$

and we apply the exponential function to both sides of Eq. (2.19). Then, we get

$$\prod_{j=1}^{n} f(x_j,\theta) = \prod_{j=1}^{n} C(x_j)\exp\{A(\theta)g(x_j) + B(\theta)\}\,.$$

This means that, from the condition (2.15) which says when the FCR-bound is reached, we have obtained the possible form of the density. All the steps can be made rigorous and also the converse can be proved. These results are summarized in the next Definition and Theorem.

Definition 2.14. A family of densities $\{f(x,\theta), \theta \in \Theta \subset \mathbb{R}\}$, where

$$f(x,\theta) = C(x)\exp\{A(\theta)g(x) + B(\theta)\} \tag{2.20}$$

is said to be a *one parameter exponential family.*

Note that the support S introduced in hypothesis (ii) of Theorem 2.3 is determined by $C(x)$, which does not depend on θ, as it should be.

Theorem 2.4. *An efficient estimator exists if and only if the statistical model has a density of the form* (2.20)*, $A(\theta)$ has a continuous nonvanishing derivative on Θ and the characteristic to be estimated is of type $\kappa(\theta) = -cB'(\theta)/A'(\theta) + d$. Moreover this estimator can be expressed as $(c/n)\sum_{j=1}^{n} g(X_j) + d$, where $c, d \in \mathbb{R}$.*

Example 2.27 (from [16] exercise 2.6 page 44). Let $X_1,\dots,X_n$ be i.i.d. $\sim f(x,\theta)$ with

$$f(x,\theta) = \frac{x+1}{\theta(\theta+1)}\,\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x), \qquad \theta > 0.$$

Find an efficient estimator of the characteristic $\kappa(\theta) = (3+2\theta)(2+\theta)/(\theta+1)$.

In order to use Eq. (2.15), let us observe that, for $x_1 > 0,\dots,x_n > 0$,

$$\log L_\theta(x_1,\dots,x_n) = \log\prod_{i=1}^{n} f(x_i,\theta) = \log\prod_{i=1}^{n}(x_i+1) - n\log[\theta(\theta+1)] - \frac{1}{\theta}\sum_{j=1}^{n} x_j\,,$$

hence,

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L_\theta(x_1,\dots,x_n) &= -n\,\frac{2\theta+1}{\theta(\theta+1)}+\frac{\sum_{j=1}^n x_j}{\theta^2}\\ &= \frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n x_j}{n}-\frac{\theta(2\theta+1)}{\theta+1}\right)=\frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n(x_j+6)}{n}-\kappa(\theta)\right). \qquad (2.21)\end{aligned}$$

The last equality follows from $\kappa(\theta)-6=\theta(2\theta+1)/(\theta+1)$. Equation (2.21) shows that Eq. (2.15) holds true if we put $a(n,\theta)=\frac{n}{\theta^2}$ and $T=\sum_{j=1}^n(X_j+6)/n$. Moreover,

$$\mathbb{E}_\theta\left[\frac{\sum_{j=1}^n(X_j+6)}{n}\right]=\mathbb{E}_\theta[X_1]+6=\int_0^\infty\frac{x(x+1)}{\theta(\theta+1)}\,\mathrm{e}^{-x/\theta}\,\mathrm{d}x+6=\kappa(\theta)$$

which shows that T is unbiased.

From Theorem 2.3 and (2.21) we conclude that $T=(1/n)\sum_{j=1}^n(X_j+6)$ is the efficient estimator of $\kappa(\theta)$.

Note that this is in line with what we just mentioned about efficient estimators in the case of exponential families of distributions. In fact, $f(x,\theta)$ belongs to the exponential family $\{f(x,\theta),\theta\in(0,\infty)\}$ with $g(x)=x$, $A(\theta)=-1/\theta$, $B(\theta)=-\log(\theta(\theta+1))$ and $C(x)=(x+1)\,\mathbf{1}_{(0,+\infty)}(x)$. Moreover, $-B'(\theta)/A'(\theta)+6=\kappa(\theta)$.

2.5.2 Fréchet-Cramer-Rao inequality in the multivariate case

In this section we briefly treat the Fréchet-Cramer-Rao inequality in the case that the parameter θ is m-dimensional, i.e. $\theta=\begin{pmatrix}\theta_1\\ \vdots\\ \theta_m\end{pmatrix}$ with $m\ge 2$. In the m-dimensional case, the Fisher information is an $m\times m$ matrix and, in place of the variance of a univariate estimator, the covariance matrix of an m-dimensional unbiased estimator of the parameter θ is considered: $T=\begin{pmatrix}T_1\\ \vdots\\ T_m\end{pmatrix}$ is an m-dimensional statistics such that $\mathbb{E}_\theta[T_j]=\theta_j,\ \forall j=1,\dots,m$.

Regularity conditions (i)-(vii) are easily generalized to the m-dimensional settings. In place of the first derivative $\frac{\partial}{\partial\theta}\log f(X_1,\theta)$, the vector of the first partial derivatives $\begin{pmatrix}\frac{\partial}{\partial\theta_1}\log f(X_1,\theta)\\ \vdots\\ \frac{\partial}{\partial\theta_m}\log f(X_1,\theta)\end{pmatrix}$ is considered and finally the *Fisher information matrix* $I(\theta)$ is defined as the covariance matrix of this vector. If C_T denotes the covariance matrix of the vector T, the analogue of inequality (2.13) is

$$C_T-\frac{1}{n}\,I(\theta)^{-1}\ge 0\,,$$

in the sense that the left hand side matrix is semi-defined positive.

The extension of Fréchet-Cramer-Rao inequality to the m-dimensional case is useful, for example, to investigate the efficiency of the estimators $\overline{X}, S^2$ of the mean and the variance of a normal population. In this case, the two-dimensional Fisher information is

$$I((\mu,\sigma^2)) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix},$$

whose inverse is

$$I^{-1}((\mu,\sigma^2)) = 2\sigma^6 \begin{bmatrix} 1/(2\sigma^4) & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix},$$

while the covariance matrix of $T = \begin{pmatrix} \overline{X} \\ S^2 \end{pmatrix}$ is

$$C_{(\overline{X},S^2)} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{bmatrix}.$$

By comparing, we see that $\overline{X}$ is efficient for estimating the mean, while S^2 is not efficient for estimating σ^2. Anyway, it can be proved that S^2 is the UMVU estimator of σ^2.

2.6 Methods of finding estimators

In this section we detail two classical methods of finding estimators: the method of moments and the method of maximum likelihood.

2.6.1 Method of moments

The method of moments is, perhaps, the oldest method of finding estimators, dating back to Karl Pearson in 1894 [10].

Let $X_1, \dots, X_n$ be i.i.d. random variables with density $f(x,\theta)$ where $\theta = (\theta_1, \dots, \theta_m) \in \Theta \subset \mathbb{R}^m$.

Suppose that $\mathbb{E}_\theta[X_1], \mathbb{E}_\theta[X_1^2] \dots, \mathbb{E}_\theta[X_1^m]$ do exist;[3] the quantity $\mu_r(\theta) := \mathbb{E}_\theta[X_1^r]$ is called *moment of order* r. Let us introduce the m statistics $M_1, \dots, M_m$ given by the m *sample moments*

$$M_r = \frac{1}{n}\sum_{j=1}^{n} X_j^r, \qquad r = 1, \dots, m.$$

[3] In the absolutely continuous case, if $\int_{\mathbb{R}} |x^r| f(x,\theta)\,\mathrm{d}x < \infty$, then $\mathbb{E}_\theta[X_1^r]$ exists and $\mathbb{E}_\theta[X_1^r] = \int_{\mathbb{R}} x^r f(x,\theta)\,\mathrm{d}x$. In the discrete case the analogous considerations hold, with integrals substituted by sums.

In particular, the first sample moment is the sample mean: $M_1 = \overline{X}$. Let us now consider the system of m equations

$$\begin{cases} \mu_1(\theta) = M_1 \\ \vdots \\ \mu_m(\theta) = M_m \end{cases} \tag{2.22}$$

in the m unknowns $\theta_1, \dots, \theta_m$. If the system (2.22) admits solutions $\widehat{\theta_1}, \dots, \widehat{\theta_m}$, then $\widehat{\theta_1}, \dots, \widehat{\theta_m}$ are statistics, in that they depend only on the sample moments $M_1, \dots, M_m$. Hence, we can use $\widehat{\theta_1}, \dots, \widehat{\theta_m}$ as estimators of $\theta_1, \dots, \theta_m$, respectively:

$\widehat{\theta} = (\widehat{\theta_1}, \dots, \widehat{\theta_m})$ *is, by definition, the estimator of* $\theta = (\theta_1, \dots, \theta_m)$ *found through the method of moments, or, briefly, the method of moments estimator.*

Obviously,

if $\mu_r(\theta)$ *exists, then* M_r *is an unbiased estimator of* $\kappa(\theta) = \mu_r(\theta)$.

Indeed,

$$\mathbb{E}_\theta[M_r] = \mathbb{E}_\theta\left[\frac{1}{n}\sum_{j=1}^n X_j^r\right] = \frac{1}{n}\sum_{j=1}^n \mathbb{E}_\theta[X_j^r] = \mu_r(\theta).$$

Example 2.28 (Gaussian model). Let $X_1, \dots, X_n$ be i.i.d. random variables $\sim \mathcal{N}(\mu;\sigma^2)$, $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Let us find the method of moments estimators of μ and σ^2. Since $\mu_1(\mu, \sigma^2) = \mu$ and $\mu_2(\mu, \sigma^2) = \mathbb{E}[X_1^2] = \mathrm{Var}[X_1] + \big(\mathbb{E}[X_1]\big)^2 = \sigma^2 + \mu^2$, then $\widehat{\mu}, \widehat{\sigma^2}$ are the solutions of the following system of equations

$$\begin{cases} \mu = M_1 \equiv \overline{X}\,, \\ \sigma^2 + \mu^2 = M_2\,. \end{cases}$$

We obtain

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\sigma^2} = M_2 - M_1^2 = \frac{1}{n}\sum_{j=1}^n \big(X_j - \overline{X}\big)^2 = \frac{(n-1)S^2}{n}\,.$$

Note that the method of moments estimator of the variance is biased, in that $\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{n-1}{n}\sigma^2$.

Example 2.29 (Uniform models).

(*a*) Let $X_1, \dots, X_n$ be i.i.d. $\sim \mathcal{U}(0, \theta)$, with $\theta > 0$. Since $\mathbb{E}[X_1] = \theta/2$, then the method of moments estimator of θ is $\widehat{\theta} = 2\overline{X}$.

(*b*) Let $X_1, \dots, X_n$ be i.i.d. $\sim \mathcal{U}(\theta_1, \theta_2)$, with $-\infty < \theta_1 < \theta_2 < \infty$. Since $\mathbb{E}[X_1] = (\theta_1 + \theta_2)/2$ and $\mathrm{Var}[X_1] = (\theta_2 - \theta_1)^2/12$, then $\widehat{\theta_1}, \widehat{\theta_2}$ are the solutions of the following system of equations

$$\begin{cases} \frac{\theta_1+\theta_2}{2} = M_1 \\ \frac{(\theta_2-\theta_1)^2}{12} + \frac{(\theta_2+\theta_1)^2}{4} = M_2 \end{cases} \Leftrightarrow \begin{cases} \frac{\theta_1+\theta_2}{2} = M_1 \\ (\theta_2 - \theta_1)^2 = 12(M_2 - M_1^2) = \frac{12(n-1)S^2}{n} \end{cases}$$

Considering the constrain $\theta_2 > \theta_1$, the system has a unique solution given by

$$\begin{cases} \widehat{\theta}_1 = \overline{X} - \sqrt{\frac{3(n-1)S^2}{n}} \\ \widehat{\theta}_2 = \overline{X} + \sqrt{\frac{3(n-1)S^2}{n}}. \end{cases}$$

We now show a serious drawback of the method of moments.

Example 2.30. Let $X_1, \ldots, X_n$ be i.i.d. $\sim \mathcal{U}(-\theta, \theta)$, with $\theta > 0$. Then $\mathbb{E}_\theta[X_1] = 0$, $\forall \theta$. This makes it impossible to solve the first moment equation $\mathbb{E}_\theta[X_1] = \overline{X}$. We could then decide to estimate θ by passing to the second moment equation

$$\mathbb{E}_\theta[X_1^2] \equiv \frac{\theta^2}{3} = M_2 \,,$$

which gives $\widehat{\theta} = \sqrt{3M_2}$. But this choice is arbitrary and with the same degree of arbitrariness we could decide to use the $2r$-th moment equation, obtaining

$$\mathbb{E}_\theta[X_1^{2r}] \equiv \frac{\theta^{2r}}{2r+1} = M_{2r}$$

for $r = 2, 3, \ldots$. For any r we obtain a different estimator of θ, that is, the method does not supplies a unique solution. In this case Rohatgi and Saleh [13] suggest to choose those estimators obtained by solving the moment equations of the lowest order which renders the system solvable.

2.6.2 Method of maximum likelihood

Another method of finding estimators with nice properties is the method of maximum likelihood, introduced by Ronald Fisher in 1921.

Let $X_1, \ldots, X_n$ be i.i.d. random variables with density $f(x, \theta)$ where $\theta = (\theta_1, \ldots, \theta_m)$ belongs to $\Theta \subset \mathbb{R}^m$ and let $\theta \mapsto L_\theta$ be the likelihood function.

For simplicity, let us first suppose that $f(x, \theta)$ is a discrete density. In this case, if θ is known to be θ_0, then $L_{\theta_0}(x_1, \ldots, x_n) = P_{\theta_0}(X_1 = x_1, \ldots, X_n = x_n)$ is the probability to observe $x_1, \ldots, x_n$. If instead, as in our case, θ is unknown and $x_1, \ldots, x_n$ is the observed sample, then it seems reasonable to select the value θ in Θ which identifies the density $f(x, \theta)$ that more likely generated the observed sample. In other words, it seems reasonable to choose the value $\widehat{\theta}$ that maximizes the probability to observe what we just observed. Formally, we choose $\widehat{\theta}$ such that $L_{\widehat{\theta}}(x_1, \ldots, x_n) = \max_{\theta \in \Theta} L_\theta(x_1, \ldots, x_n)$. If such $\widehat{\theta}$ exists, it depends only on $x_1, \ldots, x_n$ and therefore it can be taken as an estimate of θ. What just outlined above suggests for $\widehat{\theta}$ the name of maximum likelihood estimate of θ.

We now give the formal definition of the maximum likelihood estimator, that covers also the case of absolutely continuous statistical models.

Definition 2.15. Let $X_1, \ldots, X_n$ be a random sample, L_θ the likelihood function, $x_1, \ldots, x_n$ the observed sample and $g(x_1, \ldots, x_n)$ be a point in Θ such that

$$L_{g(x_1,\ldots,x_n)}(x_1, \ldots, x_n) = \max_{\theta \in \Theta} L_\theta(x_1, \ldots, x_n).$$

The statistics $\widehat{\theta} = g(X_1, \dots, X_n)$ is said to be a *maximum likelihood estimator of* θ. We will use the acronym MLE for *Maximum Likelihood Estimator.*

It is possible that the likelihood function has no maximum or that it has more than one maximum; therefore it is possible that there aren't maximum likelihood estimators, or that the maximum likelihood estimator is not unique. The maximum are to be found by the usual calculus techniques. In particular, note that the likelihood function L_θ and its logarithm $\log L_\theta$ have the same maximum points (where L_θ is null, $\log L_\theta = -\infty$).

If Θ is an interval and $\log L_\theta(x_1, \dots, x_n)$ is differentiable in Θ, then the maximum points are to be found among those for which the partial derivative $\frac{\partial}{\partial\theta}\log(L_\theta)$ is null. Among these points, the relative maxima are to be selected and compared with the values in the extreme points of the interval. Analogous considerations hold if θ is a vector in $\Theta \subset \mathbb{R}^m$: in this case we have to solve the system of *likelihood equations*

$$\frac{\partial}{\partial\theta_j}\log L_\theta(x_1, \dots, x_n) = 0\,, \qquad j = 1, \dots, m.$$

If we want to estimate a function $\kappa(\theta)$ instead of θ itself, we can consider the *induced likelihood function* defined as

$$L^*_\lambda(x_1, \dots, x_n) := \sup_{\{\theta\in\Theta\,:\,\kappa(\theta)=\lambda\}} L_\theta(x_1, \dots, x_n).$$

Here $\lambda \in \Lambda$, the set of all possible values of $\kappa(\theta)$, i.e. $\Lambda = \{\lambda \in \mathbb{R} : \lambda = \kappa(\theta) \text{ for some } \theta \in \Theta\}$. It is natural to call *maximum likelihood estimator of* $\kappa(\theta)$ a statistics that maximizes L^*_λ.

If $\hat\theta(x_1, \dots, x_n)$ denotes a MLE of θ, we have

(a) $\sup_{\{\theta\in\Theta\,:\,\kappa(\theta)=\lambda\}} L_\theta(x_1, \dots, x_n) \le \sup_{\theta\in\Theta} L_\theta(x_1, \dots, x_n)$, because $\{\theta \in \Theta \;:\; \kappa(\theta) = \lambda\}$ is a subset of Θ,

(b) $L_{\hat\theta}(x_1, \dots, x_n) = \sup_{\{\theta\in\Theta\,:\,\kappa(\theta)=\kappa(\hat\theta(x_1,\dots,x_n))\}} L_\theta(x_1, \dots, x_n)$, because $\hat\theta$ maximises L_θ and $\hat\theta \in \{\theta \in \Theta \;:\; \kappa(\theta) = \kappa(\hat\theta(x_1, \dots, x_n))\}$.

Then, we have

$$\begin{aligned} L^*_\lambda(x_1, \dots, x_n) &= \sup_{\{\theta\in\Theta\,:\,\kappa(\theta)=\lambda\}} L_\theta(x_1, \dots, x_n) \\ &\le \sup_{\theta\in\Theta} L_\theta(x_1, \dots, x_n) = L_{\hat\theta}(x_1, \dots, x_n) \\ &= \sup_{\{\theta\in\Theta\,:\,\kappa(\theta)=\kappa(\hat\theta(x_1,\dots,x_n))\}} L_\theta(x_1, \dots, x_n) = L^*_{\kappa(\hat\theta)}(x_1, \dots, x_n). \end{aligned} \tag{2.23}$$

This says that, if $\hat\theta(x_1, \dots, x_n)$ is a MLE of θ, then $\kappa(\hat\theta)$ is a MLE of $\kappa(\theta)$. This property is known as the *invariance property of the maximum likelihood estimators.* In brief, this property says that, once we have a MLE of θ, we have a MLE of any function of θ.

Remark 2.31 (Invariance property of MLE). Let $\kappa(\theta)$ be a population characteristic and let $\widehat{\theta}(X_1,\dots,X_n)$ be a MLE of θ. Then, $\kappa\big(\widehat{\theta}(X_1,\dots,X_n)\big)$ is a maximum likelihood estimator of $\kappa(\theta)$.

Example 2.32 (Gaussian model, unknown mean and variance). Let $X_1,\dots,X_n$ be i.i.d. $\sim \mathcal{N}(\mu;\sigma^2)$, $(\mu,\sigma^2)\in\Theta=\mathbb{R}\times(0,\infty)$. Let us estimate both μ and σ^2 by the method of maximum likelihood. By (2.9) the likelihood function is

$$L_{\mu,\sigma^2}(x_1,\dots,x_n)=\Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}}\exp\Big\{-\frac{(n-1)s^2}{2\sigma^2}-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\Big\}$$

where $s^2=\sum_{j=1}^n(x_j-\overline{x})^2/(n-1)$. $L_{\mu,\sigma^2}(x_1\dots,x_n)$ is strictly positive $\forall x_1,\dots,x_n$, μ, σ^2; therefore, $\log L_{\mu,\sigma^2}(x_1\dots,x_n)$ is well defined and it is given by

$$\log L_{\mu,\sigma^2}(x_1\dots,x_n)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{(n-1)s^2}{2\sigma^2}-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\,.$$

We already pointed out that L_{μ,σ^2} is maximum if and only if $\log L_{\mu,\sigma^2}$ is maximum. The stationary points of the function $\log L_{\mu,\sigma^2}$ (the points where the first partial derivatives of $\log L_{\mu,\sigma^2}$ are both null) are the solutions of the system

$$\begin{cases}\frac{\partial}{\partial\mu}\log L_{\mu,\sigma^2}(x_1\dots,x_n)\equiv\frac{n(\overline{x}-\mu)}{\sigma^2}=0\\ \frac{\partial}{\partial\sigma^2}\log L_{\mu,\sigma^2}(x_1\dots,x_n)=-\frac{n}{2\sigma^2}+\frac{(n-1)s^2+n(\overline{x}-\mu)^2}{2\sigma^4}=0\end{cases}\quad\Leftrightarrow\quad\begin{cases}\mu=\overline{x}\\ \sigma^2=\frac{n-1}{n}\,s^2.\end{cases}$$

Since $L_{\mu,\sigma^2}(x_1,\dots,x_n)\to 0$ as $\mu\to\pm\infty$, then $\overline{x}$ maximizes $L_{\mu,\sigma^2}(x_1,\dots,x_n)$, $\forall\sigma^2$. Moreover, $L_{\overline{x},\sigma^2}(x_1,\dots,x_n)\to 0$ as $\sigma^2\to 0$ and $\sigma^2\to+\infty$.

Thus, we conclude that $\overline{X}$ is the ML estimator of μ, and $\frac{(n-1)S^2}{n}$ is the ML estimator of σ^2; $\overline{X}$ is an unbiased estimator of μ, while $\frac{(n-1)S^2}{n}$ is a biased estimator of σ^2.

Note that, in the Gaussian model, the likelihood method and the method of moments give the same estimators both for the mean and for the variance.

