Introduction to Statistics

Michele Gianfelice
Department of Mathematics
University of Calabria
Ponte Pietro Bucci - Cubo 30B
I-87036 Arcavacata di Rende (CS)
gianfelice@mat.unical.it

February the 3^{rd}, 2011

synopsis

- **Descriptive Statistics**
  - Univariate Data Samples
    - Modal Value and Sample Mode
    - Sample Median
    - Sample Mean
    - Sample Variance and Sample Standard Deviation
    - Linear Transformations
  - Bivariate Samples
    - Sample Covariance and Sample Correlation Coefficient
- **Introduction to Inferential Statistics**
  - Random Variables
  - Outcome of an experiment as a random variable
  - The Maximum Likelihood method
    - Univariate Linear Regression
    - The Weighted Mean
  - Introduction to parametric Hypothesis Tests
    - The Pearson χ^2 Test
    - The t-Test
    - Test on the Variance of Gaussian Sample

How to organize a data sample in order to extract qualitative information about the performed experiment?
Let X be an experimental quantity (e.g. a physical quantity such as the lenght or the weight of an object, but also the rate of growth of a bacterial population). We denote by x the real number representing the outcome of the mesurement procedure of X.
Repeting $N \ (N \in \mathbb{N})$ times the experiment whose outcome is a measured value of X, being careful, every single time, to replicate the exact same experimental conditions, we obtain

$$\mathcal{C}_N(X) := \{x_1, .., x_N\}$$

a collection (*sample*) of size N of measured values of X (*data*).

Example

X is the lifetime (in days) of laboratory animal exposed to a pathogen. $N = 36$ is the number of animals so that the sample $\mathcal{C}_{36}(X)$ is the collection of positive integers given in the following table.

$$\begin{array}{ccccccccc} 82 & 89 & 94 & 110 & 74 & 122 & 112 & 95 & 100 \\ 78 & 65 & 60 & 90 & 83 & 87 & 75 & 114 & 85 \\ 69 & 94 & 124 & 115 & 107 & 88 & 97 & 74 & 72 \\ 68 & 83 & 91 & 90 & 102 & 77 & 125 & 108 & 65 \end{array} . \tag{1}$$

Replicating the same experiment we expect to obtain the same outcome, that is the same measured value for X. Since this is does not always happen, relevant information on the performed experiment can be gained looking at the set of values assumed by the mesures of X, i.e. the values of the sample data.

We can partition the data sample into disjoint subsets representing the the collection of data having the same values $\mathcal{C}_N(X) = \bigvee_{y\in\mathbb{R}} \mathcal{F}_N^y(X)$, where $\mathcal{F}_N^y(X) := \{x \in \mathcal{C}_N(X) : x = y\}$.

Definition

$F_y := \left|\mathcal{F}_N^y(X)\right|$ the number of data having value y is said *(absolute) frequency of* y.

To see which value of the data is more frequent (*typical*) we can plot the graph of the function

$$\mathbb{R} \ni y \longmapsto F_y \in \{0, 1, .., N\}$$

which is called *frequencies histogram.*
The frequencies histogram depends on the size of the data sample N.

Definition

$f_y := \frac{F_y}{N}$ is said *relative frequency of* y.

To compare the amount of information on the phenomenon under investigation given by data sets of different sizes it is useful to plot the function

$$\mathbb{R} \ni y \longmapsto f_y \in [0, 1] \quad ,$$

which is called *relative frequencies histogram* and which is independent of the sample size.

Exercise

plot the relative frequencies histogram of the data sample (1).

When the values assumed by the sample data spread out over an interval or over $\mathbb{R}$ and the sample size is large, relative frequencies histograms doesn't give back good information on the distribution of the measured values of X. Therefore, it is more convenient to partition the data sample into classes collecting data whose values range in an interval rather than those assuming a single value.

To do this let $\{a_k\}_{k\in\mathbb{Z}}$ to be an increasing sequence of real numbers and set $\mathbb{R} = \bigcup\limits_{k\in\mathbb{Z}} (a_k, a_{k+1}]$.

Then $\mathcal{C}_N(X) = \bigvee_{k\in\mathbb{Z}} \mathcal{B}_N^k(X)$, where

$$\mathcal{B}_N^k(X) := \{y \in \mathcal{C}_N(X) : y \in (a_k, a_{k+1}]\} \ ,$$

set $f_k := \frac{\left|\mathcal{B}_N^k(X)\right|}{N}$ and plot the function

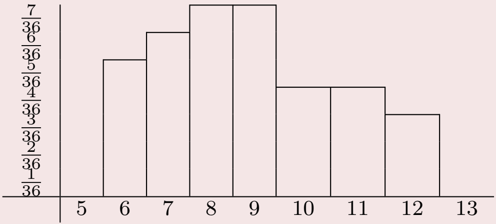
$$\mathbb{R} \ni x \longmapsto \phi(x) := \sum_{k\in\mathbb{Z}} f_k \mathbf{1}_{(a_k, a_{k+1}]}(x) \in [0,1]$$

where $\mathbf{1}_A$ denotes the indicator function of the set $A \subset \mathbb{R}$, which represents the relative frequencies histogram of the data belonging to the subsamples realizing the partition $\mathcal{C}_N(X)$.

