I. STATISTICAL MODELS

One very important environment in which the formalism of probability theory plays a leading role is the science of extracting information from **data**: mathematical statistics. Observations, that is measured data, are used to **infer** the values of some parameters necessary to complete a mathematical description of an experiment.

The simplest situation we can imagine is to measure n-times the same quantity (or the same set of quantities), performing all the measurements under the same experimental conditions and in such a way that the result of any measure does not affect the results of the others. The outcome of the experiment is thus a collection of **data**:

$$(\underline{x}_1 \dots \underline{x}_n) \tag{1}$$

where $\underline{x}_i \in \mathbb{R}^k$ is the result of the i-th measure.

Naturally, even if we are very careful in the preparation of the experimental setup, we cannot expect that, if we repeated the whole set of n measures, we would find the same data (1): some randomness unavoidably exists.

It is thus natural to use the language of probability theory to describe the experiment. The i-th measure can be modeled by a random variable X_i, defined on some probability space $(\Omega, \mathcal{F}, P)$ and the whole outcome of the experiment, the data $(\underline{x}_1 \dots \underline{x}_n)$, can be viewed as realizations of a collection $(X_1, \dots, X_n)$ of random variables, **independent** and **identically distributed**. The requirement of independence is suggested by the assumption that the result of any measure does not affect the results of the others while the one of identical distribution translates the idea that the measurements are performed under the same experimental conditions.

Sometimes, depending on the measurement procedure, one has an idea about the law of the random variables X_i: in could be Binomial, Poisson, Exponential, Normal, Uniform and so on. However, in general, the actual parameters characterizing the law are not known but can be **inferred** from the data $(\underline{x}_1 \dots \underline{x}_n)$.

This typical situation justifies the following definition:

Definizione 1 *A* **statistical model** *is a family:*

$$\{(\Omega, \mathcal{F}, P_{\underline{\theta}})\}_{\underline{\theta} \in \Theta}$$

of probability spaces sharing the same sample space and the same collection of events. The probability measures $P_{\underline{\theta}} : \mathcal{F} \to [0,1]$ depend on a parameter $\underline{\theta}$ taking values in a set $\Theta \subseteq \mathbb{R}^m$.

A statistical model describes the preparation procedure of the experiment; the results of the experiment, as anticipated above, are realizations of a multidimensional random variable:

$$X = (X_1 \dots X_n) \tag{2}$$

where the components $X_i : \Omega \to E \subset \mathbb{R}^k$ are **independent** and **identically distributed**. Such a random variable X is called a **sample** of **rank** n.

Within a statistical model, the law of X_i, describing the measurement procedure:

$$\mathcal{B}(\mathbb{R}^k) \ni B \rightsquigarrow \mu_{\underline{\theta}}(B) = P_{\underline{\theta}}(X_i \in B) \tag{3}$$

naturally depends on $\underline{\theta}$ and can be related to a density $p_{\underline{\theta}}(\underline{x})$, discrete of continuous.

As usual in probability theory, the actual precise definition of the triplet $(\Omega, \mathcal{F}, P_{\underline{\theta}})$ is in general omitted once the law of the sample is specified.

Some examples of models that are frequently employed are summarized in the following table, in which we indicate the range of the measurements E, the set Θ and the density $p_{\underline{\theta}}(\underline{x})$.

model	E	Θ	$p_{\underline{\theta}}$
Bernoulli	$\{0,1\}$	$(0,1)$	$\theta^{1-x}(1-\theta)^x$
gaussian	$\mathbb{R}$	$\mathbb{R} \times (0,\infty)$	$\frac{1}{\sqrt{2\pi\theta_1}}\, e^{-\frac{(x-\theta_0)^2}{2\theta_1}}$
exponential	$(0,\infty)$	$(0,\infty)$	$\theta_0\, e^{-\theta_0 x}$

All such models are examples of a wide class of models, called ***s*-parameters exponential models**, characterized by densities of the form:

$$p_{\underline{\theta}}(\underline{x}) = e^{-\underline{f}_1(\underline{\theta})\cdot\underline{f}_2(\underline{x})}\, e^{-f_3(\underline{\theta})}\, f_4(\underline{x}) \tag{4}$$

where $\underline{f}_1 : \Theta \to \mathbb{R}^s$, $\underline{f}_2 : \mathbb{R}^k \to \mathbb{R}^s$, $f_3 : \Theta \to \mathbb{R}$ ed $f_4 : \mathbb{R}^k \to [0,\infty)$.

II. ESTIMATORS

One of the main goals of mathematical statistics is to use the **data** $(\underline{x}_1 \dots \underline{x}_n)$ to **estimate** functions $\tau(\underline{\theta})$ of the parameter $\underline{\theta}$, useful to complete the probabilistic description of the experiment. For this purpose, suitable functions have to be applied to the data; keeping in mind that the data are viewed as realizations of a sample, the following definition is quite natural:

Definizione 2 *A* **statistic** $\mathcal{T}$ *is an s-dimensional random variable of the form:*

$$\Omega \ni \omega \rightsquigarrow \mathcal{T}(\omega) = t(X_1(\omega), \dots, X_n(\omega))$$

where $t : \mathbb{R}^k \times \cdots \times \mathbb{R}^k \to \mathbb{R}^s$ *is a measurable function which* **does not depend** *on the parameter* $\underline{\theta}$ *and* $(X_1 \dots X_n)$ *is a sample of rank* n*.*

The definition (2) of a statistic describes the manipulations we make to the data. When a statistic $\mathcal{T}$ is used to infer a value for a given function of the parameter $\underline{\theta}$, $\tau(\underline{\theta})$, we say that $\mathcal{T}$ is an **estimator** of $\tau(\underline{\theta})$, while $t(\underline{x}_1 \dots \underline{x}_n)$ is called **pointwise estimation** of $\tau(\underline{\theta})$.

III. THE EMPIRIC MEAN

The most natural statistic one considers when dealing with a set of data is the **mean**. Inside our formalism, we build up the estimator $\mathcal{M}$:

$$\mathcal{M} = m(X_1 \dots X_n) = \frac{\sum_{i=1}^n X_i}{n} \tag{5}$$

for the unknown quantity:

$$\mu(\underline{\theta}) = E_{\underline{\theta}}(X_i) = \int dx x p_{\underline{\theta}}(x)$$

We start our presentation of mathematical statistics from the analysis of this estimator, which will help us to introduce some basic notions.

Intuitively, given the set of data $(x_1, \dots, x_n)$, one would like to write something like $\mu(\underline{\theta}) \simeq \frac{\sum_{i=1}^n x_i}{n}$.

Let's give a precise meaning to such an operative procedure.

$\mathcal{M}$ is a random variable, with a law depending on the law of the sample; in particular the expected value is readily computed:

$$E_{\underline{\theta}}(\mathcal{M}) = \frac{\sum_{i=1}^n E_{\underline{\theta}}(X_i)}{n} = \mu(\underline{\theta})$$

and coincides with $\mu(\underline{\theta})$. So $\mathcal{M}$ has expected value equal to the quantity we wish to infer. This is an important property of the estimator, called **unbiasedness**, defined in the following:

Definizione 3 *An estimator* $\mathcal{T}$ *of a function* $\tau(\underline{\theta})$ *is called* **unbiased** *if:*

$$E_{\underline{\theta}}(\mathcal{T}) = \tau(\underline{\theta}), \quad \forall \underline{\theta} \tag{6}$$

What about the "error"? In other words, what do we expect about the spreading of the realizations of $\mathcal{M}$ around the expected value $\mu(\underline{\theta})$? This is controlled by the variance of $\mathcal{M}$, which we have already computed in the chapter of probability. Letting:

$$\sigma^2(\underline{\theta}) = Var_{\underline{\theta}}(X_i) = \int dx (x - \mu(\underline{\theta}))^2 p_{\underline{\theta}}(x)$$

we have:

$$Var_{\underline{\theta}}(\mathcal{M}) = \frac{\sigma^2(\underline{\theta})}{n}$$

As we have already learnt when studying the law of large numbers, the following inequality holds:

$$P_{\underline{\theta}}\left(|\mathcal{M} - \mu(\underline{\theta})| > \eta\right) \leq \frac{\sigma^2(\underline{\theta})}{n\eta^2}$$

for any $\eta > 0$. This means that, provided that the $\sigma^2(\underline{\theta}) < +\infty$ for all $\underline{\theta}$, increasing the number of data, i.e. the rank of the sample, the spreading of $\mathcal{M}$ around the expected value $\mu(\underline{\theta})$ becomes smaller and smaller, making the number $\frac{\sum_{i=1}^{n} x_i}{n}$ nearer and nearer to $\mu(\underline{\theta})$ for any realization of the sample.

This is another useful property of an estimator, **consistency**, expressed in general in the following:

Definizione 4 *An estimator $\mathcal{T}$ of a function $\tau(\underline{\theta})$ is called* **consistent** *if it converges in probability to $\tau(\underline{\theta})$ if the rank of the sample tends to $+\infty$.*

In order to proceed further, we need some assumption about the law of the X_i. A typical situation, quite always presented in textbooks, is the case when the X_i are **normal** with **known** variance σ^2. This can be a good model if we know a priori the sensibility of a given instrument, and is given by:

$$\theta \rightsquigarrow \mu_\theta(B) = P(X_i \in B) = \int_B dx \frac{\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \tag{7}$$

where the parameter $\theta \equiv \mu(\theta)$ is to be inferred from the data, while σ^2 is a fixed parameter, which we assume to know a priori. In such case, $\mathcal{M}$ is normal, being a linear combination of normal random variables. In particular, we have:

$$\mathcal{Z} = \frac{\mathcal{M} - \theta}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \tag{8}$$

This means that the statistic $\mathcal{M}$ is normally distributed around the unknown mean θ with a known variance, descreasing with the rank of the sample. Introducing the **quantiles** of a standard normal law, we can write the exact result:

$$P_\theta\left(-\phi_{1-\frac{\alpha}{2}} \leq \mathcal{Z} \leq \phi_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \tag{9}$$

or, equivalently:

$$P_\theta\left(\mathcal{M} - \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \leq \theta \leq \mathcal{M} + \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha \tag{10}$$

It is important to understand the meaning of this equality: let's fix $1 - \alpha = 0.95 = 95\%$, the **confidence level**; in such case, we have to use the quantile $\phi_{1-\frac{\alpha}{2}} = \phi_{0.975} = 1.96$. The **random interval**:

$$\left[\mathcal{M} - 1.96\sqrt{\frac{\sigma^2}{n}}, \mathcal{M} + 1.96\sqrt{\frac{\sigma^2}{n}}\right] \tag{11}$$

contains the unknown expected value θ, with probability $1-\alpha = 95\%$: for that reason, it is called **confidence interval** at the level $1-\alpha = 95\%$ for the parameter θ. This is a particular example of the following very general definition:

Definizione 5 *Given two real valued statistics $\mathcal{A}, \mathcal{B}$, the random interval $[\mathcal{A}, \mathcal{B}]$ is called* **confidence interval** *at the level $1-\alpha \in (0,1)$ of a function $\tau(\underline{\theta})$ if:*

$$P_{\underline{\theta}}(\tau(\underline{\theta}) \in [\mathcal{A}, \mathcal{B}]) \geq 1-\alpha \qquad \forall\ \underline{\theta} \in \Theta$$

To summarize, if our data $(x_1, \ldots, x_n)$ can be modelled with normal random variables with unknown expected value θ and known variance σ^2, the real number:

$$\overline{x} = \frac{\sum_{i=1}^n x_i}{n} \tag{12}$$

is a pointwise estimation of θ. Moreover, letting:

$$\delta\overline{x} = \sqrt{\frac{\sigma^2}{n}} \tag{13}$$

the interval:

$$[\overline{x} - 1.96\delta\overline{x}, \overline{x} + 1.96\delta\overline{x}] \tag{14}$$

is an estimation of a confidence interval at the level 95% for the unknown θ.