Remark 2.33 (Sample variance, MSE, unbiasedness). To estimate the variance σ^2 of a population we have two candidates: the sample variance S^2 and the estimator of the method of moments $\widehat{\sigma^2}$, which is also the MLE in the case of a Gaussian population. If we set

$$W:=\frac{1}{\sigma^2}\sum_{j=1}^n\big(X_j-\overline{X}\big)^2\,,$$

we can write

$$S^2=\frac{\sigma^2}{n-1}\,W\,,\qquad\widehat{\sigma^2}=\frac{\sigma^2}{n}\,W\,.$$

The sample variance is unbiased, while $\widehat{\sigma^2}$ is biased; this does not necessarily implies that S^2 is always preferable to $\widehat{\sigma^2}$.

Indeed, let us compute the two MSE's in the case of a random sample from a Gaussian population. We know that $W\sim\chi^2(n-1)$ and from the table at pg. 8 we get $\mathbb{E}[W] = n-1$, $\mathrm{Var}[W] = 2(n-1)$. Then, we obtain

$$\begin{aligned}
\mathbb{E}[S^2] = \sigma^2\,, &\qquad \mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{n-1}{n}\,\sigma^2\,,\\
\mathrm{Var}[S^2] = \frac{2\sigma^4}{n-1}\,, &\qquad \mathrm{Var}\left[\widehat{\sigma^2}\right] = \frac{2(n-1)\sigma^4}{n^2}\,,\\
\mathrm{MSE}_{S^2} = \mathrm{Var}[S^2] = \frac{2\sigma^4}{n-1}\,, &\qquad \mathrm{MSE}_{\widehat{\sigma^2}} = \mathrm{Var}\left[\widehat{\sigma^2}\right] + \left(\mathbb{E}\left[\widehat{\sigma^2}\right] - \sigma^2\right)^2 = \frac{(2n-1)\sigma^4}{n^2}\,,\\
\mathrm{MSE}_{S^2} - \mathrm{MSE}_{\widehat{\sigma^2}} &= \left(\frac{1}{n} + \frac{2}{n-1}\right)\frac{\sigma^4}{n} > 0\,.
\end{aligned}$$

We know from Section 2.5.2 that S^2 is the UMVUE of σ^2 and that the Cramer-Rao lower bound for the variance of unbiased estimators is $\mathrm{LB} = \frac{2\sigma^4}{n}$. Note that we have $\mathrm{MSE}_{S^2} > \mathrm{LB} > \mathrm{MSE}_{\widehat{\sigma^2}}$.

Exercise 2.34 (Gaussian model, known mean or variance).

(a) Let $X_1, \dots, X_n$ be a random sample from a Gaussian distribution with unknown mean μ and known variance σ^2. Verify that $\overline{X}$ is the MLE of μ.

(b) Let $X_1, \dots, X_n$ be a random sample from a Gaussian distribution with known mean μ and unknown variance σ^2. Verify that the MLE of σ^2 is $\displaystyle\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2$.

In the applications it is often impossible to solve analytically the maximization problem involved in finding maximum likelihood estimators. Therefore it is necessary to find solutions by using numerical methods.

Example 2.35. Let $X_1, \dots, X_n$ be a random sample extracted from a population with Cauchy density centered at $\theta \in \mathbb{R}$, that is

$$f(x, \theta) = \frac{1}{\pi}\frac{1}{1 + (x-\theta)^2}\,, \qquad \theta \in \mathbb{R}.$$

Then

$$\begin{aligned}
L_\theta(x_1, \dots, x_n) &= \frac{1}{\pi^n}\prod_{j=1}^{n}\frac{1}{1 + (x_j - \theta)^2},\\
\log L_\theta(x_1, \dots, x_n) &= -n\log\pi - \sum_{j=1}^{n}\log[1 + (x_j - \theta)^2],\\
\frac{\partial \log L_\theta}{\partial\theta} &= 2\sum_{j=1}^{n}\frac{x_j - \theta}{1 + (x_j - \theta)^2}\,.
\end{aligned}$$

In this example, as in many other cases, we can determine the MLE only by numerical methods. The most commonly used is the iterative method due to Newton-Raphson: the $t+1$ step is

$$\widehat{\theta}^{(t+1)} = \widehat{\theta}^{(t)} - \left.\frac{\partial \log L_\theta}{\partial\theta}\right|_{\theta=\widehat{\theta}^{(t)}}\left[\left.\frac{\partial^2 \log L_\theta}{\partial\theta\partial\theta^T}\right|_{\theta=\widehat{\theta}^{(t)}}\right]^{-1}, \quad t = 0, 1, \dots$$

where $\widehat{\theta}^{(0)}$ is an assigned initial value and $\frac{\partial^2 \log L_\theta}{\partial\theta\partial\theta^T}$ is the Hessian matrix (in the general case the parameter $\theta^T = (\theta_1, \dots, \theta_m)$ is m-dimensional). It is assumed that the Hessian matrix has maximum rank.

If the likelihood function $L_\theta(x_1, \dots, x_n)$ is continuous (in θ) and Θ is a bounded and closed set, then a ML estimator exists. Otherwise, if this (sufficient) condition is not satisfied, *a ML estimator does not necessarily exist.*

Example 2.36. Let us suppose we toss a coin. We know that the two faces of the coin are different, but we do not know whether the coin is fair or not. Let us toss the coin n times and let $x_1, \dots, x_n$ be the n outcomes. Determine a maximum likelihood estimator of the probability of obtaining the sequence $\{T, H, T\}$ in three successive tosses of the coin.

Let θ be the probability of obtaining head. Then θ is a number in the interval $(0, 1)$ and $x_1, \dots, x_n$ is the outcome of the random sample $X_1, \dots, X_n$ where $X_i \sim \text{Be}(\theta)$, $\theta \in (0, 1)$. We want to estimate the characteristic

$$\begin{aligned}\kappa(\theta) &= P_\theta(X_{n+1} = 0, X_{n+2} = 1, X_{n+3} = 0) = \\ &= P_\theta(X_{n+1} = 0)P_\theta(X_{n+2} = 1)P_\theta(X_{n+3} = 0) = \theta(1-\theta)^2.\end{aligned}$$

The likelihood function of the model is

$$L_\theta(x_1, \dots, x_n) = \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}, \qquad \theta \in (0, 1).$$

For any $\theta \in (0, 1)$, we have that $L_\theta \in (0, 1)$ and therefore $\log L_\theta$ is well defined. Then, we can maximize $\log L_\theta = \sum_{j=1}^n x_j \log(\theta) + (n - \sum_{j=1}^n x_j) \log(1-\theta)$:

$$\frac{\partial}{\partial\theta}\log L_\theta = \frac{\sum_{j=1}^n x_j}{\theta} - \frac{n - \sum_{j=1}^n x_j}{1-\theta} \geq 0 \;\Leftrightarrow\; \theta \leq \overline{x} \text{ and } \frac{\partial}{\partial\theta}\log L_\theta = 0 \;\Leftrightarrow\; \theta = \overline{x}.$$

It follows that: (a) if the sample outcome $x_1, \dots, x_n$ contains at least one head and one tail, then $\overline{X}$ is the maximum likelihood, (b) if the sample outcome $x_1, \dots, x_n$ contains only heads, then $L_\theta(1, \dots, 1) = \theta^n$ does not have maximum in $\Theta = (0, 1)$ (it has sup 1 which is not also a maximum), and therefore a maximum likelihood estimator does not exist, (c) if the sample outcome $x_1, \dots, x_n$ contains only tails, then $L_\theta(0, \dots, 0) = (1-\theta)^n$ does not have maximum in $\Theta = (0, 1)$ (it has sup 0 which is not also a maximum), and again a maximum likelihood estimator does not exist.

But if we take $\Theta = [0, 1]$, the maximum likelihood estimator of $\theta(1-\theta)^2$ exists and, by the invariance property, it is given by $\overline{X}(1-\overline{X})^2$.

The ML estimator is not necessarily unique.

Example 2.37. Let $X_1, \dots, X_n$ be i.i.d. $\sim \mathcal{U}[\theta - 1/2, \theta + 1/2]$, with $\theta \in \mathbb{R}$, that is $f(x, \theta) = \mathbf{1}_{[\theta-1/2,\theta+1/2]}(x)$. The likelihood function is

$$\begin{aligned}L_\theta(x_1, \dots, x_n) &= \prod_{j=1}^n \mathbf{1}_{[\theta-1/2,\theta+1/2]}(x_j) = \mathbf{1}_{[\theta-1/2,\theta+1/2]}(x_{(1)}) \times \mathbf{1}_{[\theta-1/2,\theta+1/2]}(x_{(n)}) \\ &= \mathbf{1}_{[x_{(n)}-1/2,x_{(1)}+1/2]}(\theta)\end{aligned}$$

where $x_{(1)} = \min\{x_1, \dots, x_n\}$ and $x_{(n)} = \max\{x_1, \dots, x_n\}$. The function $\theta \mapsto L_\theta$ is constantly equal to 1 on the interval $[x_{(n)} - 1/2, x_{(1)} + 1/2]$ and is null outside of the interval: hence any point in $[x_{(n)} - 1/2, x_{(1)} + 1/2]$ maximizes L_θ.

The estimators obtained by the method of moments and those obtained by the method of maximum likelihood are in general different.

Example 2.38. Let $X_1, \dots, X_n$ be i.i.d. $\sim \mathcal{U}[0, \theta]$, with $\theta > 0$. The likelihood function of the model is

$$L_\theta(x_1 \dots, x_n) = \frac{1}{\theta^n} \mathbf{1}_{[x_{(n)}, +\infty)}(\theta)\,, \quad x_1, x_2, \dots, x_n > 0,$$

(see Eq. (2.11)). $\theta \mapsto L_\theta$ is strictly decreasing on the interval $[x_{(n)}, \infty)$, therefore $x_{(n)}$ is the unique maximum for $L_\theta(x_1, \dots, x_n)$ and $X_{(n)}$ is the maximum likelihood estimator of θ.

Remember that in 2.6.1 we proved that the method of moments estimator of θ is $2\overline{X}$.

In order to decide which one between $X_{(n)}$ and $2\overline{X}$ is better, we can calculate the mean square error (MSE). In Example 2.26 we obtained the mean and the variance of $X_{(n)}$, reported hereafter:

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n\theta}{n+1}, \qquad \operatorname{Var}_\theta[X_{(n)}] = \frac{n\theta^2}{(n+1)^2(n+2)}\,.$$

Therefore,

$$\begin{aligned} \operatorname{MSE}(X_{(n)}) &= \operatorname{Var}_\theta[X_{(n)}] + \big(\mathbb{E}_\theta[X_{(n)}] - \theta\big)^2 = \frac{n\theta^2}{(n+1)^2(n+2)} + \left(\frac{n}{n+1} - 1\right)^2 \theta^2 \\ &= \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned}$$

With regard to the estimator $2\overline{X}$, it is unbiased and its variance is

$$\operatorname{Var}_\theta[2\overline{X}] = \frac{4\operatorname{Var}_\theta[X_1]}{n} = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}$$

which implies

$$\operatorname{MSE}(2\overline{X}) = \operatorname{Var}_\theta[2\overline{X}] = \frac{\theta^2}{3n}.$$

So, if $n = 1$, then $2X_1$ and $X_{(1)} = X_1$ have the same MSE and we choose $2X_1$ since it is unbiased. On the other hand, for $n \geq 2$, the estimator $X_{(n)}$ is preferable because its MSE is uniformly (in θ) smaller than the MSE of $2\overline{X}$.

As seen in the example above, the two methods of finding estimators illustrated so far produce in general different estimators. We expect that the estimator of maximum likelihood is better than the estimator of moments, in that the method of moments does not exploit all information contained in the sample and in the distribution function which generates the sample. In fact, the method of moments makes use only of the sample moments, which synthesizes the sample, and of the moments of the distribution $F(x, \theta)$. On the contrary, the maximum likelihood estimator exploits all sample information and all theoretical information (through the likelihood function).

2.6.3 Properties of maximum likelihood estimators

We now show some properties of maximum likelihood estimators, which clarify in which sense ML estimators are good estimators.

Proposition 2.5. *If the regularity conditions that imply Cramer-Rao inequality hold true and there exists an efficient estimator T of $\kappa(\theta)$ (an estimator whose variance is the Cramer-Rao lower bound), then T is the essentially unique ML estimator of $\kappa(\theta)$.*

Proof. By the regularity conditions, it follows that the MLE $\widehat{\theta}$ satisfies the equation

$$\frac{\partial}{\partial\theta}\log L_\theta(x_1,\ldots,x_n)=0. \tag{2.24}$$

On the other hand, since T is efficient, then

$$P_\theta\Big[\frac{\partial}{\partial\theta}\log L_\theta(x_1,\ldots,x_n)=a(\theta,n)\big(T-\kappa(\theta)\big)\Big]=1, \qquad \forall\theta\in\Theta. \tag{2.25}$$

By the above equation and by (2.18), it follows that

$$\begin{aligned}0<nI(\theta)=\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log L_\theta(x_1,\ldots,x_n)\right]&=\mathrm{Var}_\theta\left[a(\theta,n)\big(T-\kappa(\theta)\big)\right]\\&=\big(a(\theta,n)\big)^2\,\mathrm{Var}_\theta[T],\end{aligned}$$

which implies that the function $a(\theta,n)$ is non zero for all $\theta\in\Theta$. Therefore, by substituting $a(\theta,n)\big(T-\kappa(\theta)\big)$ to $\frac{\partial}{\partial\theta}\log L_\theta(x_1\ldots,x_n)$ in (2.24), we obtain that necessarily $P_\theta[\kappa(\widehat{\theta})=T]=1, \quad \forall\theta\in\Theta$. But T, being efficient, is also UMVUE and therefore it is essentially unique. □

Finally, the asymptotic properties of the MLE are summarized in the following Proposition.

Proposition 2.6. *Let $X_1,\ldots,X_n,\ldots$ be a sequence of i.i.d. random variables with common density function $f(x,\theta)$, $\theta\in\Theta$, and let $\{T_n\}_n$ be the sequence of ML estimators of $\kappa(\theta)$. If $f(x,\theta)$ satisfies the regularity conditions which guarantees the Cramer-Rao inequality and some other regularity conditions (existence and continuity of the second and third partial derivative of $f(x,\theta)$ with respect to θ), then the sequence $\{T_n\}_n$ is*

1. *asymptotically unbiased for $\kappa(\theta)$,*
2. *consistent for $\kappa(\theta)$,*
3. *asymptotically Gaussian with asymptotic mean $\kappa(\theta)$ and asymptotic variance $\dfrac{[\kappa'(\theta)]^2}{nI(\theta)}$, that is*

$$\lim_{n\to\infty}P\left(\frac{T_n-\kappa(\theta)}{\sqrt{\frac{(\kappa'(\theta))^2}{nI(\theta)}}}\le z\right)=\Phi(z), \qquad \forall z\in\mathbb{R}.$$

For more details both on the regularity hypotheses and on the proof of this result, see Section §25 on pages 228-233 in Borovkov [1].

Proposition 2.6 says that *an ML estimator is asymptotically efficient and, therefore, asymptotically UMVUE.*

The hypotheses necessary for the validity of Proposition 2.6 are satisfied by nearly all models considered in these lecture notes; the exceptions are those models where the support of the density $f(x,\theta)$, i.e. the set $\{x : f(x,\theta) > 0\}$, depends on θ.

We conclude this section with two exercises.

Example 2.39. Let $X_1, \dots, X_n$ be a random sample from a population with density

$$f(x;\theta) = \frac{1}{\theta}\, x^{-\left(1+\frac{1}{\theta}\right)} \mathbf{1}_{[1,+\infty]}(x), \qquad \theta > 0$$

1. Find the MLE $\widehat{\theta}_n$ of θ.
2. Find the MLE $\widehat{\tau}_n$ of $\tau(\theta) = \frac{1}{\theta}$.
3. Show that $\widehat{\tau}_n$ is asymptotically unbiased and consistent for $\tau(\theta)$.
4. Determine the asymptotic c.d.f. of $\widehat{\tau}_n$.

1. We have that

$$\ell_n(\theta) := \log L_\theta(x_1,\dots,x_n) = \log\left[\prod_{i=1}^n f(x_i;\theta)\right] = -n\log\theta - \left(1+\frac{1}{\theta}\right)\sum_{i=1}^n \log x_i,$$

and $\dfrac{\partial \ell_n}{\partial\theta} \equiv -\dfrac{n}{\theta} + \dfrac{1}{\theta^2}\displaystyle\sum_{i=1}^n \log x_i = 0$ has solution $\hat{\theta} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \log x_i$. Moreover,

$$\frac{\partial^2 \ell_n}{\partial\theta^2} \equiv \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^n \log x_i < 0 \quad\Longleftrightarrow\quad n - \frac{2}{\theta}\sum_{i=1}^n \log x_i < 0$$

is verified by $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n \log x_i$. It follows that

$\widehat{\theta}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \log X_i$ is the MLE of θ.

2. By the invariance property, $\widehat{\tau}_n = 1/\widehat{\theta}_n$ is the MLE of $\tau(\theta)$.

3. If X has density $f(x;\theta)$, then $Y = \log X$ has exponential density with parameter (mean) θ. Indeed, for $y > 0$ we have

$$\begin{aligned} P[Y \le y] = P[\log X \le y] = P\left[X \le \mathrm{e}^y\right] \\ = \frac{1}{\theta}\int_1^{\mathrm{e}^y} \frac{1}{x^{1+1/\theta}}\,\mathrm{d}x = \left[-\frac{1}{x^{1/\theta}}\right]_1^{\mathrm{e}^y} = 1 - \mathrm{e}^{-y/\theta}. \end{aligned}$$

Hence, $\sum_{i=1}^n \log X_i$ has density $\Gamma(n,\theta)$ and $\widehat{\theta}_n \sim \Gamma(n, \frac{\theta}{n})$. This implies

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_n] &= \int_0^\infty \frac{1}{x} \cdot \frac{(\frac{n}{\theta})^n}{\Gamma(n)} x^{n-1} \mathrm{e}^{-x\frac{n}{\theta}} \mathrm{d}x = \int_0^\infty \frac{n}{\theta} \cdot \frac{(\frac{n}{\theta})^{n-1}}{(n-1)\Gamma(n-1)} x^{n-1-1} \mathrm{e}^{-x\frac{n}{\theta}} \mathrm{d}x \\
&= \frac{n}{n-1} \cdot \frac{1}{\theta} \to \frac{1}{\theta} \qquad \text{as } n \to +\infty,
\end{aligned}$$

hence, $\{\widehat{\tau}_n\}_n$ is asymptotically unbiased for τ. Moreover

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_n^2] &= \int_0^\infty \left(\frac{n}{\theta}\right)^2 \frac{(\frac{n}{\theta})^{n-2}}{(n-1)(n-2)\Gamma(n-2)} x^{n-2-1} \mathrm{e}^{-x\frac{n}{\theta}} \mathrm{d}x = \frac{n^2}{(n-1)(n-2)} \cdot \frac{1}{\theta^2}\,, \\
\mathrm{Var}[\hat{\tau}_n] &= \frac{n^2}{(n-1)(n-2)} \cdot \frac{1}{\theta^2} - \left(\frac{n}{n-1} \cdot \frac{1}{\theta}\right)^2 = \frac{n^2}{(n-1)^2(n-2)\theta^2} \to 0,
\end{aligned}$$

hence, $\{\widehat{\tau}_n\}_n$ is a MSE-consistent sequence of estimators of τ.

4. Since $\log X_1 \sim \Gamma(1,\theta)$, we have

$$\begin{aligned}
nI(\theta) = \mathrm{Var}_\theta \left[\frac{\partial \log L_\theta(X_1,\dots,X_n)}{\partial\theta}\right] &= \mathrm{Var}_\theta \left[-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n \log X_i\right] \\
&= \mathrm{Var}_\theta \left[\frac{1}{\theta^2}\sum_{i=1}^n \log X_i\right] = \frac{n}{\theta^4}\mathrm{Var}_\theta[\log(X_1)] = \frac{n\theta^2}{\theta^4} = \frac{n}{\theta^2}\,.
\end{aligned}$$

Moreover, $[\tau'(\theta)]^2 = (-1/\theta^2)^2 = \theta^{-4}$ and hence $\frac{(\tau'(\theta))^2}{nI(\theta)} = \frac{1}{n\theta^2}$. Since $\widehat{\tau}_n$ is the ML estimator of a regular statistical model, it follows

$$\lim_{n\to\infty} P\left[\sqrt{n}\theta(\widehat{\tau}_n - 1/\theta) \le z\right] = \Phi(z) \qquad \forall z \in \mathbb{R}.$$

Example 2.40 (Uniform model). Let $X_1,\dots,X_n$ be i.i.d. $\sim \mathcal{U}[0,\theta]$, with $\theta > 0$. Example 2.38 showed that the MLE of θ is the maximum observation $X_{(n)}$. We now investigate the asymptotic properties of $X_{(n)}$.

- $X_{(n)}$ is asymptotically unbiased. In fact:

$$\lim_{n\to\infty} \mathbb{E}_\theta[X_{(n)}] = \lim_{n\to\infty} \frac{n\theta}{n+1} = \theta\,.$$

- $X_{(n)}$ is MSE-consistent. In fact:

$$\lim_{n\to\infty} \mathrm{Var}_\theta[X_{(n)}] = \lim_{n\to\infty} \frac{n\theta^2}{(n+1)^2(n+2)} = 0\,.$$

- $X_{(n)}$ is not asymptotically Gaussian. In fact, if we consider the standardization of $X_{(n)}$, i.e. the random variable

$$Z_n := \frac{X_{(n)} - E_\theta(X_{(n)})}{\sqrt{\mathrm{Var}_\theta[X_{(n)}]}} = \frac{X_{(n)} - \frac{n\theta}{n+1}}{\sqrt{\frac{n\theta^2}{(n+1)^2(n+2)}}} = \sqrt{\frac{n+2}{n}}\left(\frac{n+1}{\theta}X_{(n)} - n\right)$$

it can be proved that the limit c.d.f. of Z_n is

$$F(z) = \begin{cases} e^z & z \leq 0 \\ 1 & z > 0 \end{cases}$$

i.e. $-Z_n$ is asymptotically exponential with parameter 1.