Exercise

Let $\forall k \in \mathbb{Z},\ a_k = 10k$. Plot ϕ for the sample data (1) and compare this plot with the one of the previous exercise. Did we gain any information?

Solution



.

Subsample partition

The general criterion to construct a meaningful partiton of $\mathcal{C}_N(X)$ is to choose the sequence $\{a_k\}_{k\in\mathbb{Z}}$ realizing the partition of $\mathbb{R}$ in such a way that:

- $a_{k+1} - a_k$ is independent of k;
    - the distribution of the relative frequencies of data belonging to a given subsample $\mathcal{B}_N^k(X)$ is nearly homogeneous, that is if $x, y \in (a_k, a_{k+1}]$, then $f_x \simeq f_y$,
    - the distribution of the relative frequencies of data belonging to different subsamples are different.

Definition

Statistics are numerical quantities, computed from the data values, summarizing the information which can be extracted by the data sample.

Most commonly used statistics are:

Modal Values and Sample Mode

The *modal values* are those values of the data that occur at the highest frequency. If there is only one of such values this is called *sample mode.*

In general multi modal histograms are generated by data samples being the union of two or more subsamples, each of which give rise to a unimodal frequencies histograms.

Sample Median

Rearrange the data sample in increasing order. Denoting by $\widehat{\mathcal{C}}_N(X) = \{\hat{x}_1, .., \hat{x}_N\}$ the rearranged sample, the *sample median* is so defined

$$\hat{x} := \begin{cases} \frac{\left(\hat{x}_{\frac{N}{2}} + \hat{x}_{\frac{N}{2}+1}\right)}{2} & if\, N\, is\, even \\ \hat{x}_{\frac{N+1}{2}} & if\, N\, is\, odd \end{cases} .$$

Hence half of the data lie to left of $\hat{x}$ and half to the right.

Sample Mean

The *sample mean* is so defined

$$\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i \ .$$

Notice that, if

$$\bar{x}_k := \frac{1}{\left|\mathcal{B}_N^k\left(X\right)\right|} \sum_{i=1,..,N \,:\, x_i \in \mathcal{B}_N^k(X)} x_i$$

is the sample mean of the class $\mathcal{B}_N^k\left(X\right)$,

$$\bar{x} = \sum_{k\in\mathbb{Z}} f_k \bar{x}_k$$

which can be seen as the center of mass of the relative frequencies histogram of the data belonging to the $\mathcal{B}_N^k\left(X\right)$'s.

Sample Variance and Sample Standard Deviation

The *sample variance* is so defined

$$s_X^2 := \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$$
$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i^2-\bar{x}^2\right)$$

and the quantity $s_X := \sqrt{s_X^2}$ is called *sample standard deviation.*

Notice that S_X is a measure of the deviation of the data values from the sample mean. As a matter of fact,

$$x_i = \bar{x} + (x_i - \bar{x}) \ , \quad i = 1, .., N \, .$$

Since $\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x}) = 0$, it would be useful to compute the average distance of the data values from the sample mean $\frac{1}{N}\sum_{i=1}^{N}|x_i-\bar{x}|$ but,

$$\frac{1}{N}\sum_{i=1}^{N}|x_i-\bar{x}| = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i-\bar{x})^2} \leq \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} \leq s_X \, .$$

so s_X overestimates $\frac{1}{N}\sum_{i=1}^{N}|x_i-\bar{x}|$.

Linear transformations

Linear relations among experimental quantities play a fundamental role in data analysis, also because exponential and power-law functional dependence between two of such quantities can be reduced to linear relations.

If A, B are known constants and X, Y are experimental quantities,

$$Y = Ae^{BX} \Longrightarrow \log Y = BX + \log A$$

$$Y = AX^{B} \Longrightarrow \log Y = B \log X + \log A$$

Therefore, if $Y = AX + B$, with A, B known constants, we have:

- if $\{\tilde{x}_i\}_{i=1,..,K}$, $1 \leq K \leq N$, are the modal values of $\mathcal{C}_N\left(X\right)$, then $\{\tilde{y}_i\}_{i=1,..,K}$, where $\tilde{y}_i = A\tilde{x}_i + B$, are the modal values of $\mathcal{C}_N\left(Y\right)$;
- $\hat{y} = A\hat{x} + B$;
- $\bar{y} = A\bar{x} + B$;
- $s_Y^2 = A^2 s_X$, $s_Y = |A|\, s_X$.