IV. COCHRAN THEOREM AND ESTIMATION OF THE VARIANCE

A far more general situation emerges in the case that the sample components are normal with both expectation and variance unknown.

$$\underline{\theta} = (\theta_0, \theta_1) \rightsquigarrow \mu_{\underline{\theta}}(B) = P(X_i \in B) = \int_B dx \frac{\exp\left(-\frac{(x-\theta_0)^2}{2\theta_1}\right)}{\sqrt{2\pi\theta_1}} \tag{15}$$

We show now how to use the data to estimate the unknown functions $\mu(\underline{\theta})$ and $\sigma^2(\underline{\theta})$, and to obtain two intervals in $\mathbb{R}$ containing respectively the two functions with a given confidence level.

Besides the statistic $\mathcal{M}$ introduced above, which is an unbiased and consistent estimator of $\mu(\underline{\theta}) = \theta_0$, we introduce the following estimator:

$$\mathcal{S}^2 = s^2(X_1 \ldots X_n) = \frac{\sum_{i=1}^n (X_i - \mathcal{M})^2}{n-1} \tag{16}$$

of $\sigma^2(\underline{\theta}) = \theta_1$. The presence of $n-1$ in the denominator makes $\mathcal{S}^2$ unbiased, as follows from the following calculation:

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	477	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

Table I: the 100 measurements of the velocity of light in air by A. Michelson (1879), from [**?**]; the given values plus 299000 are the original measurements in km/s

$$
\begin{aligned}
E_{\underline{\theta}}(\mathcal{S}^2) = E_{\underline{\theta}}\left[\frac{\sum_{i=1}^n X_i^2 - n\,\left(\frac{\sum_{j=1}^n X_j}{n}\right)^2}{n-1}\right] = \\
= \frac{nE_{\underline{\theta}}(X_i^2) - \frac{1}{n}E_{\underline{\theta}}\left(\left(\sum_{j=1}^n X_j\right)^2\right)}{n-1} = \\
= \frac{nE_{\underline{\theta}}(X_i^2) - \frac{1}{n}\left(Var_{\underline{\theta}}\left(\sum_{j=1}^n X_j\right) + \left(E_{\underline{\theta}}\left(\sum_{j=1}^n X_j\right)\right)^2\right)}{n-1} = \\
= \frac{nE_{\underline{\theta}}(X_i^2) - Var_{\underline{\theta}}\left(X_i\right) - n\left(E_{\underline{\theta}}\left(X_i\right)\right)^2}{n-1} = Var_{\underline{\theta}}(X_i) = \sigma^2(\underline{\theta})
\end{aligned}
$$

$\mathcal{S}^2$ is also consistent, as can be easily verified using the law of large numbers. The pointwise estimation of $\sigma^2(\underline{\theta})$ is thus:

$$
s^2(x_1 \dots x_n) = \frac{\sum_{i=1}^n \left(x_i - \frac{\sum_{j=1}^n x_j}{n}\right)^2}{n-1} \tag{17}
$$

In the following we will discover the law of the random variable:

$$
\mathcal{R} = \frac{\mathcal{M} - \theta_0}{\sqrt{\frac{\mathcal{S}^2}{n}}} \tag{18}
$$

which will turn out to follow a **Student law** with $n-1$ degrees of freedom, and this will allow us to build up confidence intervals for the mean, θ_0.

Moreover, it will turn out that the random variable:

$$\frac{n-1}{\theta_1}\,\mathcal{S}^2 \tag{19}$$

follows a **chi-square law** with $n-1$ degrees of freedom, allowing to build up confidence intervals for the variance, θ_1.

The rigorous justification of such results relies on the following:

Teorema 6 (Cochran) *Let $Y = (Y_1 \dots Y_n)$ be an n-dimensional normal random variable, $Y \sim N(\underline{0}, \mathbb{I})$. Moreover, let $E_1 \dots E_s$ be orthogonal vector subspaces of $\mathbb{R}^n$, such that $\bigoplus_{j=1}^{s} E_j = \mathbb{R}^n$. We denote $\Pi_1 \dots \Pi_s$ the linear projectors onto such subspaces. Then:*

1. *the random variables $\Pi_j X$, $j = 1 \dots s$, are independent.*
2. *the random variables $|\Pi_j X|^2$, $j = 1 \dots s$, has a chi-square law $\chi^2(d_j)$, where $d_j = dim(E_j)$ is the dimension of E_j.*

Proof. *Let $B = \{e_1 \dots e_n\}$ be the canonical base of $\mathbb{R}^n$, and let's write $Y = \sum_{i=1}^{n} Y_i e_i$.*

Moreover, if $\tilde{B}_1 \dots \tilde{B}_s$ are orthonormal basis of the subspaces $E_1 \dots E_s$, the set of vectors $\tilde{B} = \tilde{B}_1 \cup \dots \cup \tilde{B}_s = \{\tilde{e}_1 \dots \tilde{e}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ and we may write $Y = \sum_{j=1}^{n} \tilde{Y}_j \tilde{e}_j$ where:

$$\begin{pmatrix} \tilde{Y}_1 \\ \dots \\ \tilde{Y}_n \end{pmatrix} = \begin{pmatrix} (\tilde{e}_1|e_1) & \dots & (\tilde{e}_1|e_n) \\ \dots & \dots & \dots \\ (\tilde{e}_n|e_1) & \dots & (\tilde{e}_n|e_n) \end{pmatrix} \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}$$

the matrix Γ with matrix elements $\Gamma_{ij} = (\tilde{e}_i|e_j)$ being orthogonal. It follows that the random variable $(\tilde{Y}_1 \dots \tilde{Y}_n)$ is normal $N(\Gamma\underline{0} = \underline{0}, \Gamma\mathbb{I}\Gamma^\tau = \mathbb{I})$. é normale con media $\Gamma\underline{0} = \underline{0}$ e matrice di covarianza $\Gamma\mathbb{I}\Gamma^T = \mathbb{I}$. Thus, the components $\tilde{Y}_i$ are standard normal and independent.

$$\Pi_j Y = \sum_{i=1}^{n} (\tilde{e}_i|\Pi_j Y)\tilde{e}_i = \sum_{\tilde{e}_i \in \tilde{B}_j} \tilde{Y}_i \tilde{e}_i$$

implies that the $\Pi_j Y$ are independent. Finally:

$$|\Pi_j Y|^2 = \sum_{\tilde{e}_i \in \tilde{B}_j} \tilde{Y}_i^2 \sim \chi^2(d_j)$$

since the $\tilde{Y}_i$ are standard normal and independent.

Let's apply the Cochran theorem to the statistics $\mathcal{M}$ and $\mathcal{S}^2$. We let:

$$\tilde{e}_1 = \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \dots \\ \frac{1}{\sqrt{n}} \end{pmatrix}$$

and define $\tilde{E}_1$ as the one-dimensional subspace of $\mathbb{R}^n$ spanned by $\tilde{e}_1$. Moreover, we let $\tilde{E}_2$ be the orthogonal complement of $\tilde{E}_1$. Starting from the sample $X = (X_1, \dots, X_n)$, we define the n-dimensional random variable:

$$Y = \frac{X - \theta_0\sqrt{n}\,\tilde{e}_1}{\sqrt{\theta_1}} \tag{20}$$

which, by construction, follows the law $Y \sim N(\underline{0}, \mathbb{I})$. Using the notations of Cochran theorem, we have:

$$\Pi_1\, Y = \frac{\mathcal{M} - \theta_0}{\sqrt{\theta_1/n}}\,\tilde{e}_1 \qquad \Pi_2\, Y = \frac{X - \sqrt{n}\,\mathcal{M}\,\tilde{e}_1}{\sqrt{\theta_1}}$$

We know that $\Pi_1\, Y$ and $\Pi_2\, Y$ are **indipendent**. Moreover:

$$|\Pi_2\, Y|^2 = \frac{1}{\theta_1}\sum_{i=1}^{n}(X_i - \mathcal{M})^2 = \frac{(n-1)\mathcal{S}^2}{\theta_1} \tag{21}$$

has a law $\chi^2(n-1)$ and is **independent** on $\Pi_1\, Y$. It follows that the random variable:

$$\mathcal{R} = \frac{\left(\frac{\mathcal{M}-\theta_0}{\sqrt{\theta_1}/\sqrt{n}}\right)}{\sqrt{|\Pi_2\, Y|^2/(n-1)}} = \frac{\mathcal{M} - \theta_0}{\sqrt{\mathcal{S}^2/n}} \tag{22}$$

has a **Student law** $\mathcal{R} \sim t(n-1)$ with $n-1$ degrees of freedom.

We let $t_{1-\alpha/2}(n-1)$ be the **quantile** of order $1-\alpha/2$ of the Student law $t(n-1)$, defined by the following:

$$P_{\underline{\theta}}(\mathcal{R} \leq t_{1-\alpha/2}(n-1)) = 1 - \alpha/2 \tag{23}$$

Since the Student law is even ($\mathcal{R}$ and $-\mathcal{R}$ have the same law), we have $t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$ and thus we can write:

$$P_{\underline{\theta}}(-t_{1-\alpha/2}(n-1) \leq \mathcal{R} \leq t_{1-\alpha/2}(n-1)) = 1 - \alpha \tag{24}$$

We conclude that the random interval:

$$[\mathcal{M} - t_{1-\alpha/2}(n-1)\sqrt{\frac{\mathcal{S}^2}{n}}, \mathcal{M} + t_{1-\alpha/2}(n-1)\sqrt{\frac{\mathcal{S}^2}{n}}] \tag{25}$$

is a **confidence interval** at the level $1-\alpha$ for the mean $\mu(\underline{\theta}) = \theta_0$. If we replace the estimators with the pointwise estimations, we provide an estimation for the confidence interval which, using the widely employed notations:

$$x_{best} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \delta x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_{best})^2}{n(n-1)}}$$

is:

$$[x_{best} - t_{1-\alpha/2}(n-1)\delta x, x_{best} + t_{1-\alpha/2}(n-1)\delta x]$$

The most typical choice is the level 95%, which means $1-\alpha = 0.95$, that is $\alpha = 0.05$: we have to use the quantile $t_{1-\alpha/2}(n-1) = t_{0.975}(n-1)$, which, in our case $n = 100$, is nearly 1.985.