The proof follows (for fans of the technicalities):

$$\begin{aligned} P_\theta[Z_n \leq z] &= P_\theta\left[X_{(n)} \leq \frac{\theta}{n+1}\Big(z\sqrt{\frac{n}{n+2}} + n\Big)\right] \\ &= \begin{cases} 0 & \text{if } \frac{\theta}{n+1}\left(z\sqrt{\frac{n}{n+2}} + n\right) \leq 0 \\ \left(\frac{z\sqrt{\frac{n}{n+2}}+n}{n+1}\right)^n & \text{if } 0 < \frac{\theta}{n+1}\left(z\sqrt{\frac{n}{n+2}} + n\right) < \theta \quad \text{[Example 2.26]} \\ 1 & \text{if } \frac{\theta}{n+1}\left(z\sqrt{\frac{n}{n+2}} + n\right) \geq \theta \end{cases} \\ &\simeq \begin{cases} 0 & \text{if } \frac{\theta z}{n} + \theta \leq 0 \\ \left(1 + \frac{z}{n}\right)^n & \text{if } 0 < \frac{\theta z}{n} + \theta < \theta \\ 1 & \text{if } \frac{\theta z}{n} + \theta \geq \theta \end{cases} \to \begin{cases} e^z & \text{if } z \leq 0 \\ 1 & \text{if } z > 0 \end{cases} \qquad \text{for } n \to \infty. \end{aligned}$$

Chapter 3

Interval estimation

3.1 Interval estimation of the mean and the variance of a Gaussian population

This section presents interval estimation of the mean and the variance of a Gaussian population. Section 3.2 generalizes the concept of confidence interval to the case of a general statistical model.

Let us recall that the quantiles of the standard normal distribution are denoted by the letter "z": $z_\gamma \equiv \Phi^{-1}(\gamma)$ is the quantile of order γ of the distribution $\mathcal{N}(0;1)$. Similarly, we write $t_\gamma(k)$ for the quantile of order γ of the Student t distribution with k degrees of freedom and $\chi^2_\gamma(k)$ for the quantile of order γ of the χ^2 distribution with k degrees of freedom. The notations for the quantiles can change according to the authors: z_γ is also written as ϕ_γ, $t_\gamma(k)$ is also written as $t_k(\gamma)$ or as $t(k;\gamma)$, $\chi^2_\gamma(k)$ is also written as $\chi^2_k(\gamma)$ or as $\chi^2(k;\gamma)$.

3.1.1 Interval estimation of the mean

Let $X_1, \ldots, X_n$ be a random sample from a Gaussian distribution with parameters μ, σ^2, that is $X_1, \ldots, X_n$ are i.i.d. random variables $\sim \mathcal{N}(\mu, \sigma^2)$. Suppose that the standard deviation is known, $\sigma = 0.8$. In Chapter 2, the use of $\overline{X}$ as an estimator of μ is circumstantiated. The statistics $\overline{X}$ is a *point estimator* of μ. Point estimation is somehow not completely satisfying for the following reason. Since $\overline{X}$ is an absolutely continuous variable, it follows that $P(\overline{X} = c) = 0$, $\forall c \in \mathbb{R}$. Therefore, the probability that $\overline{X}$ assumes the true unknown value of μ is null, whatever the value of μ is. As we now show, we can measure in probabilistic terms the error that we make when we use for μ the value of $\overline{X}$ corresponding the sample realization.

Let $\epsilon > 0$ be such that

$$P\left(-\epsilon < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \epsilon\right) = 0.95. \tag{3.1}$$

If n and σ are known, then ϵ is uniquely determined by the equation $2\Phi(\epsilon) - 1 = 0.95$, that is

$$2\Phi(\epsilon) - 1 = 0.95 \ \ \textbf{iff} \ \ \Phi(\epsilon) = \frac{1+0.95}{2} = 0.975 \ \ \textbf{iff} \ \ \epsilon = \Phi^{-1}(0.975) = z_{.975} \simeq 1.96.$$

A priori, i.e. before we actually observe the sample realization, we are quite confident that, **independently on the sample realization**, the error we make when estimating μ with $\overline{X}$ is at most $1.96\,\sigma/\sqrt{n}$. If, for example, the sample is of size 4, the error is $1.96 \times 0.8/\sqrt{4} = 0.784$. In this statement, the degree of reliance or confidence measured in term of probability is equal to 95%.

Equality (3.1) can be rewritten as

$$P\left(T_1 < \mu < T_2\right) = 0.95\,,$$

where

$$T_1 = \overline{X} - z_{(1+0.95)/2}\,\frac{\sigma}{\sqrt{n}}\,, \qquad T_2 = \overline{X} + z_{(1+0.95)/2}\,\frac{\sigma}{\sqrt{n}}\,.$$

In other words, *a priori*, with probability 95% the unknown (but deterministic) true value of μ is contained in the **random interval** (T_1, T_2). **The extreme values** T_1, T_2 **are statistics** which depend *a*) on the sample $X_1, \ldots, X_n$; *b*) on known information on the c.d.f. of the sample (in this case, σ); *c*) on the degree of confidence (*confidence level* or *coefficient* — in this case, 95%).

If we actually make the experiment and observe, for example,

$$x_1 = 4.87, \ x_2 = 5.06, \ x_3 = 2.8, \ x_4 = 5.32,$$

then we get that $\overline{X}$, T_1 and T_2 take the values $\overline{x} = 4.5125$, $t_1 = 4.5125 - 0.784 = 3.7285$ and $t_2 = 4.5125 + 0.784 = 5.2965$. Both the random interval (T_1, T_2) and the numeric interval $(t_1, t_2) = (3.7285, 5.2965)$ are called a *0.95 confidence interval for the parameter* μ.

The meaning of the confidence level is the following: if we extract a sample of size 4 a great number of times and each time we calculate the corresponding confidence interval $(\overline{x}-0.784, \overline{x}+0.784)$, we expect that more or less 95% of these intervals contain the true value of the mean μ. It is worthwhile to underline that, when we say that the we are 0.95 confident that the interval contains the true value of the parameter, we are talking about the procedure illustrated above, while, once we have actually observed the sample and constructed the interval, this interval either contains the true value of the parameter, or not. In other words the confidence level is assured by the procedure. Once the experiment is made, there is no probability involved anymore.

The general definition of *confidence interval* is given in Definition 3.2; in the following remarks, we give the explicit expressions for the confidence intervals for the mean and the variance of a Gaussian population.

Remark 3.1. Let $x_1, \ldots, x_n$ be a sample realization of the random sample $X_1, \ldots, X_n$ from the distribution $\mathcal{N}(\mu; \sigma^2)$, with known variance σ^2. Fixed $\gamma \in (0,1)$, a $100\gamma\%$

confidence interval for the parameter μ is

$$\left(\overline{X} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right) \qquad \text{“before the experiment”},$$

$$\left(\overline{x} - z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right) \qquad \text{“after the experiment”}.$$

The number γ, or the percentage $100\gamma\%$, is called the *confidence level of the interval.*

Sometimes the following notations are used:

$$\mu = \overline{X} \pm z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\,, \qquad \mu = \overline{x} \pm z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\,.$$

Remark 3.2. The confidence level of a confidence interval is usually denoted by γ or by $1-\alpha$. So, pay attention that we have $\gamma = 1-\alpha$ and $(1+\gamma)/2 = 1-\alpha/2$. Typical choices are $\gamma = 0.90,\ 0.95,\ 0.99$ and, respectively, $\alpha = 0.10,\ 0.05,\ 0.01$.

Suppose now that also the variance σ^2 is unknown. In order to evaluate *a priori* the error we make when we approximate μ with $\overline{X}$, we consider $P(-\epsilon < \sqrt{n}(\overline{X}-\mu)/S < \epsilon)$, where $S = \sqrt{S^2}$ is the sample standard deviation, and we determine ϵ in such a way that

$$P\left(-\epsilon < \frac{\overline{X}-\mu}{S/\sqrt{n}} < \epsilon\right) = 0.95. \tag{3.2}$$

We have taken 0.95 just to make a numeric example; in any case it is a typical choice. By observing that $\sqrt{n}(\overline{X}-\mu)/S \sim t(n-1)$ and that the Student t-density is symmetric (with respect to 0), analogously to the case of known variance, we obtain that ϵ is the quantile of order $(1+0.95)/2$ of the Student t distribution with $n-1$ degrees of freedom, i.e. $\epsilon = t_{.975}(n-1)$.

If again we suppose that the random size is 4, we get $\epsilon = t_{.975}(3) \simeq 3.182$ and Eq. (3.2) can be rewritten as

$$P\left(\overline{X} - 3.182\frac{S}{\sqrt{n}} < \mu < \overline{X} + 3.182\frac{S}{\sqrt{n}}\right) = 0.95.$$

The last equation says that, before we observe the sample realization, the probability that the random interval $\overline{X} \pm 3.182S/\sqrt{n}$ contains μ is 0.95. Let us stress that “$\mu \in (\overline{X} - 3.182S/\sqrt{n}, \overline{X} + 3.182S/\sqrt{n})$” is a random event in that the extreme values of the interval are random. Instead, μ is not random.

If we observe the sample

$$x_1 = 4.87,\ x_2 = 5.06,\ x_3 = 2.8,\ x_4 = 5.32,$$

then $\sum_{j=1}^n x_j^2/(n-1) \simeq 28.4876$, $s^2 = 28.4876 - 4/3(4.5125)^2 \simeq 1.337$ and $s = \sqrt{1.337} \simeq 1.1565$. Hence,

$$\overline{x} - 3.182s/\sqrt{n} \simeq 3.4484 \quad \text{and} \quad \overline{x} + 3.182s/\sqrt{n} \simeq 5.5765$$

i.e. $(3.4484, 5.5765)$ is a 95% *confidence interval for* μ.

Note that we calculated S^2 by using formula (2.2).

Remark 3.3. Let $x_1, \dots, x_n$ be a sample realization of the random sample $X_1, \dots, X_n$ from the distribution $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are both unknown. Then, a confidence interval for μ of level γ is, before the experiment,

$$\left(\overline{X} - t_{(1+\gamma)/2}(n-1)\,\frac{S}{\sqrt{n}},\ \overline{X} + t_{(1+\gamma)/2}(n-1)\,\frac{S}{\sqrt{n}}\right)$$

or, after the experiment,

$$\left(\overline{x} - t_{(1+\gamma)/2}(n-1)\,\frac{s}{\sqrt{n}},\ \overline{x} + t_{(1+\gamma)/2}(n-1)\,\frac{s}{\sqrt{n}}\right).$$

As before, the following notations can be used:

$$\mu = \overline{X} \pm t_{(1+\gamma)/2}(n-1)\,\frac{S}{\sqrt{n}}, \qquad \mu = \overline{x} \pm t_{(1+\gamma)/2}(n-1)\,\frac{s}{\sqrt{n}}$$

Remark 3.4. A confidence interval estimates μ through an interval: the length of the interval gives a measure of the precision of this estimate.

a) If σ^2 *is known*, the length of the confidence interval is

$$L = 2\, z_{\frac{1+\gamma}{2}}\, \frac{\sigma}{\sqrt{n}}.$$

Note that L is not random in that it does not depend on the sample realization, but only on σ, n and γ. In particular:

- Fixed σ and n, L is an increasing function of γ. Therefore, when we construct a confidence interval we have to come to a compromise between precision and confidence level: the more confident we want to be that μ is in the interval, the less precise the interval is (i.e. the longer the interval is).
- Fixed σ and γ, L is a decreasing function of n. As n increases, the variance of $\overline{X}$ decreases, hence the interval is smaller and the estimate more precise.
- Fixed γ and n, L is an increasing function of σ. Indeed, the variance of $\overline{X}$ is an increasing function of σ: the bigger σ is, the more probable $\overline{X}$ takes values spread around μ, and therefore the bigger the interval is.

b) If σ^2 *is unknown,* the length is:

$$L = 2\, t_{\frac{1+\gamma}{2}}(n-1)\, \frac{S}{\sqrt{n}}.$$

Hence, if σ^2 is unknown, L is random, in that it depends on the sample realization through S, and its mean is

$$\mathbb{E}[L] = \frac{2t_{\frac{1+\gamma}{2}}(n-1)}{\sqrt{n}}\,\mathbb{E}[S] = \frac{2^{3/2}t_{\frac{1+\gamma}{2}}(n-1)}{\sqrt{n(n-1)}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\,\sigma.$$

Let us prove this result. If $Y = (n-1)S^2/\sigma^2$, then $Y \sim \chi^2(n-1)$ and

$$\begin{aligned}\mathbb{E}[S] &= \mathbb{E}\left[\sqrt{\frac{\sigma^2 Y}{n-1}}\right] = \frac{\sigma}{\sqrt{n-1}}\,\mathbb{E}\left[Y^{1/2}\right]\\ &= \frac{\sigma}{\sqrt{n-1}}\int_0^\infty y^{1/2} f_{\chi^2(n-1)}(y)\,\mathrm{d}y = \frac{\sigma}{\sqrt{n-1}}\int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}\,\mathrm{e}^{-\frac{y}{2}} y^{\frac{n}{2}-1}\,\mathrm{d}y\\ &= \sigma\sqrt{\frac{2}{n-1}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\int_0^\infty f_{\chi^2(n)}(y)\,\mathrm{d}y = \sigma\sqrt{\frac{2}{n-1}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\,.\end{aligned}$$

The Gamma function $\Gamma(\alpha)$ is defined in Section 1.2.4.

Remark 3.5. To determine $\mathbb{E}[L]$ when the variance is unknown, we calculated $\mathbb{E}[S_n]$, obtaining:

$$\mathbb{E}[S_n] = \sigma\sqrt{\frac{2}{n-1}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

We use S_n instead of S when we want to stress the dependence on the size of the sample. By the fact that $\mathbb{E}[S_n^2] = \sigma^2$, we have that

$$\sigma^2 - \mathbb{E}[S_n]^2 = \sigma^2 - \mathbb{E}[S_n^2] + \mathbb{E}[S_n^2] - \mathbb{E}[S_n]^2 = \mathrm{Var}[S_n] > 0\,.$$

Therefore,

$$\mathbb{E}[S_n] < \sigma\,, \qquad \forall n \geq 2\,.$$

For example, if $n = 2$ then $\mathbb{E}[S_2] = \sigma\sqrt{2/\pi} < \sigma$. Hence, the fact that S_n^2 is an unbiased estimator of the variance σ^2 does not imply that S_n is an unbiased estimator of the standard deviation σ. Nevertheless, some unbiasedness property is still true, precisely, S_n is *asymptotically unbiased*. Indeed, one can prove that, for large n, $\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \asymp 1 - \frac{1}{4n}$; then, one has

$$\lim_{n\to\infty} \mathbb{E}[S_n] = \sigma.$$

Remark 3.6. Let us compare the lengths of the two confidence intervals for the mean, given in Remark 3.4. We have

$$\begin{aligned} L_1 &= 2\,z_{\frac{1+\gamma}{2}}\,\frac{\sigma}{\sqrt{n}}\,, & z_{\frac{1+\gamma}{2}} &< t_{\frac{1+\gamma}{2}}(n-1),\\ L_2 &= 2\,t_{\frac{1+\gamma}{2}}(n-1)\,\frac{\mathbb{E}[S_n]}{\sqrt{n}}\,, & \sigma &> \mathbb{E}[S_n].\end{aligned}$$

So we have a trade off between the contributions of the quantile and of the dispersion.

Remark 3.7. The confidence intervals shown so far are symmetric with respect to the sample mean. This is due to the fact that the random variables used in the construction $\big(\sqrt{n}(\overline{X} - \mu)/\sigma$ if σ^2 is known, and $\sqrt{n}(\overline{X} - \mu)/S$ if σ^2 is unknown$\big)$ have probability densities symmetric with respect to 0, and the point where the probability density has a maximum (the *mode* of the distribution) is 0. It follows that the interval $\overline{X} \pm z_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}$ gives the best interval estimation of μ (in the case of Gaussian population and known variance) among all interval estimations of the type $(\overline{x} + a, \overline{x} + b)$ with level $100\gamma\%$.

3.1.2 Interval estimation of the variance

If μ is unknown, in order to construct a confidence interval for σ^2, we start with the random variable $(n-1)S^2/\sigma^2$. Note that it depends on the unknown parameter σ^2 which we want to estimate, but its c.d.f. does not depend on any unknown parameter, in that $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. Therefore, fixed $\gamma \in (0,1)$, we determine a, b such that

$$P\left(a < \frac{(n-1)S^2}{\sigma^2} < b\right) = \gamma. \tag{3.3}$$

We examine the following three cases, which give rise to three different $100\gamma\%$ confidence intervals for σ^2.

1. $[a = 0]$ If $a = 0$, then Eq. (3.3) reduces to

$$P\left(\frac{(n-1)S^2}{\sigma^2} < b\right) = \gamma, \tag{3.4}$$

in the unknown b; it follows necessarily that $b = \chi^2_\gamma(n-1)$, the quantile of order γ of the distribution $\chi^2(n-1)$; Eq. (3.4) is equivalent to

$$P\left(\sigma^2 > \frac{(n-1)S^2}{\chi^2_\gamma(n-1)}\right) = \gamma\,.$$

If, for example, we observe the sample realization $x_1 = 4.87$, $x_2 = 5.06$, $x_3 = 2.8$, $x_4 = 5.32$, and if $\gamma = 0.95$, then $(4-1)s^2 \simeq 3 \times 1.337$, $\chi^2_{.95}(3) \simeq 7.814728$ and we get the 95% one-sided confidence interval $(0.513, +\infty)$. In general:

Remark 3.8. Let $\gamma \in (0,1)$ and let s^2 be the value taken by S^2 at the sample realization $x_1, \ldots, x_n$ of a random sample from a $\mathcal{N}(\mu;\sigma^2)$ distribution, with unknown mean μ. Then, the statistics $\frac{(n-1)S^2}{\chi^2_\gamma(n-1)}$ is a *lower confidence limit* for the variance σ^2 of level $100\gamma\%$. We also say that $\left(\frac{(n-1)S^2}{\chi^2_\gamma(n-1)}\,, +\infty\right)$ is an upper one-sided confidence interval. The same terminology is used when the statistics is replaced by its value $\frac{(n-1)s^2}{\chi^2_\gamma(n-1)}$.

2. $[b = +\infty]$ In this case Eq. (3.3) becomes

$$P\left(a < \frac{(n-1)S^2}{\sigma^2}\right) = \gamma \tag{3.5}$$

in the unknown a; this implies $a = \chi^2_{1-\gamma}(n-1)$. Equation (3.5) is equivalent to

$$P\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\gamma}(n-1)}\right) = \gamma\,.$$

If, for example, we observe the sample realization $x_1 = 4.87$, $x_2 = 5.06$, $x_3 = 2.8$, $x_4 = 5.32$, and $\gamma = 0.95$, then $(4-1)s^2 \simeq 3 \times 1.337$, $\chi^2_{.05}(3) \simeq 0.3518$ and we get the 95% lower one-sided confidence interval $(0, 11.4)$. In general:

Remark 3.9. Let $\gamma \in (0,1)$ and let s^2 be the value taken by S^2 at the sample realization $x_1, \dots, x_n$ of a random sample from a $\mathcal{N}(\mu; \sigma^2)$ distribution, with unknown mean μ. Then, the statistics $\frac{(n-1)S^2}{\chi^2_{1-\gamma}(n-1)}$ is an *upper confidence limit* for the variance σ^2 of level $100\gamma\%$. We also say that $\left(0, \frac{(n-1)S^2}{\chi^2_{1-\gamma}(n-1)}\right)$ is a lower one-sided confidence interval. The same terminology is used when the statistics is replaced by its value $\frac{(n-1)s^2}{\chi^2_{1-\gamma}(n-1)}$.

3. $[0 < a < b < +\infty]$ In this case we solve Eq. (3.3) in the unknowns a, b in such a way that the mass $1-\gamma$ is uniformly distributed on the left and on the right of the interval (a, b). It follows that

$$a = \chi^2_{\frac{1-\gamma}{2}}(n-1) \qquad \text{and} \qquad b = \chi^2_{\frac{1-\gamma}{2}+\gamma}(n-1) = \chi^2_{\frac{1+\gamma}{2}}(n-1)\,.$$

We can then rewrite Eq. (3.3) as follows:

$$P\left(\frac{(n-1)S^2}{\chi^2_{\frac{1+\gamma}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{1-\gamma}{2}}(n-1)}\right) = \gamma. \tag{3.6}$$

Remark 3.10. Let $\gamma \in (0,1)$ and let S^2 be the sample variance of a random sample from a $\mathcal{N}(\mu; \sigma^2)$ distribution, with unknown mean and variance. Then,

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{1+\gamma}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{1-\gamma}{2}}(n-1)}\right)$$

is a $100\gamma\%$ bilateral confidence interval for σ^2. The same terminology is used when the statistics S^2 is substituted by its value s^2.

If, for example, we observe the sample realization $x_1 = 4.87$, $x_2 = 5.06$, $x_3 = 2.8$, $x_4 = 5.32$, from the c.d.f. $\mathcal{N}(\mu; \sigma^2)$ and $\gamma = 0.95$, then $\chi^2_{.975}(3) \simeq 9.35$, $\chi^2_{.025}(3) \simeq 0.216$ and we get that $\left(\frac{3\times 1.337}{9.35}, \frac{3\times 1.337}{0.216}\right) \simeq (0.4284, 18.57)$ is a 95% bilateral confidence interval for σ^2.

The data above have been generated, by using a computer, from a c.d.f. $\mathcal{N}(4.4; 0.8^2)$. Since we know the generating distribution (what does not happen in real applications) we observe that the upper bound furnished from the confidence interval is far from the true value of the variance. This happens because, in choosing the extremes of the interval, we just followed a criterion based on simplicity (symmetry of the mass $1-\gamma$), and did not take care of optimality criteria.