Nota 7 *Quite often, when the rank of the sample in large, in the expression of confidence intervals one replaces the quantile $t_{1-\alpha/2}(n-1)$ of the Student law with the ones of standard normal law $\phi_{1-\alpha/2}$ (naturally strictly independent on n). In our case such substitution would give a slightly smaller interval, being $\phi_{0.975} = 1.96$.*

For the variance, we have learnt that:

$$\frac{(n-1)\mathcal{S}^2}{\theta_1} \sim \chi^2(n-1) \tag{26}$$

so that:

$$P_{\underline{\theta}}\left(\chi^2_{\alpha/2}(n-1) \leq \frac{(n-1)\mathcal{S}^2}{\theta_1} \leq \chi^2_{1-\alpha/2}(n-1)\right) = 1-\alpha \tag{27}$$

where we have introduced the quantiles of the $\chi^2(n-1)$. We can rewrite the above expression in the following way:

$$P_{\underline{\theta}}\left(\frac{(n-1)\mathcal{S}^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \theta_1 \leq \frac{(n-1)\mathcal{S}^2}{\chi^2_{\alpha/2}(n-1)}\right) = 1-\alpha \tag{28}$$

which shows that:

$$\left[\frac{(n-1)\mathcal{S}^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)\mathcal{S}^2}{\chi^2_{\alpha/2}(n-1)}\right] \tag{29}$$

is a confidence interval at the level $1-\alpha$ for $\sigma^2(\underline{\theta}) = \theta_1$.

Typically, when a sample has rank $n > 30$, the following approximation turns out to be very accurate:

$$\chi^2_{1-\alpha/2}(n-1) \simeq \frac{1}{2}\left(\phi_{1-\alpha/2} + \sqrt{2(n-1)-1}\right)^2 \tag{30}$$

where $\phi_{1-\alpha/2}$ is the quantile of the standard normal law. In our case, at the level 95%, $\alpha = 0.05$, we have:

$$\chi^2_{1-\alpha/2}(n-1) \simeq 129.07, \quad \chi^2_{\alpha/2}(n-1) \simeq 73.77 \tag{31}$$

so that the confidence interval is $[0.77\mathcal{S}^2, 1.36\mathcal{S}^2]$.

A. Estimation of a proportion

Let's assume now that the sample $X = (X_1, \ldots, X_n)$ is made of Bernoulli random variables with parameter $\theta \in (0,1)$. We know that:

$$E_\theta(X_i) = \theta, \quad Var_\theta(X_i) = \theta(1-\theta)$$

so that the random variable:

$$\mathcal{M} = m(X_1 \ldots X_n) = \frac{\sum_{i=1}^n X_i}{n} \tag{32}$$

is an unbiased estimator for θ, that is:

$$E_\theta(\mathcal{M}) = \theta$$

Moreover, for large n, the law of:

$$\frac{\mathcal{M} - \theta}{\sqrt{\theta(1-\theta)/n}} \tag{33}$$

can be approximated by a law $N(0,1)$, for the central limit theorem. We can thus write:

$$P_\theta\left(-\phi_{1-\frac{\alpha}{2}} \leq \frac{\mathcal{M} - \theta}{\sqrt{\theta(1-\theta)/n}} \leq \phi_{1-\frac{\alpha}{2}}\right) \simeq 1 - \alpha \tag{34}$$

where, as before, $\phi_{1-\frac{\alpha}{2}}$ are the quantiles of the standard normal $N(0,1)$. We can rewrite the above formula as:

$$P^\theta\left(\mathcal{M} - q_{1-\frac{\alpha}{2}}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} \leq \theta \leq \mathcal{M} + q_{1-\frac{\alpha}{2}}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right) \simeq 1 - \alpha$$

In order to build up a confidence interval at the level $1-\alpha$ for the parameter θ we should solve the inequality:

$$-\phi_{1-\frac{\alpha}{2}} \leq \frac{\mathcal{M} - \theta}{\sqrt{\theta(1-\theta)/n}} \leq \phi_{1-\frac{\alpha}{2}}$$

with respect to θ, which is a simple exercise which we leave to the reader. In general, when n is large enough, the resulting confidence interval can be accurately approximated as:

$$\left[\mathcal{M} - \phi_{1-\frac{\alpha}{2}}\frac{\sqrt{\mathcal{M}(1-\mathcal{M})}}{\sqrt{n}}, \mathcal{M} + \phi_{1-\frac{\alpha}{2}}\frac{\sqrt{\mathcal{M}(1-\mathcal{M})}}{\sqrt{n}}\right] \tag{35}$$

V. CRAMER-RAO THEOREM

We have learnt till now to build up estimators $\mathcal{T}$ for functions $\tau(\underline{\theta})$, in particular for the mean and the variance, inside a given statistical model. We have seen that some nice properties of an estimator are unbiasedness and consistence. We are going now to explore more deeply the quality of an estimator. Naturally, the *precision* of our estimation will depend on the variance of $\mathcal{T}$, or, in higher dimensions, on its covariance matrix.

We start limiting our attention to real valued estimators and to a one dimensional parameter θ. Later we will generalize to higher dimensions.

We fix some working ipothesis, which are satisfied by a wide class of statistical models, including the exponential ones. First of all, we assume that the real valued components X_i of the sample $(X_1 \dots X_n)$ have densiy $p_\theta(\underline{x})$, which we ask to be differentiable with respect to the parameter θ. Moreover, we assume that, for any statistic $\mathcal{T} = t(X_1 \dots X_n)$,

integrable with respect to the density $p_\theta(x_1)\dots p_\theta(x_n)$, we can exchange integration and differentiation:

$$\frac{\partial}{\partial\theta}\Big[E_\theta[\mathcal{T}]\Big] = \int dx_1\dots dx_n\, t(x_1\dots x_n)\,\frac{\partial}{\partial\theta}\Big[p_\theta(x_1)\dots p_\theta(x_n)\Big]$$

A simple manipulation of the above identity leads to the following:

$$\frac{\partial}{\partial\underline{\theta}}\Big[E_\theta[\mathcal{T}]\Big] = E_\theta\left[\mathcal{T}\,\mathcal{S}\right] \tag{36}$$

where we have introduced the **score function**:

$$S = \frac{1}{p_\theta(X_1)\dots p_\theta(X_n)}\,\frac{\partial}{\partial\theta}\Big[p_\theta(X_1)\dots p_\theta(X_n)\Big] = \frac{\partial}{\partial\theta}\log\Big[p_\theta(X_1)\dots p_\theta(X_n)\Big]$$

which measures the *sensibility* of the density $p_\theta(x)$ with respect to the parameter θ.

The score function statistic has zero mean, as can be proved by using $\mathcal{T}=1$ in the identity (36):

$$\frac{\partial}{\partial\theta}\Big[E_\theta[1]\Big] = 0 = E_\theta\left[\mathcal{S}\right] \tag{37}$$

The *average sensibility* is thus measured by the variance of the score function, which is called **Fisher Information number**:

$$I(\theta) = E_\theta\left[S^2\right] = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log\Big[p_\theta(X_1)\dots p_\theta(X_n)\Big]\right)^2\right] \geq 0$$

We will show now that such Fisher information number is related to the *maximum precision* we can expect for one estimator.

We consider now an estimator $\mathcal{T} = t(X_1,\dots,X_n)$ of a quantity $\tau(\theta)$. If the estimator is **unbiased**, we have:

$$\tau(\theta) = E_\theta[\mathcal{T}] = \int dx_1\dots dx_n\, t(x_1\dots x_n)\Big[p_\theta(x_1)\dots p_\theta(x_n)\Big] \tag{38}$$

and, by construction:

$$\frac{d\tau(\theta)}{d\theta} = E_\theta\left[\mathcal{T}\,\mathcal{S}\right] = E_\theta\left[\left(\mathcal{T}-\tau(\theta)\right)\mathcal{S}\right] = Cov\left[\mathcal{T},\mathcal{S}\right] \tag{39}$$

where we have use the fact that the score function has zero mean. We can use now Cauchy-Schwartz inequality, implying that:

$$\left|\frac{d\tau(\theta)}{d\theta}\right|^2 = \left|Cov\left[\mathcal{T},\mathcal{S}\right]\right|^2 \leq Var(\mathcal{T})Var(\mathcal{S}) \tag{40}$$

so that:

$$Var_\theta(\mathcal{T}) \geq \frac{\left|\frac{d\tau(\theta)}{d\theta}\right|^2}{I(\theta)} \tag{41}$$

This is a very important inequality, due to **Cramer-Rao**. We stress that the Fisher information number is a property of the statistical model, and not of the estimator: nevertheless, it imposes a lower bound to the variance of estimators that can be built up. Naturally, the smallest is the variance, the highest is the precision of the estimation: $I(\theta)$ controls the precision of the estimators.

Definizione 8 *An estimator $\mathcal{T}$ of a quantity $\tau(\theta)$ is called* **efficient** *if:*

$$Var_\theta(\mathcal{T}) = \frac{|\frac{d\tau(\theta)}{d\theta}|^2}{I(\theta)} \tag{42}$$

Let's consider an instructive example: let's assume that the component of the sample X_i are normal with unknown mean θ and **known** variance σ^2. We have:

$$\left(\frac{\partial}{\partial\theta}\log\Big[p_\theta(X_1)\dots p_\theta(X_n)\Big]\right)^2 = \left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log\left(\frac{\exp\left(-\frac{(X_i-\theta)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma}\right)\right)^2 =$$
$$= \left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\left(-\frac{(X_i-\theta)^2}{2\sigma^2} - \log(\sqrt{2\pi}\,\sigma)\right)\right)^2 = \left(\sum_{i=1}^n \frac{X_i-\theta}{\sigma^2}\right)^2$$

We get thus, exploiting independence, the following result for the Fisher information:

$$I(\theta) = E_\theta\left[\left(\sum_{i=1}^n \frac{X_i-\theta}{\sigma^2}\right)^2\right] = nVar_\theta\left(\frac{X_i-\theta}{\sigma^2}\right) = nE_\theta\left[\left(\frac{X_i-\theta}{\sigma^2}\right)^2\right] = \frac{n}{\sigma^2} \tag{43}$$

independent on θ. On the other hand, if we consider the estimator:

$$\mathcal{M} = m(X_1\dots X_n) = \frac{\sum_{i=1}^n X_i}{n} \tag{44}$$

of $\tau(\theta) = \theta$ (whose derivative is one!), we already know that:

$$Var(\mathcal{M}) = \frac{\sigma^2}{n} = \frac{1}{I(\theta)} \tag{45}$$

so that $\mathcal{M}$ is an efficient estimator of the mean: keeping fixed the statistical model, it is not possible to build up an estimator for the mean with variance lower than $\frac{\sigma^2}{n}$.