Remark 3.11. Finally, let us treat the case of interval estimation of the variance when the mean μ is known. In this case, we take into account the known information μ by using the statistics

$$S_0^2 := \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2\,.$$

The statistics S_0^2 measures the dispersion of the sample around the true mean μ. Recall that in point (b) of Exercise 2.34 we saw that S_0^2 is the maximum likelihood estimator of σ^2 when the mean is known. Moreover, note that the random variable nS_0^2/σ^2 has $\chi^2(n)$ distribution and therefore, analogously to the case of unknown mean μ, we have the following $100\gamma\%$ confidence intervals for σ^2 when μ is known:

$$\left(\frac{\sum_{j=1}^n (X_j-\mu)^2}{\chi^2_\gamma(n)}, +\infty\right) \qquad 100\gamma\% \text{ upper one-sided confidence interval;}$$

$$\left(0, \frac{\sum_{j=1}^n (X_j-\mu)^2}{\chi^2_{1-\gamma}(n)}\right) \qquad 100\gamma\% \text{ lower one-sided confidence interval;}$$

$$\left(\frac{\sum_{j=1}^n (X_j-\mu)^2}{\chi^2_{\frac{1+\gamma}{2}}(n)}, \frac{\sum_{j=1}^n (X_j-\mu)^2}{\chi^2_{\frac{1-\gamma}{2}}(n)}\right) \qquad 100\gamma\% \text{ bilateral confidence interval.}$$

3.1.3 Simultaneous confidence region for the mean and the variance

We now use the results obtained in the preceding sections to construct a "confidence region" for a simultaneous estimation of the mean and the variance. It can be proven that $\sqrt{n}(\overline{X}-\mu)/\sigma$ and $(n-1)S^2/\sigma^2$ are independent random variables. This implies that the events A and B here below are independent:

$$A = \left\{\sqrt{n}\left|\frac{\overline{X}-\mu}{\sigma}\right| < z_{\frac{1+\sqrt{\gamma}}{2}}\right\} = \left\{n\left(\frac{\overline{X}-\mu}{z_{\frac{1+\sqrt{\gamma}}{2}}}\right)^2 < \sigma^2\right\}$$

$$B = \left\{\frac{(n-1)S^2}{\chi^2_{\frac{1+\sqrt{\gamma}}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{1-\sqrt{\gamma}}{2}}(n-1)}\right\}.$$

Moreover, reasoning as for confidence intervals, we get $P(A) = P(B) = \sqrt{\gamma}$ and, hence,

$$P\left[n\left(\frac{\overline{X}-\mu}{z_{\frac{1+\sqrt{\gamma}}{2}}}\right)^2 < \sigma^2,\ \frac{(n-1)S^2}{\chi^2_{\frac{1+\sqrt{\gamma}}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{1-\sqrt{\gamma}}{2}}(n-1)}\right] = P(A\cap B) = P(A)P(B) = \gamma.$$

It follows that

$$\left\{(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty) :\ n\left(\frac{\overline{x}-\mu}{z_{\frac{1+\sqrt{\gamma}}{2}}}\right)^2 < \sigma^2,\right.$$
$$\left.\frac{(n-1)s^2}{\chi^2_{\frac{1+\sqrt{\gamma}}{2}}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{1-\sqrt{\gamma}}{2}}(n-1)}\right\} \qquad (3.7)$$

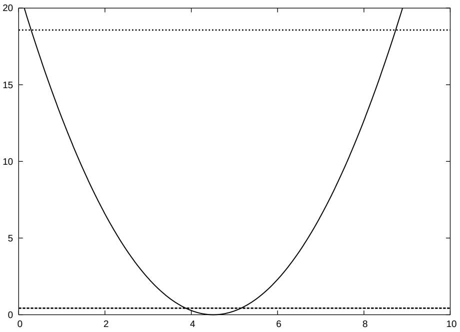
is a $100\gamma\%$ *simultaneous confidence region for* (μ, σ^2), for any sample realization $x_1, \dots, x_n$ of the random sample $X_1, \dots, X_n$ from the distribution $\mathcal{N}(\mu; \sigma^2)$ such that the sample mean and the sample variance take the values $\overline{x}$ and s^2, respectively.

If, for example, we observe the sample realization $x_1 = 4.87$, $x_2 = 5.06$, $x_3 = 2.8$, $x_4 = 5.32$, then a 95% simultaneous confidence region for (μ, σ^2) is the area of the (μ, σ^2) plane delimited by the equations:

$$\sigma^2 = 18.57$$

$$\sigma^2 = 1.0412(4.5125 - \mu)^2$$

$$\sigma^2 = 0.4284$$



3.2 Interval estimation in a general context

We now extend the concept of interval estimation to the case of a general statistical model, not necessarily Gaussian.

3.2.1 Confidence interval and confidence region

Definition 3.1. Let $X_1, \dots, X_n$ be a random sample from a population with c.d.f. $F(x; \theta)$ with $\theta \in \Theta \subset \mathbb{R}^p$. Let us consider a family of subsets of Θ, say $\{S(x_1, \dots, x_n), (x_1, \dots, x_n) \in \mathbb{R}^n\}$, obtained in correspondence of all sample realizations $(x_1, \dots, x_n)$ of $(X_1, \dots, X_n)$. Any set $S(x_1, \dots, x_n)$ can be seen as a sample realization of the random set $S(X_1, \dots, X_n)$.

Let $\gamma \in (0, 1)$. The random subset $S(X_1, \dots, X_n)$ of Θ is said to be a $100\gamma\%$ *confidence region (set) for* θ if

$$P_\theta([S(X_1, \dots, X_n) \ni \theta] \geq \gamma, \tag{3.8}$$

i.e. if the random set $S(X_1, \dots, X_n)$ contains the true value of the parameter θ with probability at least γ. Also the sample realization $S(x_1, \dots, x_n)$ of the random set $S(X_1, \dots, X_n)$ is named "confidence region of level γ for θ".

Definition 3.2. Let $X_1, \dots, X_n$ be a random sample from a population with c.d.f. $F(x; \theta)$ and $x_1, \dots, x_n$ a sample realization. Let $\kappa(\theta)$ be a one-dimensional characteristic of the population and let $T_1 = g_1(X_1, \dots, X_n)$ and $T_2 = g_2(X_1, \dots, X_n)$ be two real valued statistics; we denote by $t_1 = g_1(x_1, \dots, x_n)$ and $t_2 = g_2(x_1, \dots, x_n)$ the sample realizations of the two statistics. If

$$P_\theta[T_1 \leq \kappa(\theta) \leq T_2] \geq \gamma, \qquad \forall \theta \in \Theta\,, \tag{3.9}$$

we call both the random interval $[T_1, T_2]$ and its sample realization $[t_1, t_2]$ a $100\gamma\%$ *confidence interval for* $\kappa(\theta)$.

Definition 3.3. Let $\kappa(\theta)$ be a characteristic of the population. If T_U is a statistics such that

$$P_\theta[\kappa(\theta) \leq T_U] \geq \gamma, \qquad \forall \theta \in \Theta, \tag{3.10}$$

T_U is called a γ *upper confidence limit for* $\kappa(\theta)$. If T_L is a statistics such that

$$P_\theta[\kappa(\theta) \geq T_L] \geq \gamma, \qquad \forall \theta \in \Theta, \tag{3.11}$$

T_L is called a γ *lower confidence limit for* $\kappa(\theta)$.

In Eqs. (3.8), (3.9), (3.10) and (3.11) γ is called the *confidence level* and the probabilities in the left hand side of the equations are called the *coverage probabilities.*

Remark 3.12 (Precision). If the statistics T_1, T_2, T_U, T_L are continuous random variables, then, in order to construct the confidence intervals, in the Eqs. (3.9), (3.10) and (3.11) it is possible to consider equality in place of inequality $\geq$. In the discrete case, instead, this might be impossible. Diminishing the coverage probability without lowering the confidence level gives the possibility of obtaining a more precise interval estimation (a shorter interval).

Example 3.13. Let $X_1, \ldots, X_n$ be a random sample from a population with density $f(x, \theta) = (1/\theta)\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x)$, $\theta > 0$.

1. Find the MLE of the characteristic $\kappa(\theta) = \mathbb{E}_\theta[X]$.
2. Let T be the estimator of $\kappa(\theta)$ found in point 1. Determine the density of $\frac{2nT}{\theta}$.
3. Suppose now $n = 10$ and $\overline{x} = 3$. By using point 2, propose a 95% upper confidence bound for the characteristic $\kappa(\theta) = \mathbb{E}_\theta[X]$.

1. $\kappa(\theta) = \mathbb{E}_\theta(X) = \theta$ and $T = \overline{X}$ is an MLE of θ. Indeed, the likelihood function is

$$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, \theta) = \frac{1}{\theta^n} \exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right),$$

hence its natural logarithm is

$$\log L_\theta(x_1, \ldots, x_n) = -n \log \theta - \frac{\sum_{i=1}^n x_i}{\theta}.$$

Therefore,

$$\frac{\partial}{\partial \theta} \log L_\theta(x_1, \ldots, x_n) = \frac{1}{\theta}\left(\frac{\sum_{i=1}^n x_i}{\theta} - n\right) = 0$$

if and only if $\hat{\theta} = T = \frac{\sum_{i=1}^n x_i}{n} = \overline{X}$.

2. The variable $\sum_{j=1}^n X_j$ is a sum of i.i.d. random variables with density $\Gamma(1,\theta)$, therefore $\sum_{j=1}^n X_j \sim \Gamma(n,\theta)$. Moreover, if $W \sim \Gamma(a,b)$, then $cW \sim \Gamma(a,cb)$, $\forall c > 0$. Hence,

$$\frac{2nT}{\theta} = \frac{2\sum_{j=1}^n X_j}{\theta} \sim \Gamma(n,2) = \chi^2(2n).$$

3. Observe that

$$P_\theta\left(\theta \leq \frac{2nT}{k}\right) = P_\theta\left(\frac{2nT}{\theta} \geq k\right) = 0.95$$

if and only if $k = \chi^2_{0.05}(2n) = \chi^2_{0.05}(20) \simeq 10.9$. It follows that $(0, 2\cdot 3\cdot 10/10.9] \simeq (0, 5.505]$ is a 95% confidence interval for θ. In general, $\frac{2\sum_{j=1}^n x_j}{\chi_{1-\gamma}(2n)}$ is a $100\gamma\%$ confidence upper bound for θ, with $\gamma \in (0,1)$.

Exercise 3.14. Let $X_1,\dots,X_n$ be a random sample from a population with density $f(x,\theta) = 1/\theta \mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x)$, $\theta > 0$.

1. Propose a $100\gamma\%$ lower confidence bound for θ.
2. Propose a $100\gamma\%$ bilateral confidence interval for θ.

3.2.2 Pivotal quantity

All the examples of confidence intervals seen up to now were constructed starting from a random variable depending on the observations **and** the unknown parameter θ, but with a distribution not depending on θ.

Definition 3.4 (Pivotal quantity)**.** Let $X_1,\dots,X_n$ be a random sample from the density $f(x;\theta)$. Let $Q_\theta = q(X_1,\dots,X_n;\theta)$ be a random variable with distribution not depending on θ; Q_θ is defined to be a *pivotal quantity.*

Compare this new notion with the notion of statistics: a statistics is a random variable depending on the sample, but not on θ, whose distribution however depends on θ.

Example 3.15. (Gaussian case) If $X_1,\dots,X_n$ is a random sample form a normal population $\mathcal{N}(\mu;\sigma^2)$, then $\sqrt{n}\left(\overline{X}-\mu\right)/\sigma \sim \mathcal{N}(0;1)$, $\sqrt{n}\left(\overline{X}-\mu\right)/S \sim t(n-1)$, $\sum_{j=1}^n \left(\frac{X_j-\mu}{\sigma}\right)^2 \sim \chi^2(n)$, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ are all examples of pivotal quantities.

(Exponential case) If $X_1,\dots,X_n$ is a random sample form an exponential population $\mathcal{E}(\theta)$, then $2n\overline{X}/\theta \sim \chi^2(2n)$ is a pivotal quantity.

Remark 3.16 (Pivotal-quantity method). **First step.** If $Q_\theta = q(X_1,\dots,X_n;\theta)$ is a pivotal quantity and has a probability density, then for any fixed $\gamma \in (0,1)$ there will exist two numbers q_1, q_2 depending on γ (and not on θ) such that $P_\theta[q_1 < Q_\theta < q_2] = \gamma$.

Second step. Let us assume that it is possible to solve with respect to θ the two inequalities $q_1 < q(x_1,\dots,x_n;\theta) < q_2$ in the sense that, for any sample value $(x_1,\dots,x_n)$ and any value θ of the parameter, $q_1 < q(x_1,\dots,x_n;\theta) < q_2$ is equivalent to $t_1(x_1,\dots,x_n) < \kappa(\theta) < t_2(x_1,\dots,x_n)$, where $\kappa(\theta)$ is a characteristic of the population and $T_i = t_i(X_1, \dots, X_n)$, $i = 1, 2$, are two **statistics**. Then, $\gamma = P_\theta[q_1 < Q_\theta < q_2] = P_\theta[T_1 < \kappa(\theta) < T_2]$ and (T_1, T_2) is a $100\gamma\%$ confidence interval for $\kappa(\theta)$.

Let us note that for any fixed γ many couples of values q_1, q_2 may exist such that $P_\theta[q_1 < Q_\theta < q_2] = \gamma$ and, so, many confidence intervals can be constructed (in principle, infinitely many). We already noticed this fact in the case of the confidence intervals for the variance of a normal population, where we used the possibility of doing many choices in order to construct three types of confidence intervals.

3.2.3 Confidence intervals for large samples

Property 3 of Proposition 2.6 can be used to construct approximate confidence intervals for samples of large size. Indeed, the quantity $(T_n - \kappa(\theta)) \big/ \sqrt{\frac{(\kappa'(\theta))^2}{nI(\theta)}}$ has approximately a standard normal distribution and can be considered as an approximate pivotal quantity from which one can try to construct a confidence interval following the procedure of the previous section.

A typical example is the one of the estimation of a proportion. Let $X_1, \dots, X_n$ be a large sample from a population with density $\mathrm{Be}(\theta)$, $\theta \in (0, 1)$. Both the Central Limit Theorem and the theory of ML estimators give that the sample mean $\overline{X}_n$ is asymptotically normal with asymptotic mean θ and asymptotic variance $\theta(1-\theta)/n$. Then, for large n we have

$$P\left[-z_{1-\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} \leq z_{1-\alpha/2}\right] \simeq 1 - \alpha\,.$$

You can check that the two inequalities inside the probability are equivalent to $[T_- \leq \theta \leq T_+]$, where

$$T_\pm = \frac{2n\overline{X}_n + (z_{1-\alpha/2})^2 \pm z_{1-\alpha/2}\sqrt{4n\overline{X}_n\left(1 - \overline{X}_n\right) + (z_{1-\alpha/2})^2}}{2\left(n + (z_{1-\alpha/2})^2\right)}\,.$$

Therefore, $[T_-, T_+]$ is a confidence interval for the proportion θ of approximate level $1 - \alpha$. By discarding the terms of order higher than $1/\sqrt{n}$, one obtains the simpler, but worst, confidence interval

$$T_\pm \simeq \overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n\left(1 - \overline{X}_n\right)}{n}}\,.$$

Chapter 4

Hypothesis Tests

The subject of hypothesis tests is developed in our textbook [11]. Here we only review some terminology.

4.1 The ingredients in the theory of tests

We have a statistical model with some unknown parameters; Θ is the parameter space and θ is the generic unknown parameter (it can be one-dimensional, multi-dimensional, or even infinite-dimensional).

- We have two hypotheses on the possible values of the parameters.
  The null hypothesis: $H_0 : \theta \in \Theta_0$. The alternative hypothesis: $H_1 : \theta \in \Theta_1$.
  $\Theta_0 \cup \Theta_1 = \Theta, \ \Theta_0 \cap \Theta_1 = \emptyset$.

- Type I error: rejection of H_0 when it is true. Type II error: acceptance of H_0 when it is false. A type I error is worst than a type II error.

- A test for H_0 versus H_1 is a decision rule to reject H_0 (and to accept H_1) or to do the contrary.

  A test does not equally treat the two hypotheses, but it is constructed with a bias toward H_0 in order to reduce the possibility of the type I error. By this reason, when a test is in favour of the null hypothesis, the conclusion is considered *weak*, while, when the test is in favour of the alternative hypothesis, the conclusion is considered *strong*. When the data lead us to reject H_0, it is usual to say that they are *significant*.

- The rule to reject H_0 is given through a *critical region*. In Pestman [11] a critical region is a subset G of $\mathbb{R}^n$, the set of the possible values of an n-sample, such that if the sample $(X_1, \dots, X_n)$ falls into G, we reject H_0. We can say that we reject the null hypothesis when the event $\{(X_1, \dots, X_n) \in G\}$ is true. Also the event itself, which leads to the rejection decision, is called critical region.

- Let us consider a family of tests defined by the following decision rule, based on a suitable, fixed statistics T and on a numerical threshold x: "the null hypothesis is rejected when at the end of the experiment the event $\{T > x\}$ is true" and "do not reject the null hypothesis when at the end of the experiment the event $\{T \leq x\}$ is true". Now, $\{T > x\}$ is the *critical region* (CR) and T is called the *test statistics.* The case CR $= \{S < s\}$ is reduced to the previous one by taking $T = -S$ e $x = -s$; the case CR $= \{|S| > x\}$ is again equivalent to the first by taking $T = |S|$. The choice of a value for the threshold x picks up a single test from this family.

- The function $\pi(\theta) = P_\theta[T > x]$ is called the *power function* of the test. In principle $\pi(\theta)$ depends also on the threshold x, but we do not want to underline this dependence in the notation. When $\theta \in \Theta_0$, $\pi(\theta)$ is a probability of type I error; when $\theta \in \Theta_1$, $1 - \pi(\theta)$ is a probability of type II error and it is often denoted by β. A good test should have a large power function for $\theta \in \Theta_1$. In some books, as in Pestman [11], the power function is defined only for $\theta \in \Theta_1$.

- The supremum of the probabilities of type I error is called *size* of the test (or of the critical region), i.e. size$= \sup\limits_{\theta\in\Theta_0} \pi(\theta) = \sup\limits_{\theta\in\Theta_0} P_\theta[T > x]$.

  The test is said to be of (significance) level α if its size is less than or equal to α.

  Let us assume now that the supremum of the probabilities of type I error is a maximum and that it is realized for $\theta = \theta_0$, i.e. $\sup_{\theta\in\Theta_0} P_\theta[T > x] = P_{\theta_0}[T > x]$. Therefore, when $\{T > x\}$ is true, we reject H_0, while in the case $\{T \leq x\}$ true we cannot reject H_0. Moreover, the power is $\pi(\theta) = P_\theta[T > x]$ and, so, the size is $\sup_{\theta\in\Theta_0} P_\theta[T > x] = P_{\theta_0}[T > x]$.

- The way to treat the problem of choosing the threshold x is setting the size α, which means to keep all type I errors probabilities under a tolerable value. The typical choices are $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.10$. Fixed α, the threshold is computed in the following way. We have

$$\alpha = \sup_{\theta\in\Theta_0} P_\theta[T > x] = P_{\theta_0}[T > x] = 1 - P_{\theta_0}[T \leq x],$$

$$P_{\theta_0}[T \leq x] = 1 - \alpha\,,$$

  which gives $x = t^{(\theta_0)}_{1-\alpha}$, where $t^{(\theta_0)}_{1-\alpha}$ is the quantile of order $1-\alpha$ of the distribution of the statistics T when $\theta = \theta_0$.

- To evaluate the type II error probabilities, we have to be able to compute the power of the test as a function of θ, i.e. $\pi(\theta) = P_\theta[T > t^{(\theta_0)}_{1-\alpha}]$. Take $\theta \in \Theta_1$; then, the type II error probability, when the true value of the parameter is θ, is given by $\beta_\theta = 1 - \pi(\theta) = P_\theta[T \leq t^{(\theta_0)}_{1-\alpha}]$.

- Another useful notion is the one of *p-value* of the family of tests with test statistics T. **The p-value is the smallest level (or, better, the infimum) for which**

one should reject the null hypothesis with the data obtained in the experiment. Let us stress that from this definition one has that the p-value depends on the data and, so, it is a statistics.

Let us see how to compute the p-value. Let us indicate with $\alpha(x)$ the size of the test of threshold x and with t *the value taken by the statistics* T *in the experiment*, i.e. the value of T computed from the data. We have

$$\alpha(x) = \sup_{\theta \in \Theta_0} P_\theta[T > x] = P_{\theta_0}[T > x]$$

and we reject H_0 if $t > x$. Moreover, $\alpha(x)$ is a decreasing function of x. Therefore, we get the smallest size compatible with rejection by taking $x = t$. We have obtained: **if t is the value taken by the test statistics** T in the experiment, **the p-value is given by**

$$\text{p-value} = \alpha(t) = P_{\theta_0}[T > t].$$

The p-value is not a substitute for the level of the test. One has to fix the level according to his opinions about the consequences of type I error. Then, one can compute the p-value and compare it with the level: if the p-value is smaller than the level, he rejects H_0, otherwise not.

The p-value and the size say something about the probabilities of type I error: the size is the supremum of such probabilities. But also the type II error probabilities are important; they are given by $1 - \pi(\theta)$ for $\theta \in \Theta_1$. Therefore, also the knowledge of the power function is important.

4.2 Tests on mean and variance for normal populations

The tables of tests from page 57 to page 61 can be extracted and used during the written examinations.

4.2.1 Hypothesis tests on the mean of a normal population

$(x_1, \dots, x_n)$ = sample realization of $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu; \sigma^2)$.

σ^2 is known [Z-test]:

$\mathbf{H_0}$	$\mathbf{H_1}$	**Critical region**	**p-value**
$\mu{=}\mu_0$ $\mu{=}\mu_0$ $\mu{\leq}\mu_0$	$\mu{=}\mu_1$ with $\mu_0{<}\mu_1$ $\mu{>}\mu_0$ $\mu{>}\mu_0$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$	$1 - \Phi\left(\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$
$\mu{=}\mu_0$ $\mu{=}\mu_0$ $\mu{\geq}\mu_0$	$\mu{=}\mu_1$ with $\mu_0{>}\mu_1$ $\mu{<}\mu_0$ $\mu{<}\mu_0$	$\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{1-\alpha}$	$\Phi\left(\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\dfrac{\lvert\overline{X} - \mu_0\rvert}{\sigma/\sqrt{n}} \geq z_{1-\frac{\alpha}{2}}$	$2\left[1 - \Phi\left(\dfrac{\lvert\bar{x} - \mu_0\rvert}{\sigma/\sqrt{n}}\right)\right]$

σ^2 is unknown [t-test]:

$\mathbf{H_0}$	$\mathbf{H_1}$	**Critical region**	**p-value**
$\mu{=}\mu_0$ $\mu{=}\mu_0$ $\mu{\leq}\mu_0$	$\mu{=}\mu_1$ with $\mu_0{<}\mu_1$ $\mu{>}\mu_0$ $\mu{>}\mu_0$	$\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq t_{n-1}(1-\alpha)$	$1 - P\left(T_{n-1} \leq \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}\right)$
$\mu{=}\mu_0$ $\mu{=}\mu_0$ $\mu{\geq}\mu_0$	$\mu{=}\mu_1$ with $\mu_0{>}\mu_1$ $\mu{<}\mu_0$ $\mu{<}\mu_0$	$\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq -t_{n-1}(1-\alpha)$	$P\left(T_{n-1} \leq \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}\right)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\dfrac{\lvert\overline{X} - \mu_0\rvert}{S/\sqrt{n}} \geq t_{n-1}\left(1 - \dfrac{\alpha}{2}\right)$	$2\left[1 - P\left(T_{n-1} \leq \dfrac{\lvert\bar{x} - \mu_0\rvert}{s/\sqrt{n}}\right)\right]$

$\bar{x}$ = sample mean of $x_1, \dots, x_n$
s^2 = sample variance of $x_1, \dots, x_n$
Φ = c.d.f. of $\mathcal{N}(0; 1)$ and z_p such that $\Phi(z_p) = p$
$T_{n-1} \sim t(n-1)$ and $t_{n-1}(p)$ such that $P(T_{n-1} \leq t_{n-1}(p)) = p$.

4.2.2 χ^2-test on the variance of a normal population

μ is known: $s_0^2 = (1/n)\sum_{j=1}^{n}(x_j - \mu)^2$

$\mathbf{H_0}$	$\mathbf{H_1}$	**Critical region**	**p-value**
$\sigma^2{=}\sigma_0^2$ $\sigma^2{=}\sigma_0^2$ $\sigma^2{\leq}\sigma_0^2$	$\sigma^2{=}\sigma_1^2$ with $\sigma_0^2{<}\sigma_1^2$ $\sigma^2{>}\sigma_0^2$ $\sigma^2{>}\sigma_0^2$	$\dfrac{nS_0^2}{\sigma_0^2} \geq \chi_n^2(1-\alpha)$	$1 - F_n\left(\dfrac{ns_0^2}{\sigma_0^2}\right)$
$\sigma^2{=}\sigma_0^2$ $\sigma^2{=}\sigma_0^2$ $\sigma^2{\geq}\sigma_0^2$	$\sigma^2{=}\sigma_1^2$ with $\sigma_0^2{>}\sigma_1^2$ $\sigma^2{<}\sigma_0^2$ $\sigma^2{<}\sigma_0^2$	$\dfrac{nS_0^2}{\sigma_0^2} \leq \chi_n^2(\alpha)$	$F_n\left(\dfrac{ns_0^2}{\sigma_0^2}\right)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\dfrac{nS_0^2}{\sigma_0^2} \geq \chi_n^2\left(1-\dfrac{\alpha}{2}\right)$ or $\dfrac{nS_0^2}{\sigma_0^2} \leq \chi_n^2\left(\dfrac{\alpha}{2}\right)$	$2\min\{p_1, p_2\}$ where $p_1 = F_n\left(\dfrac{ns_0^2}{\sigma_0^2}\right)$ and $p_2 = 1 - p_1$

μ is unknown: $F_n =$ c.d.f. of $\chi^2(n)$ and $\chi_n^2(p)$ such that $F_n(\chi_n^2(p)) = p$

$\mathbf{H_0}$	$\mathbf{H_1}$	**Critical region**	**p-value**
$\sigma^2{=}\sigma_0^2$ $\sigma^2{=}\sigma_0^2$ $\sigma^2{\leq}\sigma_0^2$	$\sigma^2{=}\sigma_1^2$ with $\sigma_0^2{<}\sigma_1^2$ $\sigma^2{>}\sigma_0^2$ $\sigma^2{>}\sigma_0^2$	$\dfrac{(n-1)S^2}{\sigma_0^2} \geq \chi_{n-1}^2(1-\alpha)$	$1 - F_{n-1}\left(\dfrac{(n-1)s^2}{\sigma_0^2}\right)$
$\sigma^2{=}\sigma_0^2$ $\sigma^2{=}\sigma_0^2$ $\sigma^2{\geq}\sigma_0^2$	$\sigma^2{=}\sigma_1^2$ with $\sigma_0^2{>}\sigma_1^2$ $\sigma^2{<}\sigma_0^2$ $\sigma^2{<}\sigma_0^2$	$\dfrac{(n-1)S^2}{\sigma_0^2} \leq \chi_{n-1}^2(\alpha)$	$F_{n-1}\left(\dfrac{(n-1)s^2}{\sigma_0^2}\right)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\dfrac{(n-1)S^2}{\sigma_0^2} \geq \chi_{n-1}^2\left(1-\dfrac{\alpha}{2}\right)$ or $\dfrac{(n-1)S^2}{\sigma_0^2} \leq \chi_{n-1}^2\left(\dfrac{\alpha}{2}\right)$	$2\min\{p_1, p_2\}$ where $p_1 = F_{n-1}\left(\dfrac{(n-1)s^2}{\sigma_0^2}\right)$ and $p_2 = 1 - p_1$

4.2.3 Tests on the comparison of the means of two normal populations

Paired samples

$(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. $\sim \mathcal{N}\left(\begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}\right)$ and $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ unknown. To verify the hypotheses

- $H_0 : \mu_X \leq \mu_Y + \Delta$ versus $H_1 : \mu_X > \mu_Y + \Delta$
- $H_0 : \mu_X \geq \mu_Y + \Delta$ versus $H_1 : \mu_X < \mu_Y + \Delta$
- $H_0 : \mu_X = \mu_Y + \Delta$ versus $H_1 : \mu_X \neq \mu_Y + \Delta$

use the suitable t-test starting from the sample $X_1 - Y_1, \dots, X_n - Y_n$.

Independent samples

$(x_1, \dots, x_m)$ = sample realization of $\boldsymbol{X} = X_1, \dots, X_m$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$,
$(y_1, \dots, y_n)$ = sample realization of $\boldsymbol{Y} = Y_1, \dots, Y_n$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$
$\boldsymbol{X}$, $\boldsymbol{Y}$ independent.

σ_X^2, σ_Y^2 are known [z-test]:

$\mathbf{H_0}$	$\mathbf{H_1}$	**Critical region**	**p-value**
$\mu_X = \mu_Y + \Delta$ $\mu_X \leq \mu_Y + \Delta$	$\mu_X > \mu_Y + \Delta$	$\dfrac{\overline{X} - \overline{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \geq z_{1-\alpha}$	$1 - \Phi\left(\dfrac{\bar{x} - \bar{y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}\right)$
$\mu_X = \mu_Y + \Delta$ $\mu_X \geq \mu_Y + \Delta$	$\mu_X < \mu_Y + \Delta$	$\dfrac{\overline{X} - \overline{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \leq -z_{1-\alpha}$	$\Phi\left(\dfrac{\bar{x} - \bar{y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}\right)$
$\mu_X = \mu_Y + \Delta$	$\mu_X \neq \mu_Y + \Delta$	$\dfrac{\lvert\overline{X} - \overline{Y} - \Delta\rvert}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \geq z_{1-\frac{\alpha}{2}}$	$2\left[1 - \Phi\left(\dfrac{\lvert\bar{x} - \bar{y} - \Delta\rvert}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}\right)\right]$

The variances are unknown, but equal: $\sigma_X^2 = \sigma_Y^2$ [t-test].

$\mathbf{H_0}$	$\mathbf{H_1}$	**Critical region**	**p-value**
$\mu_X{=}\mu_Y{+}\Delta$ $\mu_X{\leq}\mu_Y{+}\Delta$	$\mu_X{>}\mu_Y{+}\Delta$ $\mu_X{>}\mu_Y{+}\Delta$	$\dfrac{\overline{X}-\overline{Y}-\Delta}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \geq t_{m+n-2}(1-\alpha)$	$1-P\left(T_{m+n-2} \leq \dfrac{\bar{x}-\bar{y}-\Delta}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}\right)$
$\mu_X{=}\mu_Y{+}\Delta$ $\mu_X{\geq}\mu_Y{+}\Delta$	$\mu_X{<}\mu_Y{+}\Delta$ $\mu_X{<}\mu_Y{+}\Delta$	$\dfrac{\overline{X}-\overline{Y}-\Delta}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \leq -t_{m+n-2}(1-\alpha)$	$P\left(T_{m+n-2} \leq \dfrac{\bar{x}-\bar{y}-\Delta}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}\right)$
$\mu_X{=}\mu_Y{+}\Delta$	$\mu_X{\neq}\mu_Y{+}\Delta$	$\dfrac{\lvert\overline{X}-\overline{Y}-\Delta\rvert}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \geq t_{m+n-2}(1-\dfrac{\alpha}{2})$	$2-2P\left(T_{m+n-2} \leq \dfrac{\lvert\bar{x}-\bar{y}-\Delta\rvert}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}\right)$

Pooled variance: $S_p^2 = \frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}$, with S_X^2 = sample variance of $\boldsymbol{X}$ and S_Y^2 = sample variance of $\boldsymbol{Y}$. $T_{m+n-2} \sim t(m+n-2)$.

4.2.4 F-test on the comparison of the variances of two normal populations

$(x_1,\ldots,x_m)$ = sample realization of $\boldsymbol{X} = X_1,\ldots,X_m$ i.i.d. $\sim N(\mu_X,\sigma_X^2)$,
$(y_1,\ldots,y_n)$ = sample realization of $\boldsymbol{Y} = Y_1,\ldots,Y_n$ i.i.d. $\sim N(\mu_Y,\sigma_Y^2)$
$\boldsymbol{X}$, $\boldsymbol{Y}$ independent.

μ_X, μ_Y **are known:**

H$_0$	**H$_1$**	**Critical region**	**p-value**
$\sigma^2_X \le \sigma^2_Y$	$\sigma^2_X > \sigma^2_Y$	$\dfrac{S^2_{0X}}{S^2_{0Y}} \ge F_{m,n}(1-\alpha)$	$1 - P\left(F_{m,n} \le \dfrac{s^2_{0X}}{s^2_{0Y}}\right)$
$\sigma^2_X \ge \sigma^2_Y$	$\sigma^2_X < \sigma^2_Y$	$\dfrac{S^2_{0X}}{S^2_{0Y}} \le F_{m,n}(\alpha)$	$P\left(F_{m,n} \le \dfrac{s^2_{0X}}{s^2_{0Y}}\right)$
$\sigma^2_X = \sigma^2_Y$	$\sigma^2_X \ne \sigma^2_Y$	$S^2_{0X}/S^2_{0Y} \ge F_{m,n}(1-\alpha/2)$ or $S^2_{0X}/S^2_{0Y} \le F_{m,n}(\alpha/2)$	$2\min\{p_1,p_2\}$ where $p_1 = P\big(F_{m,n} \le s^2_{0X}/s^2_{0Y}\big)$ and $p_2 = 1-p_1$

$$S^2_{0X} := \frac{\sum_{j=1}^m (X_j - \mu_X)^2}{m}, \qquad S^2_{0Y} := \frac{\sum_{j=1}^n (Y_j - \mu_Y)^2}{n}.$$

$F_{a,b}$ = r.v. having Fisher density with (a,b) degrees of freedom and $F_{a,b}(p)$ such that $P(F_{a,b} \le F_{a,b}(p)) = p$.

μ_X, μ_Y **are unknown:**

H$_0$	**H$_1$**	**Critical region**	**p-value**
$\sigma^2_X \le \sigma^2_Y$	$\sigma^2_X > \sigma^2_Y$	$\dfrac{S^2_X}{S^2_Y} \ge F_{m-1,n-1}(1-\alpha)$	$1 - P\left(F_{m-1,n-1} \le \dfrac{s^2_X}{s^2_Y}\right)$
$\sigma^2_X \ge \sigma^2_Y$	$\sigma^2_X < \sigma^2_Y$	$\dfrac{S^2_X}{S^2_Y} \le F_{m-1,n-1}(\alpha)$	$P\left(F_{m-1,n-1} \le \dfrac{s^2_X}{s^2_Y}\right)$
$\sigma^2_X = \sigma^2_Y$	$\sigma^2_X \ne \sigma^2_Y$	$S^2_X/S^2_Y \ge F_{m-1,n-1}(1-\alpha/2)$ or $S^2_X/S^2_Y \le F_{m-1,n-1}(\alpha/2)$	$2\min\{p_1,p_2\}$ where $p_1 = P\big(F_{m-1,n-1} \le s^2_X/s^2_Y\big)$ and $p_2 = 1-p_1$

4.3 Relation between hypothesis testing and interval estimation

The theories of hypothesis testing and interval estimation are strictly related. Here we give an idea of this relation by showing how to pass from confidence intervals to hypothesis tests and from tests to confidence regions. As usual, in the following, $X_1, \dots, X_n$ denotes a random sample from a population with unknown parameter $\theta \in \Theta$.

4.3.1 From interval estimation to hypothesis testing

From bilateral confidence intervals to bilateral hypothesis tests

Let $\kappa(\theta)$ be a population characteristic and $[T_1, T_2]$ a confidence interval of level $1-\alpha$ for $\kappa(\theta)$, i.e.

$$P_\theta[T_1 \leq \kappa(\theta) \leq T_2] = 1-\alpha, \qquad \forall \theta \in \Theta\,. \tag{4.1}$$

Let k_0 be a fixed number and consider the statistical hypotheses

$$H_0 : \kappa(\theta) = k_0\,, \qquad \text{versus} \qquad H_1 : \kappa(\theta) \neq k_0\,. \tag{4.2}$$

Then, the critical region defined as

$$\mathrm{CR} = \{k_0 \notin [T_1, T_2]\} \tag{4.3}$$

gives a test of size α for the hypotheses (4.2).

Let us prove this statement. For those values θ for which H_0 holds true we can write $\kappa(\theta)$ instead of k_0 (they are equal under H_0). By using this observation and Eq. (4.1), we get

$$\begin{aligned} \text{size} = \sup_{\theta:\kappa(\theta)=k_0} P_\theta\left(k_0 \notin [T_1,T_2]\right) &= \sup_{\theta:\kappa(\theta)=k_0} P_\theta\left(\kappa(\theta) \notin [T_1,T_2]\right) \\ &= \sup_{\theta:\kappa(\theta)=k_0} \{1 - P_\theta\left(T_1 \leq \kappa(\theta) \leq T_2\right)\} = \sup_{\theta:\kappa(\theta)=k_0} \{1-(1-\alpha)\} = \alpha\,. \end{aligned}$$

In the last step we suppressed the supremum because there is no more dependence on θ in its argument.

Example 4.1. Let $X_1, \dots, X_n$ be i.i.d. $\sim N(\mu, \sigma^2)$. If the variance is known, a $100(1-\alpha)\%$ bilateral confidence interval for the mean μ has end points $T_1 = \overline{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}$, $T_2 = \overline{X} + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}$, and, according to Eq. (4.3), the critical region for a test of size α for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ is

$$\begin{aligned} \{\mu_0 \notin [T_1, T_2]\} &= \{\mu_0 \leq T_1\} \cup \{\mu_0 \geq T_2\} \\ &= \left\{ z_{1-\alpha/2} \leq \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \right\} \cup \left\{ -z_{1-\alpha/2} \geq \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \right\} = \left\{ \left| \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \right| \geq z_{1-\alpha/2} \right\}. \end{aligned}$$

This is exactly the test given at p. 57.

In the case of unknown variance, from the confidence interval with end points $\overline{X} \pm t_{1-\alpha/2}(n-1)\,\frac{S}{\sqrt{n}}$, we get the critical region $\left\{\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| \geq t_{1-\alpha/2}(n-1)\right\}$ (again the test given at p. 57).

The unilateral case

Let $\kappa(\theta)$ be a population characteristic and T a confidence upper bound of level $1-\alpha$ for $\kappa(\theta)$, i.e.

$$P_\theta[\kappa(\theta) \leq T] = 1-\alpha, \qquad \forall\theta \in \Theta\,. \tag{4.4}$$

Let k_0 be a fixed number and consider the statistical hypotheses

$$H_0 : \kappa(\theta) \geq k_0\,, \qquad \text{versus} \qquad H_1 : \kappa(\theta) < k_0\,. \tag{4.5}$$

Then, the critical region defined as

$$\mathrm{CR} = \{k_0 > T\} \tag{4.6}$$

gives a test of size α for the hypotheses (4.5).

Let us prove this statement. Let θ_0 be such that $\kappa(\theta) = k_0$; by using Eq. (4.4), we get $P_{\theta_0}(k_0 > T) = P_{\theta_0}\big(\kappa(\theta_0) > T\big) = \alpha$. Let θ' be such that $\kappa(\theta') > k_0$; then, we get $P_{\theta'}(k_0 > T) \leq P_{\theta'}\big(\kappa(\theta') > T\big) = \alpha$. Therefore, we have

$$\text{size} = \sup_{\theta:\kappa(\theta)\geq k_0} P_\theta\,(k_0 > T) = \sup_{\theta:\kappa(\theta)=k_0} P_\theta\,(T < k_0) = \sup_{\theta:\kappa(\theta)=k_0} P_\theta\,(T < \kappa(\theta)) = \alpha\,.$$

A similar reasoning applies to confidence lower bounds. Application to the mean or to the variance of a normal population gives the unilateral tests at pp. 57, 58.

4.3.2 From hypothesis tests to confidence regions

Let us consider a family of tests for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ all at level α with critical region $(X_1, \ldots, X_n) \in G_{\theta_0}$; we have a different test for any θ_0 in Θ. The critical region is a subset of $\mathbb{R}^n$ and we denote by $A(\theta_0)$ its complement. Then

$$\mathrm{CS} := \{\theta \in \Theta :\ (X_1, \ldots, X_n) \in A(\theta)\} \tag{4.7}$$

is a random subset of Θ (it varies according to the sample realizations) such that

$$P_{\theta_0}[\mathrm{CS} \ni \theta_0] = P_{\theta_0}[(X_1, \ldots, X_n) \in A(\theta_0)] = 1 - P_{\theta_0}[(X_1, \ldots, X_n) \in G_{\theta_0}] = 1-\alpha.$$

Hence, CS is a $100(1-\alpha)\%$ confidence set (region) for θ.

Chapter 5

Non-parametric inference

We consider now the problem of making inferences on a cumulative distribution F, when, differently from what we have done so far, we suppose to be nearly completely ignorant about F. We have not an unknown finite-dimensional parameter, but we can think to an infinite-dimensional one; in this sense we speak of *non-parametric* methods of inference. We speak also of *distribution-free* methods, because in many cases the only information we need about F is whether it is absolutely continuous or discrete.

5.1 A remark on the p-value

In the following sections we shall introduce some tests for which we give the critical region for a specified level of significance. Obviously, as in the parametric case, also the notion of p-value applies and, indeed, in the examples we often compute the p-value.

In many of the following tests the null hypothesis is simple and it is worthwhile to underline a peculiarity of the p-value for the case of a simple null hypothesis.

Let H_0 be simple and let the critical region be given by $\{T > x\}$, where T is a test statistics. If the value of T in the experiment is t, the p-value is $P_{H_0}[T > t]$. If we introduce the c.d.f. of T under H_0, $F_T^0(x) = P_{H_0}[T \le x]$, then the p-value is the statistics

$$p^* = 1 - F_T^0(T)\,.$$

If F_T^0 is continuous and invertible, for any $u \in [0,1]$, we have

$$\begin{aligned}\{p^* \le u\} = \{1 - F_T^0(T) \le u\} &= \{F_T^0(T) \ge 1 - u\} \\ &= \{F_T^0(T) < 1 - u\}^{\mathrm{c}} = \{T < {F_T^0}^{-1}(1-u)\}^{\mathrm{c}}\,,\end{aligned}$$

$$\begin{aligned}P_{H_0}[p^* \le u] = 1 - P_{H_0}[T < {F_T^0}^{-1}(1-u)] &= 1 - P_{H_0}[T \le {F_T^0}^{-1}(1-u)] \\ &= 1 - F_T^0\big({F_T^0}^{-1}(1-u)\big) = 1 - (1-u) = u\,.\end{aligned}$$

Therefore, under H_0, the p-value follows a uniform distribution $\mathcal{U}(0,1)$.

5.2 Empirical cumulative distribution

See also [11]: VII.1, VII.3 (only the statement of Glivenko-Cantelli theorem).

Let $X_1, \ldots, X_n$ be a random sample extracted from the c.d.f. F. In order to estimate F, we introduce the empirical cumulative distribution associated to the sample.

Definition 5.1. The *empirical* (or *sample*) *cumulative distribution* $\hat{F}_n$, associated to the sample $X_1, \ldots, X_n$, is a function from $\mathbb{R}$ into $[0,1]$ defined by

$$\hat{F}_n(x) = \frac{\#\{j : X_j \le x\}}{n}, \qquad \forall x \in \mathbb{R}\,. \tag{5.1}$$

Let us arrange the observations $X_1, \ldots, X_n$ in increasing order of magnitude and let us denote by $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ the resulting sequence. The realizations of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ depend only on the observed sample, hence $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are statistics, named *order statistics.* You already know $X_{(1)}$ and $X_{(n)}$: $X_{(1)}$ is the minimum and $X_{(n)}$ is the maximum among the observations.