We present now the general statement of the Cramer-Rao theorem.

Teorema 9 (Cramer, Rao) *Let $\{(\Omega, \mathcal{F}, P_{\underline{\theta}})\}_{\underline{\theta}\in\Theta}$ a statistical model and $X = (X_1\dots X_n)$ a sample of rank n such that the following ipothesis hold:*

1. *the law of the components X_i of the sample $(X_1\dots X_n)$ has densiy $p_{\underline{\theta}}(\underline{x})$*

2. $p_{\underline{\theta}}(\underline{x})$ is differentiable with respect to $\underline{\theta}$

3. for any s-dimensional statistic $\mathcal{T} = t(X_1 \dots X_n)$ we can write:

$$\frac{\partial}{\partial \underline{\theta}}\Big[E_{\underline{\theta}}(\mathcal{T}_i)\Big] = \int dx_1 \dots dx_n \, t_i(\underline{x}_1 \dots \underline{x}_n) \, \frac{\partial}{\partial \underline{\theta}}\Big[p_{\underline{\theta}}(\underline{x}_1) \dots p_{\underline{\theta}}(\underline{x}_n)\Big]$$

If $\mathcal{T}$ is an estimator of the quantity $\tau(\underline{\theta})$ with finite expectation $E_{\underline{\theta}}(\mathcal{T})$, then the following matrix inequality holds:

$$cov_{\underline{\theta}}(\mathcal{T}) \geq J(\underline{\theta}) I^{-1}(\underline{\theta}) J(\underline{\theta})^T \tag{46}$$

where $J(\underline{\theta})$ is the Jacobian of $E_{\underline{\theta}}(\mathcal{T})$:

$$J_{ik}(\underline{\theta}) = \frac{\partial E_{\underline{\theta}}(\mathcal{T}_i)}{\partial \theta_k} \tag{47}$$

and $I(\underline{\theta})$ is the **Fisher Information matrix***:*

$$I_{ij}(\underline{\theta}) = E_{\underline{\theta}}\left(S_i(X_1 \dots X_n) S_j(X_1 \dots X_n)\right)$$

the **score vector** *being defined by:*

$$S(X_1 \dots X_n) = \frac{\partial}{\partial \underline{\theta}} \log \Big[p_{\underline{\theta}}(X_1) \dots p_{\underline{\theta}}(X_n)\Big]$$

Proof. *Using the constant statistic $\mathcal{T} = 1$, we see that the components of the score vector have zero mean. We introduce the matrix:*

$$A_{ik} = E_{\underline{\theta}}\Big(\big(\mathcal{T}_i - E_{\underline{\theta}}(\mathcal{T}_i)\big) \, S_k\Big) = E_{\underline{\theta}}\Big(\mathcal{T}_i \, S_k\Big) \tag{48}$$

Moreover, by inspection we see that:

$$E_{\underline{\theta}}\Big(\mathcal{T}_i \, S_k\Big) = \frac{\partial E_{\underline{\theta}}(\mathcal{T}_i)}{\partial \theta_k} = J_{ik}(\underline{\theta}) \tag{49}$$

This implies that, for each couple of vectors $\underline{a} \in \mathbb{R}^s$, $\underline{b} \in \mathbb{R}^m$:

$$(\underline{a}|A|\underline{b}) = (\underline{a}|J(\underline{\theta})|\underline{b}) \tag{50}$$

The left hand side has the explicit form:

$$(\underline{a}|A|\underline{b}) = \sum_{i=1}^{s}\sum_{j=1}^{m} a_i A_{ij} b_j = E_{\underline{\theta}}\Big((\underline{a}|\mathcal{T} - E_{\underline{\theta}}(\mathcal{T}))\,(S|\underline{b})\Big)$$

Hölder inequality implies:

$$(\underline{a}|A|\underline{b})^2 \leq E\Big((\underline{a}|\mathcal{T} - E_{\underline{\theta}}(\mathcal{T}))^2\Big)\, E\Big((S|\underline{b})^2\Big)$$

that is:

$$(\underline{a}|A|\underline{b})^2 \leq (\underline{a}|cov_{\underline{\theta}}(\mathcal{T})|\underline{a})\,(\underline{b}|I(\underline{\theta})|\underline{b}) \tag{51}$$

Finally:

$$(\underline{a}|J(\underline{\theta})|\underline{b})^2 = \underline{a}|A|\underline{b})^2 \leq (\underline{a}|cov_{\underline{\theta}}(\mathcal{T})|\underline{a})\,(\underline{b}|I(\underline{\theta})|\underline{b})$$

Choosing $\underline{b} = I(\underline{\theta})^{-1} J(\underline{\theta})^T \underline{a}$ and using the symmetry of Fisher information matrix (and of its inverse), we find:

$$(\underline{a}|J(\underline{\theta})|\underline{b})(\underline{a}|J(\underline{\theta}) I(\underline{\theta})^{-1} J(\underline{\theta})^T|\underline{a}) \leq (\underline{a}|cov_{\underline{\theta}}(\mathcal{T})|\underline{a})\,(\underline{a}|J(\underline{\theta})|\underline{b})$$

that is:

$$(\underline{a}|cov_{\underline{\theta}}(\mathcal{T})|\underline{a}) \geq (\underline{a}|J(\underline{\theta}) I(\underline{\theta})^{-1} J(\underline{\theta})^T|\underline{a})$$

A. Maximum likelihood estimators (MLE)

We have till now learnt some useful properties of estimators, determining their *precision* in inferring $\tau(\underline{\theta})$ from a set of data. A natural question is whether there exist a tool to *invent* an estimator for a particular $\tau(\underline{\theta})$. In the case of mean and variance the actual definition of the estimator is very natural, but there can be situations in which the choice is not so simple. We limit our attention to the case when the quantity $\tau(\underline{\theta})$ to be estimated is the parameter $\underline{\theta}$ itself. We assume moreover that the components of the sample $(X_1 \dots X_n)$ have density $p_{\underline{\theta}}(\underline{x})$.

Given the data $(\underline{x}_1 \dots \underline{x}_n)$ let's consider the **likelihood function**:

$$L(\underline{\theta}; \underline{x}_1 \dots \underline{x}_n) \stackrel{def}{=} p_{\underline{\theta}}(\underline{x}_1) \dots p_{\underline{\theta}}(\underline{x}_n) \tag{52}$$

Intuitively, "$L(\underline{\theta}; \underline{x}_1 \dots \underline{x}_n) d\underline{x}_1 \dots d\underline{x}_n$" is the *probability* to obtain precisely the measured data for the value $\underline{\theta}$ of the parameter. We are naturally induced to estimate the unknown parameter as the value $\underline{\theta}$ which **maximizes** such *probability*. This justifies the following:

Definizione 10 *We call* **maximum likelihood estimator** *(MLE) of the parameter $\underline{\theta}$ the statistic:*

$$\tilde{\underline{\theta}}_{ML}(X_1 \dots X_n) = \arg \max_{\underline{\theta} \in \Theta} L(\underline{\theta}; X_1 \dots X_n) \tag{53}$$

We observe that such estimator is well defined whenever, for the given data, the function $\underline{\theta} \rightsquigarrow L(\underline{\theta}; \underline{x}_1 \dots \underline{x}_n)$ has a unique maximum.

In order to give a first example, let's consider again the normal sample with unknown mean θ and known variance σ^2. In such case:

$$L(\theta; X_1 \dots X_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(- \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2\sigma^2} \right) \tag{54}$$

The maximization of such function with respect to θ leads to the following equation:

$$0 = \frac{\partial}{\partial \theta} \left(- \sum_{i=1}^{n} (X_i - \theta)^2 \right) = 2 \sum_{i=1}^{n} (X_i - \theta)$$

which implies:

$$\tilde{\underline{\theta}}_{ML}(X_1 \dots X_n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

which is exactly the empirical mean.

The MLE for some important models are summarized in the following table:

model	$\tilde{\underline{\theta}}_{ML}$	$E_{\underline{\theta}}(\tilde{\underline{\theta}}_{ML})$	$\mathrm{cov}_{\underline{\theta}}(\tilde{\underline{\theta}}_{ML})$	$I(\underline{\theta})$
bernoulli	$\frac{\sum_{i=1}^n X_i}{n}$	θ	$\frac{\theta(1-\theta)}{n}$	$\frac{n}{\theta(1-\theta)}$
gaussian	$\begin{pmatrix} \frac{\sum_{i=1}^n X_i}{n} \\ \frac{\sum_{i=1}^n X_i^2}{n} - \left(\frac{\sum_{i=1}^n X_i}{n}\right)^2 \end{pmatrix}$	$\begin{pmatrix} \theta_0 \\ \frac{n-1}{n}\theta_1 \end{pmatrix}$	$\begin{pmatrix} \frac{\theta_1}{n} & 0 \\ 0 & \frac{n-1}{n^2} 2\theta_1^2 \end{pmatrix}$	$\begin{pmatrix} \frac{n}{\theta_1} & 0 \\ 0 & \frac{n}{2\theta_1^2} \end{pmatrix}$
exonential	$\frac{n}{\sum_{i=1}^n X_i}$	$\frac{n}{n-1}\theta$	$\frac{n^2}{(n-1)^2(n-2)}\theta^2$	$\frac{n}{\theta^2}$

We observe that the MLE may be not unbiased nor efficient, but they asyntotically have these properties in the limit of large samples.

The following theorems, which we state without proof, provide general results about MLEs. The first result concernes existence of MLEs.

Teorema 11 (Wald) *If:*

1. Θ *is compact.*
2. *for each* $\underline{x}$ *the density* $p_{\underline{\theta}}(\underline{x})$ *is a continuous function of* $\underline{\theta}$*.*
3. $p_{\underline{\theta}}(\underline{x}) = p_{\underline{\theta}'}(\underline{x})$ *if and only if* $\underline{\theta} = \underline{\theta}'$
4. *there exists a positive function* $K : \mathbb{R}^k \to \mathbb{R}$*, such that the random variable* $K(X_i)$ *has finite expectation and such that, for each* $\underline{x}$ *and* $\underline{\theta}$*:*

$$\left|\log\left[\frac{p_{\underline{\theta}}(\underline{x})}{p_{\underline{\theta}'}(\underline{x})}\right]\right| \leq K(\underline{x}) \tag{55}$$

Then there exist a maximum likelihood estimator $\tilde{\underline{\theta}}_{ML}(X_1 \dots X_n)$ *that converges almost surely to* $\underline{\theta}$ *as the rank of the sample increases to* $+\infty$*.*

A strongest result is the following:

Teorema 12 (Cramer) *If:*

1. Θ *is open.*
2. *for each* $\underline{x}$*,* $p_{\underline{\theta}}(\underline{x}) \in \mathcal{C}^2(\Theta)$*, and it is possible to exchange derivative and expectation.*
3. $p_{\underline{\theta}}(\underline{x}) = p_{\underline{\theta}'}(\underline{x})$ *if and only if* $\underline{\theta} = \underline{\theta}'$
4. *there exists a function* $K(\underline{x})$ *such that* $K(X_i)$ *has finite expectation and:*

$$\|\nabla_{\underline{\theta}} \log\left(p_{\underline{\theta}}(\underline{x})\right)\| \leq K(\underline{x}) \tag{56}$$

for each $\underline{x} \in \mathbb{R}^k$*.*

Then ther exists a maximum likelihood estimator $\underline{\tilde{\theta}}_{ML}(X_1 \dots X_n)$ *that converges almost surely to* $\underline{\theta}$ *as the rank of the sample increases to* $+\infty$*, and that is asyntotically normal and efficient:*

$$\begin{aligned} \underline{\tilde{\theta}}_{ML}(X_1 \dots X_n) &\xrightarrow[n\to\infty]{a.s.} \underline{\theta} \\ \lim_{n\to\infty} \sqrt{n} \left[\underline{\tilde{\theta}}_{ML}(X_1 \dots X_n) - \underline{\theta}\right] &= N\left[\underline{0}, I(\underline{\theta})^{-1}\right] \end{aligned} \tag{57}$$

VI. HYPOTHESIS TESTS

A typical problem in mathematical statistics is to use the data to confirm or reject an ipothesis relying on a set of data. Once fixed a statistical model, an ipothesis is a statement about the parameter $\underline{\theta}$. In practice, the statistical ipothesis to be tested, called the **null ipothesis** H_0 (that in general the tester tries to *reject*) can be expressed as:

$$H_0 : \quad \underline{\theta} \in \Theta_0 \tag{58}$$

while the **alternative ipothesis** H_1, (that in general the tester tries to *establish*) can be expressed as:

$$H_1 : \quad \underline{\theta} \in \Theta_1 = \Theta - \Theta_0 \tag{59}$$

We start from the data $(\underline{x}_1 \dots \underline{x}_n)$. Performing a statistical test means choosing a subset $\Omega_R \subset \mathbb{R}^k \times \dots \times \mathbb{R}^k$, called **critical region**, such that we **reject** the null ipothesis if $(\underline{x}_1 \dots \underline{x}_n) \in \Omega_R$:

$$(\underline{x}_1 \dots \underline{x}_n) \in \Omega_R \quad \Rightarrow \quad reject\ H_0 \tag{60}$$

In such case the conclusion is that H_0 is **not** consistent with the data. Naturally, the randomness in the experiment can leads to errors: if we **reject** H_0 when H_0 is true, we say we do a **type I error**; on the other hand, if we do **not reject** H_0 when H_0 is false, we say that we do a **type II error**.

In most cases, the critical region is expressed in term of a statistic $\mathcal{T} = t(X_1, \dots, X_n)$, in the form:

$$\Omega_R = \{(\underline{x}_1 \dots \underline{x}_n) : t(\underline{x}_1 \dots \underline{x}_n) > \mathcal{T}_0\}$$

for a given treshold value $\mathcal{T}_0$.

A. Student test

One very common experimental situation is the comparison between the mean of a measured quantity and a reference value, maybe coming from a theoretical study. We assume that the data $(x_1 \dots x_n)$ can be modeled as realization of a sample $(X_1 \dots X_n)$, with one dimensional **normal** components. Introducing the statistics $\mathcal{M}$ and $\mathcal{S}^2$, respectively estimators of mean $\mu(\underline{\theta})$ and variance $\sigma^2(\underline{\theta})$, we already know that:

$$\mathcal{R} = r(X_1, \dots, X_n) = \frac{\mathcal{M} - \mu(\underline{\theta})}{\sqrt{\frac{\mathcal{S}^2}{n}}} \tag{61}$$

follows a Student law with $n-1$ degrees of freedom.

We denote μ_0 the reference value. We test the ipothesis:

$$H_0: \quad \mu(\underline{\theta}) = \mu_0 \tag{62}$$

against:

$$H_1: \quad \mu(\underline{\theta}) \neq \mu_0 \tag{63}$$

Naturally we will reject H_0 if the estimated mean is far from μ_0. **If the null ipothesis H_0 is true**, we can calculate:

$$P_{\underline{\theta}}\left(\left|\frac{\mathcal{M} - \mu_0}{\sqrt{\frac{\mathcal{S}^2}{n}}}\right| > t_{1-\frac{\alpha}{2}}(n-1)\right) = \alpha$$

for any $\alpha \in (0,1)$. At the **significance level** α, we can define the critical region as:

$$\Omega_R = \{(x_1 \dots x_n) : |r(x_1 \dots x_n)| > t_{1-\frac{\alpha}{2}}(n-1)\}$$

Typically chosen values are $\alpha = 0.10, 0.05, 0.01$, corresponding, for large samples, to the quantiles $1.645, 1.96, 2.58$. We note that α is precisely the **probability of type I error**.

Given the data $(x_1 \dots x_n)$, we can thus imediately calculate the *standardized discrepancy* with respect to the reference value: $r(x_1 \dots x_n)$. We can also, using the Student law or the normal if the sample is large enough, compute the **p value**:

$$p\ value = P_{\underline{\theta}}\left(\left|\frac{\mathcal{M} - \mu_0}{\sqrt{\frac{\mathcal{S}^2}{n}}}\right| > r(x_1 \dots x_n)\right) \tag{64}$$

under the assumption that H_0 is true. If the p value is less than or equal to the significance level α, H_0 is rejected at the significance level α; the p value is the probability to find data *worse* than the ones we have measured if H_0 is true. A small p value means that is very unlikely that H_0 is consistent with the data.

Another important class of ipothesis that are often tested have the form:

$$H_0: \quad \mu(\underline{\theta}) \leq \mu_0 \tag{65}$$

Naturally, the alternative is:

$$H_1: \quad \mu(\underline{\theta}) > \mu_0 \tag{66}$$

It is clear that we will reject the null ipothesis if we get a mean much bigger than μ_0. In order to be quantitative, we observe that:

$$\mathcal{R} = \frac{\mathcal{M} - \mu_0}{\sqrt{\frac{\mathcal{S}^2}{n}}} = \frac{\mathcal{M} - \mu(\underline{\theta})}{\sqrt{\frac{\mathcal{S}^2}{n}}} + \frac{\mu(\underline{\theta}) - \mu_0}{\sqrt{\frac{\mathcal{S}^2}{n}}} \tag{67}$$

is the sum of a Student random variable $t(n-1)$ and a term which is always negative if H_0 is true. Thus:

$$P_{\underline{\theta}}\left(\mathcal{R} > t_{1-\alpha}(n-1)\right) \leq P_{\underline{\theta}}\left(\frac{\mathcal{M} - \mu(\underline{\theta})}{\sqrt{\frac{\mathcal{S}^2}{n}}} > t_{1-\alpha}(n-1)\right) = \alpha$$

and we may set the critical region:

$$\Omega_R = \{(x_1 \dots x_n) : r(x_1 \dots x_n) > t_{1-\alpha}(n-1)\}$$

defines a statistical test of the ipothesis (65) whose probability of I type error is α.

In several situations two different estimations of averages of **independent normal** samples $X = (X_1 \dots X_n)$ and $Y = (Y_1 \dots Y_m)$ are compared. We will limit our attention to the situation in which the two independent samples share the same value for the variance.

In the simplest case, the hypothesis that the two means are equal:

$$H_0 : \quad \mu_X(\underline{\theta}) = \mu_Y(\underline{\theta}) \tag{68}$$

is tested against the alternative:

$$H_1 : \quad \mu_X(\underline{\theta}) \neq \mu_Y(\underline{\theta}) \tag{69}$$

Using Cochran theorem, it is simple to show that, if H_0 is true, the random variable:

$$\mathcal{T} = \frac{\mathcal{M}_X - \mathcal{M}_Y}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)\mathcal{S}_X^2 + (m-1)\mathcal{S}_Y^2}{n+m-2}}}$$

follows a Student law with $n+m-2$ degrees of freedom $t(n+m-2)$. The above notation is precisely the same we have used throughout this chapter a part from a label to distinguish the two samples: $\mathcal{M}_X = \frac{1}{n}\sum_{i=1}^n X_i$ and so on. We have thus:

$$P_{\underline{\theta}}\left(|\mathcal{T}| > t_{1-\frac{\alpha}{2}}(n+m-2)\right) = \alpha \tag{70}$$

providing a critical region at the significance level α of the form:

$$\Omega_R = \{(x_1 \dots x_n; y_1 \dots y_m) : |t| > t_{1-\frac{\alpha}{2}}(n-1)\}$$

where $t = \mathcal{T}(x_1 \dots x_n; y_1 \dots y_m)$. Intuitively, if the two estimations of the means turn out to be "too different", we reject the hypothesis.

If we have to test the hypothesis:

$$H_0 : \quad \mu_X(\underline{\theta}) \leq \mu_Y(\underline{\theta}) \tag{71}$$

we rely on the observation that:

$$\mathcal{T} = \frac{\mathcal{M}_X - \mathcal{M}_Y}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)\mathcal{S}_X^2 + (m-1)\mathcal{S}_Y^2}{n+m-2}}} = \frac{(\mathcal{M}_X - \theta_{0X}) - (\mathcal{M}_Y - \theta_{0Y})}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)\mathcal{S}_X^2 + (m-1)\mathcal{S}_Y^2}{n+m-2}}} + \frac{\theta_{0X} - \theta_{0Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)\mathcal{S}_X^2 + (m-1)\mathcal{S}_Y^2}{n+m-2}}}$$

is the sum of a random variable with law $t(n+m-2)$ and a term which, if H_0 is true, is always negative or equal to zero. Therefore:

$$P_{\underline{\theta}}\left(\mathcal{T} > t_{1-\alpha}(n+m-2)\right) = \alpha \tag{72}$$

and the critival region for a test at the significance level α is:

$$\Omega_R = \{(x_1 \dots x_n; y_1 \dots y_m) : t > t_{1-\alpha}(n+m-2)\}$$

where $t = \mathcal{T}(x_1 \dots x_n; y_1 \dots y_m)$.

B. Chi-Squared test

The **Chi-Squared test**, or **Goodness-of-Fit test**, due to Pearson, is a test of the ipothesis:

$$H_0 : \quad \underline{\theta} = \underline{\theta}_0$$

aiming to verify whether the probability density $p_{\underline{\theta}_0}(\underline{x})$, specified by the value $\underline{\theta}_0$ of the parameter, is a good description of the experiment we have made, given a set of data $(\underline{x}_1 \dots \underline{x}_n)$. The starting point is a partition of the range of the measurements, $\mathbb{R}^k$, in a **finite** family $\{E_j\}_{j=1}^r$ of **outcomes**, mutually disjoint such that $\bigcup_{j=1}^r E_j = \mathbb{R}^k$. The basic idea of the test is to compare the **theoretical frequencies**:

$$p_j(\underline{\theta}_0) = P_{\underline{\theta}_0}(X_i \in E_j) = \int_{E_j} d\underline{x}\, p_{\underline{\theta}_0}(\underline{x})$$

with the **empirical frequencies**:

$$f_j = \frac{\sum_{i=1}^n 1_{E_j}(\underline{x}_i)}{n}$$

giving the number of measurements fallen in the set E_j.