We can rewrite $\hat{F}_n$ in terms of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ as follows:

$$\hat{F}_n(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \frac{k}{n} & \text{if } X_{(k)} \le x < X_{(k+1)}\,,\ k = 1, \ldots, n-1 \\ 1 & \text{if } x \ge X_{(n)} \end{cases} \tag{5.2}$$

Obviously the function $\hat{F}_n$ is random: as the sample realization changes, $\hat{F}_n$ changes consequently. Moreover, it follows immediately from (5.2) that $\hat{F}_n$ depends only on the sample (hence, it is a statistics) and that it is a discrete c.d.f. (it is a non-decreasing step function).

Example 5.1. Suppose we observed the sample realization $1, 0, 1, -1, 3, 2.5, 3, 1$. Thus, $\hat{F}_8$ is:

$$\hat{F}_8(x) = \begin{cases} 0 & x < -1 \\ 1/8 & -1 \le x < 0 \\ 2/8 & 0 \le x < 1 \\ 5/8 & 1 \le x < 2.5 \\ 6/8 & 2.5 \le x < 3 \\ 1 & x \ge 3 \end{cases} \tag{5.3}$$

Let us now introduce some synthesis values of $\hat{F}_n$. We will find some sample statistics we already know. Indeed:

- The rth moment of $\hat{F}_n$ is defined as

$$M_r = \frac{1}{n} \sum_{j=1}^{n} {X_j}^r$$

i.e. it is the already known sample moment of order r.

- The mean of $\hat{F}_n$ is $M_1 = \overline{X}$, i.e. the sample mean.
- The variance of $\hat{F}_n$ is
$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2 = \frac{n-1}{n} S^2 ,$$
which is proportional to the sample variance.
- The *median*[1] of $\hat{F}_n$ is
$$\hat{q}_{1/2} = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \\ \dfrac{X_{(n/2)} + X_{(n/2+1)}}{2} & \text{if } n \text{ is even.} \end{cases}$$

Let us now analyze the probabilistic properties of the statistics $\hat{F}_n(x)$.

For a fixed x, let us introduce the random variables $Y_1, \dots, Y_n$ defined by $Y_j = \mathbf{1}_{(-\infty,x]}(X_j)$, $j = 1, \dots, n$; the random variables $Y_1, \dots, Y_n$ are i.i.d. $\sim \text{Be}(F(x))$ and $\hat{F}_n(x) = \overline{Y}$, i.e. $\hat{F}_n(x)$ is the sample mean of n observations from a Bernoulli density with parameter $F(x)$. By this remark, it is immediate to prove the following properties of $\hat{F}_n$:

1. $n\hat{F}_n(x) \sim \mathcal{B}(n, F(x))$ for any $n \geq 1$;
2. $\mathbb{E}_F(\hat{F}_n(x)) = F(x)$, $\forall F$, $\forall x$, i.e. $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$;
3. $\text{Var}_F\left(\hat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n} \to 0$, as $n \to \infty$;
4. the sequence of estimators $\{\hat{F}_n(x)\}_n$ is MSE-consistent for $F(x)$ (this follows from the previous points).

A stronger result than the one in point 4 holds. We have also that in some sense $\hat{F}_n \to F$ uniformly in x. The precise statement is the Glivenko-Cantelli theorem hereafter:

5. *Let $X_1, X_2, \dots$ be i.i.d. random variables $\sim F$. Then*
$$P\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F(x)\right| = 0\right) = 1$$
That is, the "random function" $\hat{F}_n$ is, in some sense, globally consistent as an estimator of F.

[1] The median of a distribution F is the quantile of order 1/2 of F.

6. The sequence $\{\hat{F}_n(x)\}_n$ is asymptotically gaussian, i.e.

$$\lim_{n\to\infty} P_F\left(\sqrt{n}\,\frac{\hat{F}_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]}} \le z\right) = \Phi(z), \quad \forall z \in \mathbb{R}$$

$$\forall x \in \mathbb{R} \text{ s.t. } 0 < F(x) < 1$$

In order to prove this statement it is sufficient to recall that $\hat{F}_n(x)$ is the sample mean of Bernoulli i.i.d. random variables and to apply the central limit theorem.

5.3 Goodness-of-Fit Tests

We now show some examples of problems which can be faced in the case of a single sample of n i.i.d. observations.

Let $X_1, \ldots, X_n$ be a random sample from F. We want to establish whether the distribution F which generated the sample is an assigned and completely specified distribution F_0, or not. In other words, we want to construct a procedure to verify the null simple hypothesis $H_0 : F = F_0$ versus the composite alternative hypothesis $H_1 : F \neq F_0$. For instance:

$$H_0 : F = \mathcal{N}(0,1) \quad \text{versus} \quad H_1 : F \neq \mathcal{N}(0,1),$$

or

$$H_0 : F = \mathcal{P}(2) \quad \text{versus} \quad H_1 : F \neq \mathcal{P}(2).$$

We might also be interested to verify the null composite hypothesis that F belongs to a certain family of c.d.f. $\mathcal{F}_0$ specified up to some m-dimensional parameter, that is $\mathcal{F}_0 = \{F(\cdot,\theta), \theta \in \Theta \subset \mathbb{R}^m\}$, versus the alternative hypothesis that F does not belong to $\mathcal{F}_0$. For instance:

$$H_0 : F \in \{\mathcal{N}(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\} \quad \text{versus} \quad H_1 : F \text{ is not Gaussian}$$

or

$$H_0 : F \in \{\mathcal{P}(\lambda), \lambda > 0\} \quad \text{versus} \quad H_1 : F \text{ is not Poisson.}$$

A test to verify this kind of hypothesis on F (both in the case that the hypothesized distribution for F is completely specified and the case where it is not) is called *goodness of fit test*.

The goodness-of-fit tests we are going to treat are based on a statistics which measures the deviation of the empirical c.d.f. associated to the sample $\hat{F}_n$ from the c.d.f. of the distribution specified in the null hypothesis. The exact distribution of the statistics under the null hypothesis is sometimes known, but often it is obtained trough Monte Carlo simulations. For large samples explicit asymptotical expressions are available.

5.3.1 Kolmogorov-Smirnov test

See Section VII.4 in [11].

Let $X_1, \ldots, X_n$ be a random sample from F. In order to test $H_0 : F = F_0$ versus $H_1 : F \neq F_0$, we consider the statistics

$$D_n := \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right|$$

If H_0 is true, the Glivenko-Cantelli theorem guarantees that, with probability 1, $D_n \to 0$ as n increases. Therefore, if H_0 is true, we expect small values of D_n. On the basis of this observation, we can adopt this decisional rule:

"reject H_0 if D_n is large and accept H_0 if D_n is small."

As usual, we quantify the concepts of "large, small" in terms of significance level α, solving the equation

$$P_0(D_n > k) = \alpha.$$

Hence, it is necessary to investigate the c.d.f. of D_n under H_0. The following result holds:

Proposition 5.1. *If $X_1, \ldots, X_n$ is a random sample extracted from F_0 and F_0 is continuous, then the c.d.f. of D_n does not depend on F_0 (in other words, the c.d.f. of D_n is the same in the class of all continuous c.d.f. F_0).*

Proof. For simplicity, we prove the Proposition in the case that F_0 is strictly increasing.

If F_0 is strictly increasing, then the inverse function F_0^{-1} exists. If $X \sim F_0$ then $U := F_0(X) \sim \mathcal{U}(0, 1)$, in that $F_U(u) = 0$ if $u \leq 0$, $F_U(u) = 1$ if $u \geq 1$, while for $0 < u < 1$ we have

$$F_U(u) = P[F_0(X) \leq u] = P[X \leq F_0^{-1}(u)] = F_0(F_0^{-1}(u)) = u.$$

Moreover

$$\hat{F}_n(x) = \frac{\#\{j : X_j \leq x\}}{n} = \frac{\#\{j : F_0(X_j) \leq F_0(x)\}}{n} = \frac{\#\{j : U_j \leq F_0(x)\}}{n}$$

where $U_j := F_0(X_j)$, $j = 1, \ldots, n$. Finally,

$$D_n = \sup_{x \in \mathbb{R}} \left| \frac{\#\{j : U_j \leq F_0(x)\}}{n} - F_0(x) \right| = \sup_{u \in [0,1]} \left| \frac{\#\{j : U_j \leq u\}}{n} - u \right|.$$

It follows that the c.d.f. of D_n under H_0 coincides with the distribution of the statistics D_n that you would obtain if the sample were extracted from the uniform $\mathcal{U}(0, 1)$ distribution. □

The test statistics D_n is called *Kolmogorov-Smirnov statistics* and the test above is named *Kolomogorov-Smirnov test.*

The c.d.f. of D_n under H_0 (valid whatever F_0 is, provided it is continuous) is tabulated on [11] for some values of n and α. Moreover, for $n \to \infty$, the closed expression of the asymptotic c.d.f. of D_n is known:

$$\lim_{n\to\infty} P_0\left(\sqrt{n}D_n \le z\right) = H(z)\,, \qquad \forall z \in \mathbb{R}\,,$$

where

$$H(z) = 1 - 2\sum_{i=0}^{\infty}(-1)^{i-1}\mathrm{e}^{-2i^2z^2}\,.$$

Example 5.2. Consider the following sample of size 4:

$$1.126, 3.104, 2.577, 2.372.$$

By using the Kolmogorov-Smirnov test, verify the hypothesis H_0 that the sample is extracted from an exponential distribution with parameter 1 at significance levels 1% and 10%.

Let us calculate the Kolmogorov-Smirnov statistics D_4:

$x_j =$	1.126	2.372	2.577	3.104
$F_0(x_j) =$	$1-\mathrm{e}^{-1.126}$ $\simeq 0.6757$	$1-\mathrm{e}^{-2.372}$ $\simeq 0.9067$	$1-\mathrm{e}^{-2.577}$ $\simeq 0.9240$	$1-\mathrm{e}^{-3.104}$ $\simeq 0.9551$
$\hat{F}_4(x_j) =$	0.25	0.5	0.75	1
$\lvert\hat{F}_4(x_j) - F_0(x_j)\rvert =$	0.4257	0.4067	0.174	0.0449
$\lvert\hat{F}_4(x_{j-1}) - F_0(x_j)\rvert =$	$\boxed{0.6757}$	0.6567	0.4240	0.2051

hence $D_4 = 0.6757$. The p-value is 0.05186, therefore we accept H_0 at level 1%, while we reject H_0 at level 10%.[2]

Confidence bands for F. The fact that the c.d.f. of $D_n = \sup_x |\hat{F}_n(x) - F(x)|$ is the same in the class of all continuous F, makes it possible to construct a "confidence band for F" in the following way: let us fix γ and n and let us calculate, through the statistical tables (or through an appropriate software), the quantile of order γ of the c.d.f. of the Kolmogorov-Smirnov statistics, i.e. the value $q_n(\gamma)$ such that $P[D_n \le q_n(\gamma)] = \gamma$. This equality is equivalent to

$$P_F\left[\hat{F}_n(x) - q_n(\gamma) \le F(x) \le \hat{F}_n(x) + q_n(\gamma), \quad \forall x \in \mathbb{R}\right] = \gamma$$

[2] The sample has been generated from the $\Gamma(3,1)$ distribution. Using the package *ctest* of the software *R* (`http://cran.r-project.org`) we obtain:

```
ks.test(c(1.126, 3.104, 2.577, 2.372), "pgamma",1, 1)
One-sample Kolmogorov-Smirnov test
data:  c(1.126, 3.104, 2.577, 2.372)
D = 0.6757, p-value = 0.05186
alternative hypothesis:  two.sided
```

and, so,

$$\Big\{\text{continuous c.d.f. } F \text{ such that } \hat{F}_n(x) - q_n(\gamma) \le F(x) \le \hat{F}_n(x) + q_n(\gamma), \quad \forall x \in \mathbb{R}\Big\}$$

is a *confidence band of level γ for F.*

For example: if $\gamma = 0.95$ and $n = 8$ then $q_\gamma \simeq 0.4543$. If we consider the value of the "random function" $\hat{F}_8$ for the sample realization $(x_1, \dots, x_8)$ in Example 5.1, then the set of all continuous c.d.f. F such that $\hat{F}_8(x) - 0.4543 \le F(x) \le \hat{F}_8(x) + 0.4543$, $\forall x \in \mathbb{R}$, is a confidence band of level 95% for F.

5.3.2 χ^2 test for categorical or discrete data

See Section III.3 in [11].

Differently from the Kolmogorov-Smirnov test, which can be used only if the data are continuous and the null hypothesis is simple, the χ^2 goodness-of-fit test we are going to introduce is also suitable for hypothesis problems of type *a)* $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ for discrete data and *b)* $H_0 : F \in \mathcal{F}_0$ versus $H_1 : F \notin \mathcal{F}_0$.

Let us begin with the case $H_0 : F = F_0$ where we know that F is discrete with k jumps in the points $a_1, \dots, a_k$, which we suppose to be known. That is, we know that $X_1 \sim F$ assumes with positive probability only the values $a_1, \dots, a_k$. This problem is only apparently non-parametric; it is actually parametric, in that the number of parameters that define F is $2k - 1$: k possible values for X (that we supposed to be known) and $k - 1$ heights of the jumps, given by

$$p_1 = P_F(X_1 = a_1), \dots, p_{k-1} = P_F(X_1 = a_{k-1}).$$

Indeed, since $\sum_{i=1}^k p_i = 1$, the last jump height $p_k = 1 - \sum_{i=1}^{k-1} p_i$ is uniquely determined once the first $k - 1$ are known.

Let

$$p_{01} := P_{F_0}(X_1 = a_1), \ \dots, \ p_{0k} := P_{F_0}(X_1 = a_k)$$

be the values for $p_1, \dots, p_k$ specified by the null hypothesis and let us consider the problem

$$H_0 : p_i = p_{0i} \ \ \forall i = 1, \dots, k \quad \text{versus} \quad H_1 : p_i \neq p_{0i} \ \text{ for some } i.$$

Given the random sample $X_1, \dots, X_n$ extracted from F we calculate the *sample absolute frequency* of any of the a_i's, that is we calculate the number of the observations equal to a_i:

$$N_i = \#\{j : X_j = a_i\} \qquad i = 1, \dots, k$$

and the *sample relative frequency*

$$\hat{p}_{ni} = \frac{N_i}{n} \qquad i = 1, \dots, k\,,$$

which is nothing but the discrete density of the empirical c.d.f. $\hat{F}_n$.

At this point we measure the deviation of $\hat{F}_n$ from F_0 by using the statistics

$$Q_n := \sum_{i=1}^{k} \frac{(N_i - np_{0i})^2}{np_{0i}} \tag{5.4}$$

known as *Pearson statistics.* If H_0 is true, then $N_i \sim \mathcal{B}(n, p_{0i})$ and, so, $\mathbb{E}_0[N_i] = np_{0i}$ for any $i = 1, \dots, k$. Hence, under the null hypothesis H_0, we expect "small" values of Q_n. On the basis of this remark, we can adopt the following decisional rule:

if Q_n in (5.4) is large, reject H_0; if Q_n is small, accept H_0.

For large samples, we are able to approximately determine the critical values of the test statistics using the following asymptotic result:

Proposition 5.2. *Let $X_1, X_2, \dots$ be i.i.d. random variables $\sim F_0$. Then, as $n \to \infty$, the c.d.f. of $\sum_{i=1}^{k} \frac{(N_i - np_{0i})^2}{np_{0i}}$ converges to the chi-square c.d.f. with $k-1$ degrees of freedom $\chi^2(k-1)$.*

Finally, if n is large, a test of significance level approximately α is

reject H_0 if $Q_n > \chi^2_{1-\alpha}(k-1)$ and accept H_0 if, on the contrary, $Q_n \le \chi^2_{1-\alpha}(k-1)$.

Remark 5.3. A "rule of thumb" to establish if n is large enough is the following: $np_{0i} > 5$ for any $i = 1, \dots, k$.

Remark 5.4. The calculus of the value taken by the Pearson statistics is simplified by the following equalities:

$$\begin{aligned}\sum_{i=1}^{k} \frac{(N_i - np_{0i})^2}{np_{0i}} &= \sum_{i=1}^{k} \frac{N_i^2}{np_{0i}} + \sum_{i=1}^{k} \frac{(np_{0i})^2}{np_{0i}} - 2\sum_{i=1}^{k} \frac{N_i np_{0i}}{np_{0i}} = \\ &= \sum_{i=1}^{k} \frac{N_i^2}{np_{0i}} + n\sum_{i=1}^{k} p_{0i} - 2\sum_{i=1}^{k} N_i = \sum_{i=1}^{k} \frac{N_i^2}{np_{0i}} - n,\end{aligned}$$

where the last equality follows from $\sum_{i=1}^{k} p_{0i} = 1$ and $\sum_{i=1}^{k} N_i = n$.

Remark 5.5. We already observed that the Pearson statistics measures the deviation of the empirical c.d.f. $\hat{F}_n$ from the theoretical c.d.f. F_0 in terms of the deviations between the sample relative frequencies $(\hat{p}_{n1}, \dots, \hat{p}_{nk})$ and the theoretical probabilities $p_{01}, \dots, p_{0k}$. On the other hand probabilities and relative frequencies can also be defined when the data are not of *ordinal* type, but are *categorical*, that is when each observation can be classified as belonging to a certain category and there is no possible order among the various categories. For example, suppose we are conducting a sociological research about the distributions of the religions in Italy. We can use the χ^2 test considering $k-1$ categories for the $k-1$ different religions professed in Italy (the category number k is reserved to atheists) and interpreting p_{0i} as the (presumed) probability that an Italian individual chosen at random professes the religion labelled by i.

Example 5.6. The marine biologists classify the blue crabs as *young*, *adults* and *old*, depending on their dimensions. In a healthy population the proportion is usually 50% young, 30% adults, 20% old. A deviation from these proportions denotes disequilibrium in the ecosystem. In a small bay, 58 young crabs, 33 adult crabs and 39 old crabs are observed. Can the population of the crabs in the bay be considered healthy?

Let us verify the hypothesis

$$H_0 : p_1 = 0.5,\ p_2 = 0.3,\ p_3 = 0.2\,.$$

Our data are $N_1 = 58$, $N_2 = 33$, $N_3 = 39$ and $n = 58 + 33 + 39 = 130$. The value of the test statistics is

$$Q_{130} = \frac{\frac{58^2}{0.5} + \frac{33^2}{0.3} + \frac{39^2}{0.2}}{130} - 130 \simeq 8.177.$$

The p-value of the test is $1 - F_{\chi^2_{3-1}}(8.177) = 1 - F_{\mathcal{E}(2)}(8.177) = \mathrm{e}^{-8.177/2} \simeq 0.017 =$ 1.7%. Then, for example, we accept the hypothesis that the population is healthy at level 1%, but we reject the hypothesis at level 5%.[3]

5.3.3 General χ^2 test

The χ^2 goodness-of-fit test can be used to verify the goodness of fit to a discrete model (also in the case that the set of modalities of the random variable is countable) or to a continuous model; moreover it is suitable both in the case that the model is completely specified and the case where it is specified except for m unknown parameters.

In order to implement the χ^2 test, we group the data in k classes and we compare the observed frequencies of these classes with the expected values corresponding to the model specified in H_0.

χ^2 test when H_0 is simple. We start with the case $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ with F_0 completely known. Let $X_1, \dots, X_n$ be a random sample and let $A_1, \dots, A_k$ be k disjoint intervals of $\mathbb{R}$. For any $i = 1, \dots, k$ we calculate:

a) the number N_i of observations which fall in A_i, i.e. $N_i = \#\{j : X_j \in A_i\}$,

b) the theoretical probability (under H_0) that X falls in A_i, i.e. $p_{0i} = P_{F_0}(X \in A_i)$,

c) the deviation of $\hat{F}_n$ from F_0 in terms of the deviation of N_i from np_{0i} as given by the Pearson statistics (5.4).

Observe that, if H_0 is true, the mean number of observations in A_i is np_{0i}. Assuming that the sample size is large, with significance level α,

[3] Implementing the chi-square test in the R software: `chisq.test(c(58,33,39), p=c(0.5,0.3,0.2))`

```
Chi-squared test for given probabilities
data:  c(58, 33, 39)
X-squared = 8.1769, df = 2, p-value = 0.01677
```

we reject H_0 *if* $Q_n > \chi^2_{1-\alpha}(k-1)$ *and accept* H_0 *if* $Q_n \leq \chi^2_{1-\alpha}(k-1)$.

The threshold $\chi^2_{k-1}(1-\alpha)$ in the critical region of the χ^2 test is justified from the fact that the asymptotic distribution of the Pearson statistics $Q_n = \sum_{i=1}^k (N_i - np_{0i})^2/(np_{0i})$ is $\chi^2(k-1)$, as before.

Example 5.7. In order to test the goodness of a random numbers generator program, 250 numbers from the uniform distribution on $[0,1]$ are generated, obtaining:

value of X	$[0,0.2)$	$[0.2,0.4)$	$[0.4,0.6)$	$[0.6,0.8)$	$[0.8,1]$
frequency	45	53	59	43	50

What can you infer about the goodness of the program?