As usual, we interpret the numbers f_j as realizations of the statistics:

$$\mathcal{N}_j = n_j(X_1 \dots X_n) = \frac{\sum_{i=1}^n 1_{E_j}(X_i)}{n}$$

The discrepancy between theoretical and empirical frequencies builds up the **Pearson random variable**:

$$\mathcal{P} = p(X_1, \dots, X_n) = \sum_{j=1}^r n\, \frac{\left(\mathcal{N}_j - p_j(\underline{\theta}_0)\right)^2}{p_j(\underline{\theta}_0)} \tag{73}$$

The key result is the following, which we will prove at the end of this section:

Teorema 13 *If the ipothesis:*

$$H_0 : \quad \underline{\theta} = \underline{\theta}_0$$

is true, the Pearson random variable converges in distribution to a random variable $\chi^2(r-1)$, as the rank of the sample tends to $+\infty$.

Thus, assuming H_0 true, if the sample is large enough, we have:

$$P_{\underline{\theta}_0}\left(\mathcal{P} \geq \chi^2_{1-\alpha}(r-1)\right) = \alpha \tag{74}$$

so that we can define the critical region for the test of H_0 at the significance level α.

$$\Omega_R = \left\{(\underline{x}_1 \dots \underline{x}_n) : p(\underline{x}_1 \dots \underline{x}_n) > \chi^2_{1-\alpha}(r-1)\right\} \tag{75}$$

Let's now prove the basic theorem (13). Keeping in mind the idea of dealing with samples with arbitrary size, We consider the sequence of r-dimensional random variables:

$$Z_n = \sqrt{n}\left(\mathcal{N} - p(\underline{\theta}_0)\right)$$

where:

$$\mathcal{N} = \begin{pmatrix} n_1(X_1 \dots X_n) \\ \dots \\ n_r(X_1 \dots X_n) \end{pmatrix} \qquad p(\underline{\theta}_0) = \begin{pmatrix} p_1(\underline{\theta}_0) \\ \dots \\ p_r(\underline{\theta}_0) \end{pmatrix}$$

The definition:

$$\mathcal{N} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} 1_{E_1}(X_i) \\ \dots \\ 1_{E_r}(X_i) \end{pmatrix}$$

guarantees that $\mathcal{N}$ is the empirical mean of n random variables independent and identically distributed with mean:

$$E_{\underline{\theta}}\left(\begin{pmatrix} 1_{E_1}(X_i) \\ \dots \\ 1_{E_r}(X_i) \end{pmatrix}\right) = p(\underline{\theta}_0)$$

and covariance matrix:

$$\begin{aligned} \Sigma_{jk} &= E_{\underline{\theta}_0}\left(1_{E_j}(X_i)1_{E_k}(X_i)\right) - E_{\underline{\theta}_0}\left(1_{E_j}(X_i)\right) E_{\underline{\theta}_0}\left(1_{E_k}(X_i)\right) = \\ &= \delta_{jk}\, p_j(\underline{\theta}_0) - p_j(\underline{\theta}_0)\, p_k(\underline{\theta}_0) \end{aligned}$$

The multidimensional central limit theorem guarantees thus that:

$$\lim_{n\to\infty} Z_n = N(\underline{0}, \Sigma)$$

We observe now that the Pearson random variable may be expressed as an inner product in $\mathbb{R}^r$:

$$\mathcal{P} = (AZ_n | AZ_n)$$

where $A \in M_{r\times r}(\mathbb{R})$ is the diagonal matrix:

$$A = \operatorname{diag}\left(\frac{1}{\sqrt{p_1(\underline{\theta}_0)}} \dots \frac{1}{\sqrt{p_r(\underline{\theta}_0)}}\right) \tag{76}$$

Using the identities:

$$E_{\underline{\theta}_0}(AZ_n) = AE_{\underline{\theta}}(Z_n) = \underline{0} \qquad \operatorname{cov}(AZ_n) = A\operatorname{cov}(Z_n)A^T$$

we conclude that:

$$\lim_{n\to\infty} A\, Z_n = N(\underline{0}, A\Sigma A^T)$$

Performing the product of matrices we get:

$$(A\Sigma A^T)_{jk} = \delta_{jk} - \sqrt{p_j(\underline{\theta}_0)}\sqrt{p_k(\underline{\theta}_0)}$$

By inspection we see that $A\Sigma A^T$ is a projection matrix of rank:

$$\mathrm{rg}(A\Sigma A^T) = \mathrm{tr}(A\Sigma A^T) = \sum_k (A\Sigma A^T)_{kk} = r - 1$$

There exists thus an orthogonal matrix $U \in M_{r\times r}(\mathbb{R})$ such that:

$$A\Sigma A^T = U\,\mathrm{diag}(1\dots 10)\,U^T \equiv U\,\Delta\,U^T$$

The sequence UAZ_n converges thus in distribution to a random variable $N(\underline{0}, \Delta)$, whose components are independent: the first $r-1$ follow a standard normal, the last one is the constant 0.

The sequence $(UAZ_n|UAZ_n)$ converges thus in law to a $\chi^2(r-1)$ and:

$$\lim_{n\to\infty} \mathcal{P} = \lim_{n\to\infty} (AZ_n|AZ_n) = \lim_{n\to\infty} (UAZ_n|UAZ_n) \sim \chi^2(r-1) \qquad (77)$$

This completes the proof.

C. Kolmogorov-Smirnov test

The weak point of the Pearson χ^2 test is the necessity of introducing the partition $\{E_j\}_{j=1}^r$ of the outcomes, which is quite arbitrary. In this section we will describe a different approach due to Kolmogorov and Smirnov. The aim is again to test the hypothesis:

$$H_0: \quad \underline{\theta} = \underline{\theta}_0$$

We will assume to deal with a sample $(X_1 \dots X_n)$ made of one-dimensional random variables with cumulative distribution function $F_{\underline{\theta}} : \mathbb{R} \to [0,1]$ that is **continuous** and **stricty increasing** for all $\underline{\theta}$.

We introduce now the **empirical cumulative distribution function** of the sample:

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \Theta(x - X_i)$$

$\tilde{F}_n(x)$, for all $x \in \mathbb{R}$, is a random variable counting the number of outcomes smaller or equal to x. The following calculation shows that $\tilde{F}_n(x)$ is an unbiased and consistent estimator of the "true" cumulative distribution function $F_{\underline{\theta}}$:

$$E_{\underline{\theta}}[\tilde{F}_n(x)] = \frac{1}{n}\sum_{i=1}^n E_{\underline{\theta}}[\Theta(x - X_i)] = P_{\underline{\theta}}(X \le x) = F_{\underline{\theta}}(x)$$

$$\mathrm{var}_{\underline{\theta}}[\tilde{F}_n(x)] = \frac{1}{n^2}\sum_{ij=1}^n E_{\underline{\theta}}[\Theta(x - X_i)\Theta(x - X_j)] - F_{\underline{\theta}}(x)^2 = \frac{F_{\underline{\theta}}(x)(1 - F_{\underline{\theta}}(x))}{n}$$

Moreover, we are going now to show the following important result:

Teorema 14 (Glivenko-Cantelli) *The* **Kolmogorov-Smirnov random variable***:*

$$\sup_{x\in\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}}(x)|$$

converges in probability to zero, that is:

$$\lim_{n\to\infty} P_{\underline{\theta}}\Big(\sup_{x\in\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}}(x)| \leq \epsilon \Big) = 1$$

$\forall\ \epsilon > 0$.

Proof. *Let's fix $\epsilon > 0$, choose $k \in \mathbb{N}$, $k \geq \frac{1}{2\epsilon}$ and consider the points $x_j = F^{-1}_{\underline{\theta}}\left(\frac{j}{k}\right)$ with $j = 0 \dots k$. Then:*

$$F_{\underline{\theta}}(x_{j+1}) - F_{\underline{\theta}}(x_j) = \frac{j+1}{k} - \frac{j}{k} = \frac{1}{k} \leq \epsilon$$

As we have observed above, in each point x_j the empirical cumulative distribution function $\tilde{F}_n(x_j)$ converges in probability to $F_{\underline{\theta}}(x_j)$. Then the random variable:

$$\Delta_k = \max_{j=0\dots k} |\tilde{F}_n(x_j) - F_{\underline{\theta}}(x_j)|$$

converges in probability to zero. Since $\forall\ x \in \mathbb{R}$ there exists one and only one j such that $x \in [x_{j-1}, x_j)$ we can write:

$$\tilde{F}_n(x) - F_{\underline{\theta}}(x) \leq \tilde{F}_n(x_j) - F_{\underline{\theta}}(x_{j-1}) = \tilde{F}_n(x_j) - F_{\underline{\theta}}(x_j) + F_{\underline{\theta}}(x_j) - F_{\underline{\theta}}(x_{j-1})$$

$$|\tilde{F}_n(x) - F_{\underline{\theta}}(x)| \leq |\tilde{F}_n(x_j) - F_{\underline{\theta}}(x_j)| + |F_{\underline{\theta}}(x_j) - F_{\underline{\theta}}(x_{j-1})| \leq \Delta_k + \epsilon$$

The fact that the last member is independent of x allows to write:

$$\sup_{\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}}(x)| \leq \Delta_k + \epsilon$$

which implies:

$$P_{\underline{\theta}}\Big(\sup_{x\in\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}}(x)| \leq 2\epsilon \Big) \geq P_{\underline{\theta}}\big(\Delta_k + \epsilon \leq 2\epsilon\big) = P_{\underline{\theta}}\big(\Delta_k \leq \epsilon\big)$$

This completes the proof since Δ_k converger in probability to zero:

$$\lim_{n\to\infty} P_{\underline{\theta}}\left(\sup_{\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}}(x)| \leq 2\epsilon \right) \geq 1$$

Let's consider now the random variable:

$$\sqrt{n}\sup_{x\in\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}}(x)| = \sqrt{n}\sup_{x\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} \Theta(x - X_i) - F_{\underline{\theta}}(x)|$$

for the given sample. Since $F_{\underline{\theta}}$ is invertible by construction, we can write:

$$\sqrt{n}\sup_{\mathbb{R}}|\tilde{F}_n(x)-F_{\underline{\theta}}(x)| = \sqrt{n}\sup_{[0,1]}\left|\tilde{F}_n(F_{\underline{\theta}}^{-1}(t))-t\right| = \sup_{[0,1]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Theta(t-F_{\underline{\theta}}(X_i))-\sqrt{n}t\right|$$

Now, let's define:

$$\tilde{B}_t^{(n)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Theta(t-F_{\underline{\theta}}(X_i))-\sqrt{n}t$$

The key point is that $F_{\underline{\theta}}(X_1),\dots,F_{\underline{\theta}}(X_n)$ are independent and **uniform** in $(0,1)$, as we have shown in the first chapter. It is immediate to see that:

$$\tilde{B}_0^{(n)} = \tilde{B}_1^{(n)} = 0$$

Moreover:

$$E\left[\tilde{B}_t^{(n)}\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 du\Theta(t-u)-\sqrt{n}t = 0$$

and, if $s<t$:

$$\begin{aligned}
E\left[\tilde{B}_t^{(n)}\tilde{B}_s^{(n)}\right] &= \frac{1}{n}\sum_i\int_0^1 du\Theta(t-u)\Theta(s-v)du+\\
&+\frac{1}{n}\sum_{i\neq j=1}^{n}\int_0^1 du\int_0^1 dv\Theta(t-u)\Theta(s-v)+nts+\\
&-s\sum_{i=1}^{n}\int_0^1 du\Theta(t-u)-t\sum_{i=1}^{n}\int_0^1 dv\Theta(s-v)=\\
&= s+(n-1)ts+nts-nts-nts=\\
&= s(1-t)
\end{aligned}$$

Finally, the central limit theorem guarantees that $\tilde{B}_t$, in the limit $n\to+\infty$ becomes normal. When we will introduce the theory of stochastic processes, we will call the process:

$$\tilde{B}_t = \lim_{n\to+\infty}\tilde{B}_t^{(n)}$$

brownian bridge. We have thus shown that:

$$\lim_{n\to\infty}\sqrt{n}\sup_{\mathbb{R}}|\tilde{F}_n(x)-F_{\underline{\theta}}(x)| = \sup_{[0,1]}|\tilde{B}(t)| \tag{78}$$

where $\tilde{B}(t)$ is a brownian bridge. This is very useful since the following technical result, of which we will omit the proof, holds:

$$P\big(\sup_{[0,1]}|\tilde{B}(t)|\leq x\big) = 1-2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} \tag{79}$$

Table II: quantiles $D_{1-\alpha}$ of the random variable $\sup_{[0,1]} |\tilde{B}(t)|$

$1-\alpha$	$D_{1-\alpha}$
0.99	1.627
0.98	1.518
0.95	1.358
0.90	1.222
0.85	1.138
0.80	1.073

The reader may refer to the following table of the quantiles of the random variable $\sup_{[0,1]} |\tilde{B}(t)|$:

The critical region of the Kolmogorov-Smirnov test is:

$$\Omega_R = \left\{ (x_1 \dots x_n) \, : \, \sqrt{n} \sup_{\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}_0}(x)| > D_{1-\alpha} \right\} \tag{80}$$

Given the data, the tester, assuming that $\underline{\theta} = \underline{\theta}_0$ evaluates the empirical cumulative distribution function $\tilde{F}_n(x)$ and finds the real number $\sqrt{n} \sup_{\mathbb{R}} |\tilde{F}_n(x) - F_{\underline{\theta}_0}(x)|$; if such positive number is bigger than $D_{1-\alpha}$, the tester rejects the hypothesis at the significance level α.

D. Estimators of covariance and correlation

During experiments a very important issue is the existence of correlations among different quantities that are measured. Let's consider the the simplest situation, when only two quantities are measured: this results into two sets of data, $(x_1 \dots x_n)$ and $(y_1 \dots y_n)$, which we view as realizations of two samples $(X_1 \dots X_n)$ e $(Y_1 \dots Y_n)$.

We wish to estimate the covariance:

$$\mathrm{cov}_{\underline{\theta}}(X_i Y_i) = E_{\underline{\theta}}(X_i Y_i) - E_{\underline{\theta}}(X_i) E_{\underline{\theta}}(Y_i) = \mu_{XY}(\underline{\theta}) - \mu_X(\underline{\theta}) \mu_Y(\underline{\theta}) \tag{81}$$

Let's define the estimator:

$$\begin{aligned} \mathcal{C} = \frac{n}{n-1} \mathcal{M}_{XY}(X_1 Y_1 \dots X_n Y_n) - \frac{n}{n-1} \mathcal{M}_X(X_1 \dots X_n) \mathcal{M}_Y(Y_1 \dots Y_n) = \\ = \frac{1}{n-1} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} X_i \right) \left(\sum_{i=1}^{n} Y_i \right) \end{aligned} \tag{82}$$

This is an unbiased estimator for $\mu_{XY}(\underline{\theta}) - \mu_X(\underline{\theta}) \mu_Y(\underline{\theta})$, as can be seen from the following calculation:

$$\begin{aligned} E_{\underline{\theta}}(\mathcal{C}) = \frac{\sum_{i=1}^{n} E_{\underline{\theta}}(X_i Y_i)}{n-1} - \frac{\sum_{ij=1}^{n} E_{\underline{\theta}}(X_i Y_j)}{n(n-1)} = \\ = E_{\underline{\theta}}(X_i Y_i) - E_{\underline{\theta}}(X_i) E_{\underline{\theta}}(Y_i) = \mu_{XY}(\underline{\theta}) - \mu_X(\underline{\theta}) \mu_Y(\underline{\theta}) \end{aligned}$$

Moreover, the law of large numbers guarantees that $\mathcal{C}$ is also consistent.

A very interesting quantitative information about correlation, very often used in data analysis, is the **Pearson correlation coefficient**:

$$\rho(\underline{\theta}) = \frac{\mu_{XY}(\underline{\theta}) - \mu_X(\underline{\theta})\mu_Y(\underline{\theta})}{\sqrt{\sigma_X^2(\underline{\theta})\,\sigma_Y^2(\underline{\theta})}}$$

which is a real number, $-1 \leq \rho(\underline{\theta}) \leq 1$, is zero if the quantities are non correlated and reaches the value ± 1 when there exists a linear relationship between the two quantities.

A typical estimator for $\rho(\underline{\theta})$ is:

$$\mathcal{R} = \frac{\mathcal{M}_{XY} - \mathcal{M}_X\mathcal{M}_Y}{\sqrt{\mathcal{S}_X^2\,\mathcal{S}_Y^2}} \tag{83}$$

This natural estimator is a quite complicated function of the samples: it is highly non trivial to evaluate its expectation or to build up confidence intervals. It is useful to introduce here a well established technique, the propagation of errors, which will help us to study the estimator $\mathcal{R}$. The first observation is that:

$$\mathcal{R} = g\left(\mathcal{M}_X, \mathcal{M}_Y, \mathcal{M}_{X^2}, \mathcal{M}_{Y^2}, \mathcal{M}_{XY}\right)$$

where:

$$g(x_1, x_2, x_3, x_4, x_5) = \frac{x_5 - x_1x_2}{\sqrt{(x_3 - x_1^2)(x_4 - x_2^2)}}$$

and we know the properties of the statistics $\mathcal{M}_X, \mathcal{M}_Y, \mathcal{M}_{X^2}, \mathcal{M}_{Y^2}, \mathcal{M}_{XY}$. What can we learn about $\mathcal{R}$?

The approach we will follow relies on the important theorem:

Teorema 15 (Propagation of errors) *Let $\{Z_n\}_{n=0}^{\infty}$ be a sequence of k dimensional random variables converging almost surely to a constant vector $z \in \mathbb{R}^k$ and such as the sequence:*

$$\frac{Z_n - z}{1/\sqrt{n}} \tag{84}$$

converges in distribution to a normal random variable $N(0, \Sigma)$ for a given matrix Σ.

If $g : \mathbb{R}^k \to \mathbb{R}$ is a function of class C^1 in a neighborhood of z, then the sequence:

$$\frac{g(Z_n) - g(z)}{1/\sqrt{n}} \tag{85}$$

converges in distribution to a normal random variable $N(0, \nabla g(z)\,\Sigma\,\nabla g(z)^T)$.

Dimostrazione. *The proof relies on a first order Taylor expansion with Lagrange rest:*

$$g(Z_n(\omega)) = g(z) + \nabla g(Z_n^\star(\omega))(Z_n(\omega) - z)$$

where $Z_n^\star(\omega)$ lies, for all ω, between z and $Z_n(\omega)$. Exploiting the continuity of ∇g (g is of class C^1 by construction), we have thus:

$$\lim_{n\to\infty} \sqrt{n}\,(g(Z_n) - g(z)) = \nabla g(z) \lim_{n\to\infty} \sqrt{n}\,(Z_n - z)$$

and the thesis follows from the fact that the right hand side has law $N(0, \nabla g(z)\,\Sigma\,\nabla g(z)^T)$.

Now, the sequence $(\mathcal{M}_X, \mathcal{M}_Y, \mathcal{M}_{X^2}, \mathcal{M}_{Y^2}, \mathcal{M}_{XY})$ converges almost surely to $(\mu_X(\underline{\theta}), \mu_Y(\underline{\theta}), \mu_{X^2}(\underline{\theta}), \mu_{Y^2}(\underline{\theta}), \mu_{XY}(\underline{\theta}))$ when the rank of the samples tends to $+\infty$. Since:

$$\mathcal{R} = g\,(\mathcal{M}_X, \mathcal{M}_Y, \mathcal{M}_{X^2}, \mathcal{M}_{Y^2}, \mathcal{M}_{XY})$$

where:

$$g(x_1, x_2, x_3, x_4, x_5) = \frac{x_5 - x_1 x_2}{\sqrt{(x_3 - x_1^2)(x_4 - x_2^2)}}$$

is continuous $(\mu_X(\underline{\theta}), \mu_Y(\underline{\theta}), \mu_{X^2}(\underline{\theta}), \mu_{Y^2}(\underline{\theta}), \mu_{XY}(\underline{\theta}))$, then $\mathcal{R}$ converges almost surely to $\rho(\underline{\theta})$ (the interested reader can try to show this continuous mapping theorem!). This guarantees the consistency of the estimator, since almost sure convergence implies convergence in probability. Moreover, since:

$$\lim_{n\to\infty} E_{\underline{\theta}}(\mathcal{R}) = \rho(\underline{\theta})$$

$\mathcal{R}$ is also an asyntotically unbiased estimator.

If the components (X_i, Y_i) follow a normal law, we can also use the propagation of errors to find the law of $\mathcal{R}$ and to provide confidence intervals for $\rho(\underline{\theta})$. In order to simplify the notations, we let $\mu_X(\underline{\theta}) = 0$ e $\mu_Y(\underline{\theta}) = 0$.