$A_i =$	$[0,0.2)$	$[0.2,0.4)$	$[0.4,0.6)$	$[0.6,0.8)$	$[0.8,1]$
$N_i =$	45	53	59	43	50
under $H_0 : F = \mathcal{U}(0,1)$, $n \times p_{0i} =$	$250 \cdot 0.2 = 50$	50	50	50	50

The value of the Pearson statistics is:

$$Q_{250} = \frac{45^2 + 53^2 + 59^2 + 43^2 + 50^2}{50} - 250 = 3.28\,.$$

The data have been grouped in 5 classes; hence, asymptotically, $Q_{250} \sim \chi_4$ and the p-value is $\simeq 1 - F_{\chi_4}(3.28) \simeq 1 - 0.4879 = 0.5121$: since the p-value is quite large, we can be rather confident about the goodness of the random generator program; in fact we accept H_0 at all usual significance levels.[4]

Remark 5.8. One of the most severe implementation problems of the χ^2 test for continuous or countably discrete data is the choice of the number k of disjoint intervals $A_1, \ldots, A_k$ and of their location on the real line. In the last 80 years (starting with Fisher works in the twenties [6,7], moving to Mann and Wald in 1942 [8]) many rules have been elaborated for choosing k in a way that would not reduce the power of the test. Until now, the most commonly used rule is the one due to Mann and Wald, who proposed to choose k as a function of the sample size and the significance level according to the formula $k \simeq 4(2n^2/z_{1-\alpha})^{1/5}$ and the intervals $A_1, \ldots, A_k$ as intervals with the same probability under H_0, that is such that $P_{F_0}(A_i) = 1/k$, for any $i = 1, \ldots, k$. Recent studies (at the beginning of the nineties) suggest to modify the Mann-Wald rule dividing by 4 and fixing $\alpha = 5\%$ (that is $z_{1-\alpha} \simeq 1.96$), hence some people choose k as the integer part of $n^{2/5}$ (see Del Barrio *et alii* 2000 [4]).

Remark 5.9. An efficiency problem concerning the χ^2 test arises when the data are of continuous type: the grouping of the data generates a loss of information and a consequent reduction of the power of the test (equivalently an increase of the II type error probabilities) with respect to others goodness-of-fit tests, like the Kolmogorov-Smirnov test.

[4] Using R:

```
chisq.test(c(45 , 53 , 59 ,43 ,50),p=c(0.2,0.2,0.2,0.2,0.2))
Chi-squared test for given probabilities
data:  c(45, 53, 59, 43, 50)
X-squared = 3.28, df = 4, p-value = 0.5121
```

Another problem: fixed the classes $A_1, \dots, A_k$, the Pearson statistics does not distinguish among different c.d.f. which assign the same probabilities to those classes. So, the power against that distributions in the alternative hypothesis which have the same discretized probabilities as the distribution in the null hypothesis is exactly zero.

On the other hand, the grouping of the data is also a good quality of the χ^2 test, in fact, it is possible to implement the test only knowing the grouped data, it is not necessary to know the single observations and, indeed, often only the grouped data are reported and are at disposal of the statistician.

The χ^2 test when H_0 is composed. Let be $\mathcal{F}_0 = \{F(\cdot, \theta), \theta \in \Theta \subset \mathbb{R}^m\}$ and let us consider

$$H_0 : F \in \mathcal{F}_0 \qquad \text{versus} \qquad H_1 : F \notin \mathcal{F}_0.$$

We assume that $k > m + 1$ and, for any $i = 1, \dots, k$, we calculate $N_i =$ the number of observations which fall in A_i, and $p_i(\theta_1, \dots, \theta_m)$ the probability that X_1 falls in A_i when $F \in \mathcal{F}_0$ and the value of the parameter is $(\theta_1, \dots, \theta_m)$. We then proceed as follows.

First step We estimate the m unknown parameters $\theta_1, \dots, \theta_m$ with the maximum likelihood method based on $N_1, \dots, N_k$. The random vector $(N_1, \dots, N_k)$ of the sampling frequencies has a *multinomial density* with parameters $p_1, \dots, p_k$, that is

$$P(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! \times \cdots \times n_k!} \, p_1^{n_1} \cdots p_k^{n_k},$$

$$\text{if } n_1, \dots, n_k = 0, 1, \dots, n \text{ and } \sum_{i=1}^{k} n_i = n$$

and, therefore, the logarithm of the likelihood function of the sample $(N_1, \dots, N_k)$ is

$$\log L_{\theta_1, \dots, \theta_m}(n_1, \dots, n_k) = \text{const} + \sum_{i=1}^{k} n_i \log p_i(\theta_1, \dots, \theta_m).$$

It follows that the MLE of $\theta_1, \dots, \theta_m$ (for grouped data) are the solutions of the system of equations

$$\sum_{i=1}^{k} \frac{n_i}{p_i(\theta_1, \dots, \theta_m)} \frac{\partial p_i(\theta_1, \dots, \theta_m)}{\partial \theta_j} = 0\,, \qquad j = 1, \dots, m\,,$$

which is almost always only numerically solvable. Let now $\hat{\theta}_1, \dots, \hat{\theta}_m$ be the MLE of $\theta_1, \dots, \theta_m$, based on the sample $N_1, \dots, N_k$.

Second Step. We calculate $\hat{p}_i := p_i(\hat{\theta}_1, \ldots, \hat{\theta}_m)$, for $i = 1, \ldots, k$, and

$$Q_n^* = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}. \tag{5.5}$$

The asymptotic c.d.f. of the statistics Q_n^* defined by (5.5) under H_0 (composite) is still χ^2. With regard to the degrees of freedom, Fisher (1924) proved that, compared to the aforementioned case where H_0 is simple, for any estimated parameter one degree of freedom is lost. Precisely, the following result holds:

Proposition 5.3. *If*

1. *the functions $p_i(\theta_1, \ldots, \theta_m)$ have all first and second partial derivatives and they are continuous,*
2. *$p_i(\theta_1, \ldots, \theta_m) \geq c > 0$, $\forall i = 1, \ldots, k$, $\forall \theta_1, \ldots, \theta_m$,*
3. *the matrix $\{\frac{\partial p_i(\theta_1,\ldots,\theta_m)}{\partial \theta_j}\}_{i=1,\ldots,k; j=1,\ldots,m}$ is of maximum rank m,*

then, under H_0, Q_n^* defined in (5.5) *is asymptotically $\chi^2(k-1-m)$ distributed.*

Proof. See Theorem 15.4 and Corollary 17.1 page 422 in Borovkov [1]. □

Third Step. The test is then executed as in the case of simple null hypothesis, that is *"if $Q_n^* > \chi^2_{1-\alpha}(k-1-m)$, reject $H_0 : F \in \mathcal{F}_0$ and accept $H_1 : F \notin \mathcal{F}_0$".*

Exercise 5.10 (Exercise n.50 page 480 in Mood, Graybill, Boes [9]). According to a certain genetic model, the proportion of individuals having the four existing different blood groups is:

0	A	B	AB
q^2	$p^2 + 2pq$	$r^2 + 2qr$	$2pr$

with $p, q, r \geq 0$ and $p + q + r = 1$.

Verify the goodness of the genetic model, using the following sample:

0	A	B	AB
374	436	132	58

5.4 Tests of independence and of concordance

Let X, Y be two random variables with joint c.d.f. H: $(X, Y) \sim H$ and let F, G be the marginal c.d.f. of X, Y respectively, i.e.

$$F(x) = \lim_{y \to +\infty} H(x, y)$$
$$G(y) = \lim_{x \to +\infty} H(x, y)$$

In this section we treat the problem of how to establish whether

- the characters X and Y are independent, or
- the characters X and Y are *concordant*, that is, as X increases (decreases), also Y increases (decreases), or
- the characters X and Y are *discordant*, that is, as X increases (decreases), Y decreases (increases).

The tests which can be used to solve this kind of problems are called tests of independence and tests of concordance.

We will treat mainly the problem within the non parametric setting, the only parametric case we consider is when H is jointly Gaussian, treated in 5.4.3.

5.4.1 χ^2 test of independence

See Section III.4 “The χ^2-test on statistical independence” pages 171-176 in [11].

Exercise 5.11. Do Exercise 22 pages 180-181 in [11].

Remark 5.12. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a two-dimensional random sample extracted by the c.d.f. H. The random function

$$\hat{H}_n(x, y) = \frac{\#\{i : X_i \leq x \text{ and } Y_i \leq y\}}{n} \qquad \forall x \in \mathbb{R},\ \forall y \in \mathbb{R}$$

is called the *empirical cumulative distribution function associated to the sample* (X_1, Y_1), …, (X_n, Y_n). If all we know about H is the sample, then $\hat{H}_n$ is the best we can do for estimating H. If instead the hypothesis H_0 : “X, Y *are independent*” is true, then we use as an estimator of H the product $\hat{F}_n \cdot \hat{G}_n$ of the empirical c.d.f. $\hat{F}_n$ and $\hat{G}_n$ associated to the two samples extracted from F and G, respectively.

The Pearson statistics T for the test of independence defined on page 176 in Pestman [11], can be interpreted as a (quadratic) synthesis of the distance between the c.d.f. $\hat{H}_n$ and $\hat{F}_n \cdot \hat{G}_n$. Hence, again, also this test is based on the empirical (joint and marginal) distribution functions.

5.4.2 Kendall’s test of independence and concordance

As stated in the general introduction of Section 5.4, we are studying the problem of establishing whether two characters X and Y are dependent. We will speak of positive dependence or concordance when as one variable increases (decreases), also the other variable increases (decreases). On the contrary, we will speak of negative dependence or discordance when as one variable increases (decreases), the other variable decreases (increases).

Kendall’s τ

Kendall translated the idea of positive and negative dependence as following. Let (X_1, Y_1), (X_2, Y_2) be two independent copies of the vector $(X, Y) \sim H$ and let us define

$$\pi_c := P[\text{“}(X_1 - X_2) \text{ and } (Y_1 - Y_2) \text{ have the same sign”}] = P[(X_1 - X_2)(Y_1 - Y_2) > 0] \tag{5.6}$$

and

$$\pi_d := P[\text{“}(X_1 - X_2) \text{ and } (Y_1 - Y_2) \text{ have opposite signs”}] = P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{5.7}$$

Definition 5.2. The characters X, Y are said to be *perfectly concordant* if $\pi_c = 1$, that is X, Y are perfectly concordant if $X_1 < X_2$ implies essentially $Y_1 < Y_2$ and $X_1 > X_2$ implies essentially $Y_1 > Y_2$.
The characters X, Y are said to be *perfectly discordant* if $\pi_d = 1$, that is X, Y are perfectly discordant if $X_1 < X_2$ implies essentially $Y_1 > Y_2$ and $X_1 > X_2$ implies essentially $Y_1 < Y_2$.

If F and G are continuous, then $\pi_d = 1 - \pi_c$.
Indeed, if F and G are continuous, then $P[X_1 - X_2 = 0] = P[Y_1 - Y_2 = 0] = 0$, hence

$$\begin{aligned} P[(X_1 - X_2)(Y_1 - Y_2) = 0] &= P[\{X_1 - X_2 = 0\} \cup \{Y_1 - Y_2 = 0\}] \\ &\leq P[X_1 - X_2 = 0] + P[Y_1 - Y_2 = 0] = 0 \end{aligned}$$

and

$$1 = P[(X_1 - X_2)(Y_1 - Y_2) > 0] + P[(X_1 - X_2)(Y_1 - Y_2) < 0] = \pi_c + \pi_d.$$

Definition 5.3. The measure of association between the random variables X, Y defined by

$$\tau := \pi_c - \pi_d \equiv 2\pi_c - 1$$

is known as *Kendall's τ*.

Proposition 5.4. *The following properties of Kendall's τ hold true:*

1. *For any $(X, Y) \sim H$, $-1 \leq \tau \leq 1$;*
2. *X, Y are perfectly concordant if and only if $\tau = 1$;*
3. *X, Y are perfectly discordant if and only if $\tau = -1$;*
4. *if X, Y are independent, then $\tau = 0$.*

Proof. We only prove 4. If X, Y are independent, then

$$\begin{aligned} \pi_c &= P_H((X_1 - X_2)(Y_1 - Y_2) > 0) \\ &= P_H(X_1 - X_2 > 0, Y_1 - Y_2 > 0) + P_H(X_1 - X_2 < 0, Y_1 - Y_2 < 0) \\ &= P_H(X_1 - X_2 > 0) P_H(Y_1 - Y_2 > 0) + P_H(X_1 - X_2 < 0) P_H(Y_1 - Y_2 < 0) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2} \end{aligned}$$

It can be analogously proved that $\pi_d = 1/2$, hence $\tau = 0$. □

Remark 5.13. Note that, the condition $\tau = 0$ does not imply that X, Y are independent, unless H is Gaussian. In Section 5.4.3, we will come back to this point. If $\tau = 0$ we just say that there is no association between X, Y.

Remark 5.14. It can be proved[5] that, if F is the marginal c.d.f. of X and G is the marginal c.d.f. of Y, then

1. X, Y are perfectly concordant if and only if $(X, Y) \sim H^+(x, y) = \min(F(x), G(y))$;
2. X, Y are perfectly discordant if and only if $(X, Y) \sim H^-(x, y) = \max(F(x) + G(y) - 1, 0)$;
3. The marginal c.d.f. of H^+ and H^- are F e G;
4. Finally, the class of all c.d.f. H such that $\tau = 0$ has been completely characterized in terms of the marginal c.d.f.

Kendall's independence test

Using the Kendall's τ we can face the following kinds of problems:

$$H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau > 0 \tag{5.8}$$

to verify whether there is concordance;

$$H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau < 0 \tag{5.9}$$

to verify whether there is discordance;

$$H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau \neq 0 \tag{5.10}$$

to verify that there is no association between X and Y. A test for problem (5.10) (where H_1 is bilateral) can also be red as a test of independence, in that H_1 is compatible only with a situation of non independence.

On the basis of the preceding observations, we now construct the Kendall's test of independence and concordance.

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a two-dimensional random sample extracted from an unknown joint distribution H. Let us count how many concordant and discordant data couples $(X_i, Y_i), (X_j, Y_j)$ $(i < j,\ i, j = 1, \dots, n)$ are found in the sample (over all possible $n(n-1)/2$ couples), in the following way:

First step. Order the couples of the sample on the basis of the character X. Formally: if $X_{(1)}, \dots, X_{(n)}$ are the order statistics of $X_1, \dots, X_n$ and $Y_{[k]}$ is the value of the variable Y coupled to (*concomitant to*) $X_{(k)}$, for $k = 1, \dots, n$, then the sample is reordered in the following way: $(X_{(1)}, Y_{[1]}), \dots, (X_{(n)}, Y_{[n]})$.

[5] If interested, see [2]

Example 5.15 (Example 2 in Pestman [11] page 245). If we have the sample of size 5 given by:

$$(3.7, 5.4), (2.1, 3.6), (4.2, 1.1), (3.2, 1.9), (2.3, 4.8),$$

the reordered sample is:

$$(2.1, 3.6), (2.3, 4.8), (3.2, 1.9), (3.7, 5.4), (4.2, 1.1).$$

Second step. Count how many times (over the $n(n-1)/2$ couples of indexes (i, j), with $i < j$) there is concordance:

$$C = \#\{(i, j) \text{ such that } i, j = 1, \dots, n \text{ with } i < j \text{ and } Y_{[i]} < Y_{[j]}\}$$

and how many times there is discordance:

$$D = \#\{(i, j) \text{ such that } i, j = 1, \dots, n \text{ with } i < j \text{ and } Y_{[i]} > Y_{[j]}\}.$$

Third step. Compute the *Kendall's sample concordance coefficient*, defined by

$$R_K := \frac{2(C-D)}{n(n-1)}$$

One has easily:

1. $C \sim \mathcal{B}(n(n-1)/2, \pi_c)$, $D \sim \mathcal{B}(n(n-1)/2, \pi_d)$
2. R_K is an unbiased estimator of $\tau = \pi_c - \pi_d$;
3. if there is total concordance, then $C = n(n-1)/2$, $D = 0$ and $R_K = 1$. If there is total discordance, then $C = 0$, $D = n(n-1)/2$ and $R_K = -1$;
4. if $\tau = 0$, $C - D$ (and hence R_K) is a symmetric (with respect to 0) random variable, because in this case any couple has the same probability of being concordant or discordant.

If $\tau = 0$ the probability that the statistics R_K exhibit values "closed" to -1 or to 1 is small, in that there is no association between X and Y. In other words, we can decide about the association between X and Y by estimating whether the values of the statistics $C - D$ are closed to one of the extremes $-n(n-1)/2$ and $n(n-1)/2$ or are far from both the extreme values.

Following this intuition, we construct the critical regions of size α:

$$\{C - D > q_{C-D}(1-\alpha)\} \text{ is a critical region for } H_0 : \tau = 0 \text{ versus } H_1 : \tau > 0$$
$$\{C - D < -q_{C-D}(1-\alpha)\} \text{ is a critical region for } H_0 : \tau = 0 \text{ versus } H_1 : \tau < 0$$
$$\{|C - D| > q_{C-D}(1-\frac{\alpha}{2})\} \text{ is a critical region for } H_0 : \tau = 0 \text{ versus } H_1 : \tau \neq 0.$$

In the above lines, $q_{C-D}(a)$ denotes the quantile of order a of the c.d.f. of $C-D$ under H_0 and $q_{C-D}(a) = -q_{C-D}(1-a)$, $\forall a \in (0,1)$, since, under the null hypothesis $\tau = 0$, $C-D$ is symmetric with respect to 0.

We need therefore to investigate the distribution of $C-D$ when $\tau = 0$, that is under H_0: for small values of n the quantiles of $C-D$ are tabulated, for example in Conover [3] pages 543–544. For big values of n, the following result holds: if $\tau = 0$, then

$$\lim_{n\to\infty} P\left(3\sqrt{\frac{n(n-1)}{2(2n+5)}}R_K \le z\right) = \Phi(z)\,, \qquad \forall z \in \mathbb{R}.$$

Hence, for large samples it is possible to approximate the quantile q_{C-D} using the quantile z_a of the standard Gaussian , precisely:

$$q_{C-D}(a) \simeq z_a\sqrt{\frac{n(n-1)(2n+5)}{18}}$$

Example 5.16 (Continuation of Example 2 in Pestman [11] page 245). Given the sample of size 5:

$(3.7, 5.4), (2.1, 3.6), (4.2, 1.1), (3.2, 1.9), (2.3, 4.8)$,
the ordered sample is:

$(2.1, 3.6), (2.3, 4.8), (3.2, 1.9), (3.7, 5.4), (4.2, 1.1)$
and $R_K = -2/10$ since

	2	3	4	5
$-\text{signum}(Y_{[1]} - Y_{[j]}) =$	$+1$	-1	$+1$	-1
$-\text{signum}(Y_{[2]} - Y_{[j]}) =$		-1	$+1$	-1
$-\text{signum}(Y_{[3]} - Y_{[j]}) =$			$+1$	-1
$-\text{signum}(Y_{[4]} - Y_{[j]}) =$				-1

Let us consider $H_1 : \tau \neq 0$. With $n = 5$, the p-value is $P(|C-D| \ge 2) \simeq 0.817$ and therefore we accept H_0[6].

Exercise 5.17. Do Exercise n. 14 page 255 in Pestman [11], substituting to questions *i*), *ii*) the following ones:

i) Compute the outcome of the sample Kendall's coefficient of concordance (R_K).

ii) Test at a level significance of $\alpha = 0.10$ the null hypothesis $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$.

[6] using the software *R*:

```
x<-c(3.7,2.1,4.2,3.2,2.3);
y<-c(5.4,3.6,1.1,1.9,4.8)
cor.test(x, y, alternative = c("two.sided"), method =c("kendall"), exact = NULL)
Kendall's rank correlation tau
data:  x and y T = 4, p-value = 0.8167 alternative hypothesis:  true tau is not equal
to 0
sample estimates:  tau -0.2
```

5.4.3 Test of independence and concordance for Gaussian data

Proposition 5.5. *If $(X,Y) \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$, then X,Y are independent if and only if the linear correlation coefficient*

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

is zero. Moreover, if $(X,Y) \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$, then $\tau(X,Y)=0$ if and only if $\rho(X,Y)=0$.

Hence, in the case of joint Gaussian data, the test of independence can be translated in a test $H_0 : \rho(X,Y) = 0$ versus $H_1 : \rho(X,Y) \neq 0$. The test of concordance is translated in a test on the positive linear correlation: $H_0 : \rho(X,Y) = 0$ versus $H_1 : \rho(X,Y) > 0$, while the test of discordance is simply a test on the negative linear correlation: $H_0 : \rho(X,Y) = 0$ versus $H_1 : \rho(X,Y) < 0$.

Let us see how we can make inference about the linear correlation coefficient ρ starting by a bivariate random sample $(X_1,Y_1),\dots,(X_n,Y_n)$ extracted from a Gaussian population with means μ_X,μ_Y, variances σ_X^2,σ_Y^2 and linear correlation coefficient ρ, **all unknown**. An estimator of ρ is given by the *sample or empirical correlation coefficient*:

$$R = \frac{\sum_{j=1}^n (X_j-\bar{X})(Y_j-\bar{Y})}{\sqrt{\sum_{j=1}^n (X_j-\bar{X})^2 \sum_{j=1}^n (Y_j-\bar{Y})^2}}$$

so called because it is nothing but the correlation coefficient of the bivariate empirical distribution $H_n(x,y)$ associated to the sample $(X_1,Y_1),\dots,(X_n,Y_n)$. Hence $-1 \le R \le 1$. With regard to the c.d.f. of the estimator R the following result holds:

Theorem 5.6. *Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be i.i.d. $\sim \mathcal{N}$ and $\rho = 0$. Then*

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \sim t(n-2), \quad n \ge 3.$$

Proof. See Section 7.6 pages 325-332 in [13]. □

Since $\frac{x}{\sqrt{1-x^2}}\sqrt{n-2}$ is an increasing function of x in the interval $(-1,1)$, then a test for the null hypothesis $H_0 : \rho = 0$ can be constructed on the basis of the test statistics $\dfrac{R}{\sqrt{1-R^2}}\sqrt{n-2}$, for any alternative $H_1 : \rho \neq 0$ or $H_1 : \rho > 0$ or $H_1 : \rho < 0$. But, depending on the alternative hypothesis, the quantile to be compared with the statistics value changes accordingly. Precisely:

$$\left\{\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} > t_{1-\alpha}(n-2)\right\} \text{ is a critical region for } H_0 : \rho = 0 \text{ versus } H_1 : \rho > 0$$

$$\left\{\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} < -t_{1-\alpha}(n-2)\right\} \text{ is a critical region for } H_0 : \rho = 0 \text{ versus } H_1 : \rho < 0$$

$$\left\{\left|\frac{R\sqrt{n-2}}{\sqrt{1-R^2}}\right| > t_{1-\frac{\alpha}{2}}(n-2)\right\} \text{ is a critical region for } H_0 : \rho = 0 \text{ versus } H_1 : \rho \neq 0$$

Exercise 5.18. Let $(X_1, Y_1), \dots, (X_{10}, Y_{10})$ be a bivariate random sample extracted by a Gaussian (bivariate) population with parameters all unknown. The following values of some sample moments are found:

$$\bar{x} = -0.231,\ \bar{y} = 0.988,\ \sum_{j=1}^{10} x_j^2 = 6.54,\ \sum_{j=1}^{10} y_j^2 = 24.024,\ \sum_{j=1}^{10} x_j y_j = -5.755.$$

Propose a procedure to verify whether X and Y are independent.

5.5 Tests of randomness

A fundamental hypothesis in all the inferential procedures presented in these Lectures Notes is that the sample is a random sample, that is the hypothesis of randomness of the sample. If the n observations $X_1, \dots, X_n$ are actually i.i.d., the order of the observations does not matter; in other words, an increasing or decreasing trend in the observations is not compatible with the hypothesis of randomness of the sample.

5.5.1 Kendall test of randomness

We now show how the Kendall's concordance test can be applied to verify the null hypothesis of randomness of the sample: $H_0 : X_1, \dots, X_n$ i.i.d. $\sim F$ versus an alternative hypothesis of increasing trend, or decreasing trend, or simply the bilateral alternative: "$H_1 : X_1, \dots, X_n$ *are not i.i.d.* ".

For any $i = 1, \dots, n-1$, let us set $C_i =$ "number of observations subsequent to X_i and greater than X_i", and $D_i =$ "number of observations subsequent to X_i and smaller than X_i". Moreover, let T be the difference between the sum of all C_i's and the sum of all D_i's that is: $C = \sum_{i=1}^{n-1} C_i$, $D = \sum_{i=1}^{n-1} D_i$ and $T = C - D$.

The minimum value of T is $-n(n-1)/2$ and T takes this value in the case $X_1 > X_2 > \cdots > X_n$. On the other hand T takes the value $n(n-1)/2$ if and only if $X_1 < X_2 < \cdots < X_n$. Therefore $T \approx -n(n-1)/2$ signals a decreasing trend while $T \approx n(n-1)/2$ is only compatible with an increasing trend. Finally, if the sample is a random sample we expect approximately an equal number of concordant and discordant couples, therefore $T \approx 0$. But, T is simply the difference between the number of concordant couples and the number of discordant couples of the artificial bivariate sample $(1, X_1), \dots, (n, X_n)$. We can then apply the Kendall's concordance test. The critical regions of level α are:

1. $\{C - D > q_{C-D}(1-\alpha)\}$ for H_0: "$X_1, \dots, X_n$ is a random sample" versus H_1 : *"data exhibit an increasing trend"*

2. $\{C - D < -q_{C-D}(1-\alpha)\}$ for H_0 : "$X_1, \dots, X_n$ is a random sample" versus H_1 : *"data exhibit a decreasing trend"*

3. $\{|C - D| > q_{C-D}(1-\alpha/2)\}$ for H_0: "$X_1, \dots, X_n$ is a random sample" versus H_1 : *"data are not independent"*.

Example 5.19. Using the software R, 10 random numbers from the uniform distribution $\mathcal{U}(0,1)$ are generated, and, in arriving order, the data are the following:

0.5923 0.6944 0.6956 0.6443 0.6114 0.5073 0.0993 0.1070 0.6701 0.3607.

Can these numbers be considered as a realization of a random sample extracted from a $\mathcal{U}(0,1)$ distribution?

i	$C-D$
0.5923	+5 - 4
0.6944	+1 - 7
0.6956	+0 - 7
0.6443	+1 - 5
0.6114	+1 - 4
0.5073	+1 - 3
0.0993	+3 - 0
0.1070	+2 - 0
0.6701	+0 - 1
0.3607	$\boxed{-17 = T}$

If H_1 is bilateral, the critical region to consider is $\mathcal{G}_3$ and the p-value of the test is $2(1 - P(T \leq 17))$. Consulting the tables, for $n = 10$, the closest values to 17 are 15 and 19 with $P(T \leq 15) = 0.90$ and $P(T \leq 19) = 0.95$. Hence the p-value is between 10% and 20%. Linearly interpolating $(15, 0.90)$ and $(19, 0.95)$ we obtain p-value$\simeq$ 15% (R gives 0.1557). Hence the hypothesis of randomness is rejected only at level $\geq$ 15%. We can conclude that the random generator program is a good program.

5.5.2 Testing randomness of sequences of bits

Two batteries of statistical tests:

`http://csrc.nist.gov/rng/`
`http://random.com.hr/products/random/Diehard.html`

An example of application:

`http://www.idquantique.com/products/files/quantis-test.pdf`

5.6 Homogeneity tests

We now treat the problem to verify whether two variables X, Y are homogeneous, that is whether they are regulated by the same model. In other words, we want to verify whether X and Y have the same c.d.f. Let F be the c.d.f. of X and G the c.d.f. of Y, and let us construct a *test of homogeneity* for the null hypothesis

$$H_0 : F(x) = G(x) \qquad \forall x \in \mathbb{R}$$

versus the alternative

$$H_1 : F(x) \neq G(x) \qquad \text{for some } x \in \mathbb{R} \tag{5.11}$$

or

$$H_1 : F(x) \leq G(x) \ \ \forall x \in \mathbb{R}, \ \ \text{and } F(x) < G(x) \text{ for some } x \tag{5.12}$$

or

$$H_1 : F(x) \geq G(x) \ \ \forall x \in \mathbb{R}, \ \ \text{and } F(x) > G(x) \text{ for some } x. \tag{5.13}$$

The unilateral alternative hypotheses have the following meaning: if (5.12) is true, then X tends to be greater than Y and we say that X *stochastically dominates* Y. If instead (5.13) is true, then Y tends to be greater than X, that is Y *stochastically dominates* X.

The data can be of two types:

1. two independent random samples: $X_1, \ldots, X_m$ i.i.d. $\sim F$ and $Y_1, \ldots, Y_n$ i.i.d. $\sim G$,

2. a random sample of couples data $(X_1, Y_1), \ldots, (X_n, Y_n)$ extracted from a joint distribution H such that F is the marginal c.d.f. of X and G the marginal c.d.f. of Y.

The χ^2-test and the Kolmogorov-Smirnov test, already seen in the previous sections as goodness of fit tests, can be modified to verify $H_0 : F(x) = G(x)\ \forall x \in \mathbb{R}$. In the following sections we instead describe the *Wilcoxon-Mann-Whitney homogeneity test for independent samples* and the *Wilcoxon test of signs for coupled data.*

To avoid technicalities, we assume that F, G (and eventually also the joint distribution H) are continuous: from this hypothesis it follows that *essentially* there are no repetitions in the samples, in that $P(X_{i_1} = X_{i_2}) = P(Y_{j_1} = Y_{j_2}) = P(X_i = Y_j) = 0$, $\forall i \neq j,\ i_1 \neq i_2$ and $j_1 \neq j_2$.

5.6.1 Wilcoxon-Mann-Whitney homogeneity test for independent samples

See [11] Section VI.2 "Wilcoxon's rank-sum test",

Suppose we have two independent random samples $X_1, \ldots, X_m$ i.i.d. $\sim F$ and $Y_1, \ldots, Y_n$ i.i.d. $\sim G$. Let us collect the two samples in a single one of size $m + n$ and let us order it in increasing order. Let us record the *rank* (or *degree*) R_i of each observation X_i, that is the position of the observation in the ordered sample and let us calculate the sum of the ranks $T_X = \sum_{i=1}^{m} R_i$.

Example 5.20. If $(x_1, \dots, x_4) = (23, 10, 21, 5)$ and $(y_1, \dots, y_5) = (3, 8, 20, 25, 12)$, then

$$\begin{array}{rccccccccc}
\text{ordered global sample} = & 3_y & 5_x & 8_y & 10_x & 12_y & 20_y & 21_x & 23_x & 25_y \\
\text{ranking} = & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
X \text{ ranks} = & & R_1 & & R_2 & & & R_3 & R_4 & \\
T_X = & & 2+ & & 4+ & & & 7+ & 8 = & \boxed{21}
\end{array}$$

If all the x_i's are smaller than all the y_j's, T_X takes value $m(m+1)/2$; if, on the contrary, all the x_i's are greater than all the y_j's, then $T_X = m(2n+m+1)/2$.

If $F(x) < G(x)$, we expect that many x_i's are greater than the y_j's and therefore that T_X is "large". Viceversa, if $F(x) > G(x)$, we expect that many y_j's are greater than the x_i's and therefore that T_X is "small". Finally, if $F = G$, the x_i's and y_j's are randomly mixed and therefore we expect that with high probability T_X is far from the extreme values. In order to construct the critical region of the test we need to know the c.d.f. of T_X when $F = G$.

Let U be the total number of terms of type X greater than terms of type Y in the reunited sample:

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}_{(Y_j, \infty)}(X_i). \tag{5.14}$$

To obtain U, we first identify the values of the m order statistics $x_{(1)} < \cdots < x_{(m)}$; then we count the number a_1 of y_j's smaller than $x_{(1)}$, the number a_2 of y_j's smaller than $x_{(2)}, \dots$, and the number a_m of y_j's smaller than $x_{(m)}$. We then sum the a_i's and set $U = a_1 + a_2 + \cdots + a_m$. Note than $a_1 =$ "$\#y_j \leq x_{(1)}$"$= R_1 - 1$, $a_2 =$ "$\#y_j \leq x_{(2)}$"$= R_2 - 2, \dots, a_m =$ "$\#y_j \leq x_{(m)}$"$= R_m - m$, therefore

$$U = \sum_{i=1}^{m} (R_i - i) = T_X - \frac{m(m+1)}{2}. \tag{5.15}$$

The statistics U is called *Mann-Whitney statistics*, while the sum of the ranks T_X is called *Wilcoxon statistics.*

We now calculate the mean and the variance of T_X, when $F = G$. If $F = G$, the random variables $X_1, \dots, X_m, Y_1, \dots, Y_n$ are i.i.d. with common c.d.f. F and

$$\mathbb{E}(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{E}(\mathbf{1}_{(Y_j, \infty)}(X_i)) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(X_i > Y_j) = mnP(X_1 > Y_1) = \frac{mn}{2}$$

in that $P(X_1 = Y_1) = 0$ and $P(X_1 > Y_1) = P(X_1 < Y_1) = 1/2$, since X_1, Y_1 are independent and have the same c.d.f. F. It follows that, if $F = G$, $\mathbb{E}(T_X) = \mathbb{E}(U) + m(m+1)/2 = m(m+n+1)/2$. Analogous calculations lead to the following value for the variance:

$$\operatorname{Var}(T_X) = \operatorname{Var}(U) = \frac{mn(m+n+1)}{12}.$$

If $F = G$ and $m, n \leq 20$, the quantiles w_a of the c.d.f. of T_X[7] are tabulated (see, for example, Appendix A7 pages 536-538 in [3]). Moreover, for symmetry reasons, the following relation holds between w_a and w_{1-a}:

$$w_a = m(m + n + 1) - w_{1-a}.$$

Finally, for large samples and $F = G$, T_X is asymptotically Gaussian, in the sense that

$$\lim_{n \to \infty} P\left(\frac{T_X - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \leq z \right) = \Phi(z) \qquad \forall z \in \mathbb{R}.$$

It follows that an approximate value of the quantile of order a of T_X is

$$w_a \simeq \frac{m(m + n + 1)}{2} + z_a \sqrt{\frac{mn(m + n + 1)}{12}}$$

In the light of what we have just shown, it seems reasonable to adopt the following decisional criteria for a test of significance level α:

Reject $H_0 : F(x) = G(x)\ \forall x$ *and accept* H_1 : "$F(x) \geq G(x) \forall x \in \mathbb{R}$ *and* $F(x) > G(x)$ *for some* x" *if* $T_X < w_\alpha$

Reject $H_0 : F(x) = G(x) \forall x$ *and accept* H_1 : "$F(x) \leq G(x) \forall x \in \mathbb{R}$ *and* $F(x) < G(x)$ *for some* x" *if* $T_X > w_{1-\alpha}$

Reject $H_0 : F(x) = G(x) \forall x$ *and accept* $H_1 : F(x) \neq G(x)$ *if* $T_X \notin [w_{\alpha/2}, w_{1-\alpha/2}]$.

The described test is known as *Wilcoxon-Mann-Whitney rank-sum test.*

Remark 5.21. If, instead of considering the sum of the ranks of the terms of type X, we consider the sum of the ranks of the y_j's, T_Y, we end up with the the same test, in that one has $T_X + T_Y = (m + n)(m + n + 1)/2$.

Example 5.22 (continuation of Example 5.20). Let us verify the hypothesis $H_0 : F = G$ versus the alternative $H_1 : F \neq G$, at level $\alpha = 10\%$, using the data in Example 5.20. We have $T_X = 21$ and, with $m = 4$ and $n = 5$, we find on the tables that $w_{0.10/2} = w_{0.05} = 13$ and $w_{1-0.10/2} = w_{0.95} = m(m + n + 1) - w_{0.05} = 40 - 13 = 27$. Since $13 < 21 < 27$ we accept H_0.

Remark 5.23. The rank-sum test can be red as a non-parametric version of the t-test for the comparison of the means of two independent Gaussian populations. In fact, if the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are both Gaussian and have the same variance, then $F = G$ if and only if the means are equal and the homogeneity test is substituted by the appropriate t-test, depending on the alternative hypothesis (5.11) or (5.12) or (5.13). If $F = N(\mu_X, \sigma^2)$ and $G = N(\mu_Y, \sigma^2)$, it is easy to prove that $F > G$ is equivalent to $\mu_X < \mu_Y$, while $F < G$ if and only if $\mu_X > \mu_Y$. Anyhow, if the samples are Gaussian, the Wilcoxon-Mann-Whitney test is almost as powerful as the t-test.

[7] The tables were calculated first by Mann and Whitney in 1947 for $m, n \leq 8$; while the first version of the Wilcoxon test for the case $m = n$ is dated 1945.

Remark 5.24 (Non-parametric test on the variance). A variant of the rank-sum test due to Siegel and Tukey (1960) (see for instance [3]) can be used to compare the variances of two different and independent populations with equal means, in a non-parametric context. Siegel and Tukey modify the test of the sum of the ranks in the following way: once the observations have been reunited and ordered in increasing order, the smaller observation is labelled as 1, the greatest as 2, then the last but one is labelled as 3, the second as 4, the third as 5, the last but two as 6, and so on:

$$1,\ 4,\ 5,\ 8,\ 9,\ ,\dots,(m+n-1),(m+n),\dots\ 7,\ 6,\ 3,\ 2.$$

At this point the sum of the labels of the elements of type X is calculated. If this sum is small, this means that X is more widespread than Y, and therefore $H_0 : \mathrm{Var}(X) = \mathrm{Var}(Y)$ is rejected in favour of $H_1 : \mathrm{Var}(X) > \mathrm{Var}(Y)$. Suitable tables of quantiles are available. The Siegel-Tukey test is a sort of non-parametric version of the Fisher test for the comparison of the variances of two independent Gaussian populations.

5.6.2 Wilcoxon test of signs for coupled data

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample of coupled data extracted from a jointly continuous c.d.f. H with marginal c.d.f. F (of X) and G (of Y). Let T^+ be the number of couples where the X coordinate is greater than the Y coordinate, and let T^- be the number of couples where the Y coordinate is greater than the X coordinate. We expect $T^+ \approx n/2$ if $F = G$, T^+ "large" if $F < G$ and T^+ "small" if $F > G$. Since the couples of observations are i.i.d. , *the statistics T^+ has binomial density with parameters n and $p = P(X > Y)$*. In particular, if $H_0 : F = G$ is true, then $T^+ \sim \mathcal{B}(n, 1/2)$ in that, $P(X = Y) = 0$ for the continuity of H, and $P(X > Y) = P(Y > X) = 1/2$.

In the light of what we have just shown, it seems reasonable to adopt the following decisional criteria for a test of significance level α to verify the hypothesis $F = G$ for coupled data:

Reject $H_0 : F(x) = G(x)\,\forall x$ and accept H_1 : "$F(x) \geq G(x)\,\forall x \in \mathbb{R}$ and $F(x) > G(x)$ for some x" if $T^+ < q^+_\alpha$

Reject $H_0 : F(x) = G(x)\,\forall x$ and accept H_1 : "$F(x) \leq G(x)\,\forall x \in \mathbb{R}$ and $F(x) < G(x)$ for some x" if $T^+ > q^+_{1-\alpha}$

Reject $H_0 : F(x) = G(x)\,\forall x$ and accept $H_1 : F(x) \neq G(x)$ if $T^+ \notin [q^+_{\alpha/2}, q^+_{1-\alpha/2}]$

where q^+_a is the quantile of order a of the $\mathcal{B}(n, 1/2)$ distribution.

The test described is known as *Wilcoxon signs test.* For n large, by the Central Limit Theorem, $q_+(a) \simeq n/2 + z_a\sqrt{n}/2$

Remark 5.25. The Wilcoxon signs test can be red as a non-parametric version of the t-test to compare the means of Gaussian coupled data.

Appendix A

The Greek alphabet

Greek	letters	math	symbols	names
Α	α		α	alpha
Β	β		β	beta
Γ	γ	Γ	γ	gamma
Δ	δ	Δ	δ	delta
Ε	ε		ϵ , ε	epsilon
Ζ	ζ		ζ	zeta
Η	η		η	eta
Θ	ϑ	Θ	θ , ϑ	theta
Ι	ι		ι	iota
Κ	ϰ		κ , $\varkappa$	kappa
Λ	λ	Λ	λ	lambda
Μ	μ		μ	mu
Ν	ν		ν	nu
Ξ	ξ	Ξ	ξ	xi
Ο	ο			omicron
Π	π	Π	π , ϖ	pi
Ρ	ρ		ρ , ϱ	rho
Σ	σς	Σ	σ , ς	sigma
Τ	τ		τ	tau
Υ	υ	Υ	υ	upsilon
Φ	φ	Φ	ϕ , φ	phi
Χ	χ		χ	chi
Ψ	ψ	Ψ	ψ	psi
Ω	ω	Ω	ω	omega
F		$\digamma$		digamma

Appendix B

Statistical tables

Standard normal distribution:
Some often used values: $z_{.90} \simeq 1.282$ $z_{.95} \simeq 1.645$ $z_{.975} \simeq 1.960$ $z_{.99} \simeq 2.326$
Tables: [11] p. 523 or

```
http://web.mate.polimi.it/viste/studenti/
            pagina_docente5.php?id=7&id_insegnamento=595
http://www.york.ac.uk/depts/maths/tables/normal.pdf
```

Student t-distribution: [11] p. 524 or

```
http://www.york.ac.uk/depts/maths/tables/t.pdf
```

χ^2-distribution: [11] p. 525 or

```
http://web.mate.polimi.it/viste/studenti/
            pagina_docente5.php?id=7&id_insegnamento=595
http://www.york.ac.uk/depts/maths/tables/chisquared.pdf
```

Fisher F-distribution: [11] p. 526–529 or

```
http://www.york.ac.uk/depts/maths/tables/f.pdf
```

Mann-Whitney-Wilcoxon test statistics T_X: [3] pp. 536–538.
Binomial distribution: [11] pp. 519–522 or

```
http://www.york.ac.uk/depts/maths/tables/bintable.pdf
```

Kolmogorov-Smirnov two-sided test: [11] p. 536. For $n > 50$: $D_{.9}(n) \simeq \frac{1.22}{\sqrt{n}}$, $D_{.95}(n) \simeq \frac{1.36}{\sqrt{n}}$, $D_{.98}(n) \simeq \frac{1.52}{\sqrt{n}}$, $D_{.99}(n) \simeq \frac{1.63}{\sqrt{n}}$.
Kendal test: quantiles of the Kendal statistics $T = C - D$: [3] pp. 543–544; quantiles of the Kendal statistics R_K:

```
http://www.york.ac.uk/depts/maths/tables/kendall.pdf
```

Bibliography

[1] Borovkov, A. (1987) *Statistique mathématique*, Éditions Mir, Moscow.

[2] Cifarelli, D.M., Conti, L., Regazzini, E. (1996) On the asymptotic distribution of a general measure of monotone dependence. *Ann. Statist.* **24**, 1386–1399.

[3] Conover, W.J. (1999) *Practical Nonparametric Statistics*, Wiley, New York

[4] Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. (2000) Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. (With comments) *Test* **9**, 1–96

[5] Epifani, I., Ladelli, L., Posta, G. (2006) *Appunti per il corso di Calcolo delle Probabilità*, http://www1.mate.polimi.it/ ileepi/dispense/0506CP/

[6] Fisher, R. (1922) On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, A*, **222**, 309–368.

[7] Fisher, R.A. (1924) The conditions under which χ^2 measures the discepancy between observation and hypohesis. *J. Roy. Statist. Soc.*, **87**, 442–450.

[8] Mann, H.B. and Wald, A. (1942) On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math. Stat.*,**13**, 306–317.

[9] Mood, A.M., Graybill, F.A., Boes, D.C. (1982) Introduction to the Theory of Statistics, McGraw-Hill.

[10] Pearson, K. (1894) *Contributions to the Mathematical Theory of Evolution*, Philosophical Transactions of the Royal Society A, **185**, 71–110.

[11] Pestman, W. R. (1998) *Mathematical Statistics, An Introduction* De Gruyter.

[12] *R: A language and environment for statistical computing* R Development Core Team (2003) `http://www.R-project.org` , R Foundation for Statistical Computing Vienna, Austria

[13] Rohatgi, V.K. and Saleh, A.K. Md. E. (1999) *An Introduction to Probability and Statistics*, Wiley, New York

[14] Ross, S.M. (1987) *Introduction to Probability and Statistics for Engineers and Scientists*, Wiley, New York

[15] Ross, S.M. (2002) *A First Course in Probability*, Prentice-Hall

[16] Silvey, S.D (1975) *Statistical Inference* Chapman & Hall London