For the central limit theorem, the random variable:

$$\lim_{n\to\infty} \sqrt{n} \left[\begin{pmatrix} \mathcal{M}_X \\ \mathcal{M}_Y \\ \mathcal{M}_{X^2} \\ \mathcal{M}_{Y^2} \\ \mathcal{M}_{XY} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \sigma_X^2(\underline{\theta}) \\ \sigma_Y^2(\underline{\theta}) \\ \rho(\underline{\theta})\sigma_X(\underline{\theta})\sigma_Y(\underline{\theta}) \end{pmatrix} \right]$$

follows a normal law with covariance matrix Σ whose explitit form is:

$$\Sigma = \begin{pmatrix} \sigma_X^2 & 2\rho\sigma_X\sigma_Y & 0 & 0 & 0 \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 & 0 & 0 & 0 \\ 0 & 0 & 2\sigma_X^4 & 2\rho^2\sigma_X^2\sigma_Y^2 & 2\rho\sigma_X^3\sigma_Y \\ 0 & 0 & 2\rho^2\sigma_X^2\sigma_Y^2 & 2\sigma_Y^4 & 2\rho\sigma_Y^3\sigma_X \\ 0 & 0 & 2\rho\sigma_X^3\sigma_Y & 2\rho\sigma_Y^3\sigma_X & (1+\rho^2)\sigma_X^2\sigma_Y^2 \end{pmatrix}$$

Moreover:

$$\nabla g(0, 0, \sigma_X^2(\underline{\theta}), \sigma_Y^2(\underline{\theta}), \rho(\underline{\theta})\sigma_X(\underline{\theta})\sigma_Y(\underline{\theta})) = \begin{pmatrix} 0 \\ 0 \\ -\frac{\rho(\underline{\theta})}{\sigma_X^2(\underline{\theta})} \\ -\frac{\rho(\underline{\theta})}{\sigma_Y^2(\underline{\theta})} \\ \frac{1}{\sigma_X(\underline{\theta})\sigma_Y(\underline{\theta})} \end{pmatrix}$$

Since $\nabla g \Sigma \nabla g = (1-\rho(\underline{\theta})^2)^2$, the propagation of errors theorem guarantees that:

$$\lim_{n\to\infty} \sqrt{n}\,(\mathcal{R} - \rho(\underline{\theta})) \sim N\left(0, \left(1-\rho(\underline{\theta})^2\right)^2\right) \tag{86}$$

The reader can verify that the same result can be obtained also when $\mu_X(\underline{\theta})$, $\mu_Y(\underline{\theta})$ do not vanish.

We have thus:

$$P_\theta\left(-\phi_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\mathcal{R}-\rho(\underline{\theta})}{1-\rho(\underline{\theta})^2} \leq \phi_{1-\frac{\alpha}{2}}\right) \simeq 1-\alpha \tag{87}$$

and a confidence interval at the level $1-\alpha$ for $\rho(\underline{\theta})$ turns out to be:

$$\left[\frac{1-\sqrt{1-4z\mathcal{R}+4z^2}}{2z}, \frac{\sqrt{1+4z\mathcal{R}+4z^2}-1}{2z}\right] \tag{88}$$

where $z = \frac{\phi_{1-\frac{\alpha}{2}}}{\sqrt{n}}$.

E. Linear Regression

It is very common, in several applications, to guess an affine-linear relation between two quantities, say X and Y. The quantity X is usually called the **input variable**, and can be controlled by the experimentalist, who chooses n-values, $(x_1, \ldots, x_n)$ and, correspondingly, performes n measurements of the **response variable**, Y. The response variable is random and the experimentalist will obtain n data $(y_1, \ldots, y_n)$. If we expect a linear-affine relation between X and Y, the simplest way to describe the experiment with the language of probability theory is to model $(y_1, \ldots, y_n)$ as realizations of n random variables of the form:

$$Y_i = a + bx_i + \sigma\varepsilon_i \tag{89}$$

where the x_i enter simply as real parameters, while $\varepsilon_i \sim N(0,1)$ are standard normal independent random variables; the coefficients a and b of the linear-affine relation, and the measurement error σ, are to be inferred from the data. We thus assume that the probability distribution of the error does not depend on the value of the input variable.

We are going now to show how to build up estimations of a, b, and σ, using as starting point the input parameters $(x_1, \ldots, x_n)$ and the data $(y_1, \ldots, y_n)$.

If we organize the data in couples $(x_1, y_1) \ldots (x_n, y_n)$, the most natural strategy is to find the values of a and b minimizing the quantity:

$$F(a,b) = \sum_{i=1}^{n} |y_i - a - bx_i|^2 \tag{90}$$

We can write the above function in a more geometrical way as follows:

$$F(a,b) = \left|\underline{y} - M\begin{pmatrix} a \\ b \end{pmatrix}\right|^2 \tag{91}$$

where $\underline{y} = (y_1, \ldots, y_n)$ and $M \in M_{n\times 2}(\mathbb{R})$ is the matrix:

$$\begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \tag{92}$$

whose columns are linearly independent provided that $(x_1 \dots x_n)$ are not all equals. As (a, b) vary, the set of points $M \begin{pmatrix} a \\ b \end{pmatrix}$ is the plane E_1 in $\mathbb{R}^n$ spanned by the columns of M. Thus:

$$\min_{(a,b)\in\mathbb{R}^2} F(a, b) = \min_{\underline{p}\in E_1} |\underline{y} - \underline{p}|^2 \tag{93}$$

so that elementary geometry implies that the minimum is reached when $\underline{p} = \Pi_1 \underline{y}$, Π_1 being the projector onto the plane E_1, whose explicit form is the following:

$$\Pi_1 = M(M^T M)^{-1} M^T \tag{94}$$

We have thus:

$$\underline{p} = \Pi_1 \underline{y} = M(M^T M)^{-1} M^T \underline{y} \tag{95}$$

which, keeping in mind that $\underline{p} = M \begin{pmatrix} a \\ b \end{pmatrix}$, leads to the estimator:

$$\begin{pmatrix} \mathcal{A} \\ \mathcal{B} \end{pmatrix} (Y_1 \dots Y_n) = (M^T M)^{-1} M^T Y \tag{96}$$

or, more explicitely:

$$\begin{aligned} \mathcal{A}(Y_1 \dots Y_n) &= \frac{\mathcal{M}_{X^2}\mathcal{M}_Y - \mathcal{M}_X \mathcal{M}_{XY}}{\mathcal{M}_{X^2} - \mathcal{M}_X^2} \\ \mathcal{B}(Y_1 \dots Y_n) &= \frac{\mathcal{M}_{XY} - \mathcal{M}_X \mathcal{M}_Y}{\mathcal{M}_{X^2} - \mathcal{M}_X^2} \end{aligned} \tag{97}$$

We stress that the quantities $\mathcal{M}_X$ and $\mathcal{M}_{X^2}$ are **not** random, depending only on the input data. The random variables (97) are unbiased estimators of the parameters (a, b); in fact:

$$E\left[\begin{pmatrix} \mathcal{A} \\ \mathcal{B} \end{pmatrix}\right] = (M^T M)^{-1} M^T E[Y] = (M^T M)^{-1} M^T M \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \tag{98}$$

We have still to build up an estimator for σ^2. The idea is that such parameter determines the discrepancy between the data $(y_1, \dots, y_n)$ and the points of the **regression line** $(a+bx_1, \dots, a+bx_n)$. It is thus natural to interrelate such parameter to the minimum of the function:

$$\min_{\underline{p}\in E_1} |\underline{y} - \underline{p}|^2 = |\Pi_2 \underline{y}|^2 \tag{99}$$

Π_2 being the projector onto the $n-2$-dimensional orthogonal complement of the plane E_1. Such quantity can be interpreted as a realization of the random variable:

$$\mathcal{S}^2(Y_1 \dots Y_n) = \frac{|\Pi_2 Y|^2}{n-2} = \frac{\sum_i |Y_i - \mathcal{A} - \mathcal{B}X_i|^2}{n-2} \tag{100}$$

which is an unbiased estimator for σ^2 since:

$$E\left[|\Pi_2 Y|^2\right] = \sigma^2 E\left[|\Pi_2 \epsilon|^2\right] = \sigma^2 (n-2) \tag{101}$$

We can also estimate confidence intervals for the parameters a, b, σ^2 relying on the following result:

Proposizione 16 *The random variables $\mathcal{A}$ and $\mathcal{B}$ are independent from $\mathcal{S}^2$. Moreover:*

$$\begin{aligned} \frac{\mathcal{S}^2}{\sigma^2} &\sim \frac{\chi^2(n-2)}{n-2} \\ \frac{\mathcal{A}-a}{\sqrt{m_a}\,\mathcal{S}} &\sim t(n-2) \\ \frac{\mathcal{B}-b}{\sqrt{m_b}\,\mathcal{S}} &\sim t(n-2) \end{aligned} \tag{102}$$

where:

$$m_a = \frac{\mathcal{M}_{X^2}}{n(\mathcal{M}_{X^2} - \mathcal{M}_X^2)}, \quad m_b = \frac{1}{n(\mathcal{M}_{X^2} - \mathcal{M}_X^2)} \tag{103}$$

Proof. *Since the random variable ϵ and the linear projectors Π_1, Π_2 satisfy the hyphothesis of Cochran theorem, the random variables $\Pi_1\epsilon$ and $\Pi_2\epsilon$ are independent. Moreover, Cochran theorem guarantees that:*

$$|\Pi_2 \epsilon|^2 \sim \chi^2(n-2) \tag{104}$$

since the subspace onto which Π_2 projects has dimension $n-2$. Thus $\frac{\mathcal{S}^2}{\sigma^2}\frac{|\Pi_2\epsilon|^2}{n-2} \sim \frac{\chi^2(n-2)}{n-2}$. Since the covariance matrix of the random variable $(M^TM)^{-1}M^T\underline{Y}$ is:

$$cov[(M^TM)^{-1}M^TY] = (M^TM)^{-1}M^T cov[Y] M (M^TM)^{-1} = \sigma^2 (M^TM)^{-1} \tag{105}$$

we conclude that $\mathcal{A} \sim N(a, m_a\,\sigma^2)$, $\mathcal{B} \sim N(b, m_b\,\sigma^2)$ where $m_a = \left[(M^TM)^{-1}\right]_{11} = \frac{\mathcal{M}_{X^2}}{n(\mathcal{M}_{X^2} - \mathcal{M}_X^2)}$ and $m_b = \left[(M^TM)^{-1}\right]_{22} = \frac{1}{n(\mathcal{M}_{X^2} - \mathcal{M}_X^2)}$. The definition of the Student law together with the independence of $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{S}$ completes the proof.

We are now able to provide confidence intervals of level $1-\alpha$ for σ^2:

$$\left[\frac{n-2}{\chi^2_{1-\frac{\alpha}{2}}(n-2)}\,\mathcal{S}^2, \frac{n-2}{\chi^2_{\frac{\alpha}{2}}(n-2)}\,\mathcal{S}^2\right] \tag{106}$$

and for a and b:

$$\begin{aligned} &\left[\mathcal{A} - \sqrt{m_a}\,\mathcal{S}\,t_{1-\frac{\alpha}{2}}(n-2), \mathcal{A} + \sqrt{m_a}\,\mathcal{S}\,t_{1-\frac{\alpha}{2}}(n-2)\right] \\ &\left[\mathcal{B} - \sqrt{m_b}\,\mathcal{S}\,t_{1-\frac{\alpha}{2}}(n-2), \mathcal{B} + \sqrt{m_b}\,\mathcal{S}\,t_{1-\frac{\alpha}{2}}(n-2)\right] \end{aligned} \tag{107}$$