Bivariate samples

Suppose we perform an experiment which allow us to measure two quantities X and Y.
A sample of N measured values of X and Y denoted by

$$\mathcal{C}_N\left(X,Y\right) := \left\{\left(x_1,y_1\right),..,\left(x_N,y_N\right)\right\}$$

and $\mathcal{C}_N\left(X\right), \mathcal{C}_N\left(Y\right)$ denote the univariate associated data samples called *marginal data samples.* The Cartesian plot of $\mathcal{C}_N\left(X,Y\right)$ is called *scatter diagram* and gives us a qualitative criterion to see if there is a functional dependence between X and Y.

Sample Covariance and Sample Correlation Coefficient

A quantitative measure of the relationship between two experimental quantities X and Y are the statistics *sample covariance*

$$s_{X,Y} := \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$$

and *sample correlation coefficient*

$$r_{X,Y} := \frac{s_{X,Y}}{s_X s_Y} = \frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}}\ .$$

If $Z = AX + B, W = CY + D$, with A, B, C, D constants indipendent of the data of $\mathcal{C}_N(X, Y)$,

$$s_{Z,W} = AC s_{X,Y} \ ,$$
$$r_{Z,W} = \frac{AC}{|A|\,|C|} r_{X,Y} \ .$$

Hence,

- $r_{X,Y}$ is a pure number i.e. its value does not depend on the mesurement sytem,
- $\left|r_{X,Y}\right| \leq 1$ and $\left|r_{X,Y}\right| = 1 \Longleftrightarrow Y = AX + B$, in particular

$$r_{X,Y} = \begin{cases} +1 & if \ \ A > 0 \\ -1 & if \ \ A < 0 \end{cases} \quad .$$

In this last case the parameters A and B are called *regression parameters*.

Introduction to Inferential Statistics

Each time we perform an experiment and mesure X, despite the fact that we start with the same experimental condition we produce a different measured value of X. How to deal with this fact? We can assume the outcomes of an experiment to be that of a random variable.
Given a set Ω, let $\mathcal{F}$ be a σalgebra of subsets of Ω, that is a collection of subsets of Ω such that:

- $\Omega \in \mathcal{F}$;
- if $A \in \mathcal{F}$, then $A^c := \Omega \backslash A \in \mathcal{F}$;
- any finite or denumerable union of elements of $\mathcal{F}$ is in $\mathcal{F}$.

The couple $(\Omega, \mathcal{F})$ is called *mesurable space.* A probability measure on $(\Omega, \mathcal{F})$ is a non-negative function

$$\mathcal{F} \ni A \longmapsto \mathbb{P}(A) \in [0,1]$$

such that if $\{A_i\}_{i \in \mathbb{N}} \subset \mathcal{F}$ is a collection of mutually disjoint subsets of Ω, that is $\forall i, j \in \mathbb{N},\ i \neq j,\ A_i \cap A_j = \varnothing,\ \mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$.
The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called *probability space.* A *random variable (r.v.)* is a map $\xi : \Omega \longrightarrow \mathbb{R}$ such that, if $\mathcal{B}(\mathbb{R})$ denotes the σalgebra generated by the open subsets of $\mathbb{R}$, with respect to the Euclidean topology, $\forall B \in \mathcal{B}(\mathbb{R}),\ \xi^{-1}(B) \in \mathcal{F}$. Hence, ξ maps a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ to the probability measure $\mathbb{P}_\xi$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\forall B \in \mathcal{B}(\mathbb{R}),\ \mathbb{P}_\xi(B) = \mathbb{P}\left(\xi^{-1}(B)\right)$.

The function,

$$\mathbb{R} \ni x \longmapsto F_{\xi}\left(x\right) := \mathbb{P}\left\{\omega \in \Omega : \xi\left(\omega\right) \leq x\right\} \in [0,1]$$

is called *distribution function* of the r.v. ξ and has the following properties:

- F_{ξ} is non-decreasing;
- $\lim_{x\downarrow-\infty} F_{\xi}\left(x\right) = 0$, $\lim_{x\uparrow+\infty} F_{\xi}\left(x\right) = 1$;
- F_{ξ} is right-continuous and has left limits. Hence so F_{ξ} has at most jump discontinuities and the collection of points at which F_{ξ} is discontinuous is at most denumerable.

Notice that $\forall a, b \in \mathbb{R}$ such that $a < b$,

$$\mathbb{P}_{\xi}\left(a,b\right) = \mathbb{P}_{\xi}(a,b] = F_{\xi}\left(b\right) - F_{\xi}\left(a\right)$$

and $\forall x \in \mathbb{R}$,

$$\mathbb{P}_{\xi}\left(x\right) = \lim_{\varepsilon\downarrow 0}\left(F_{\xi}\left(x\right) - F_{\xi}\left(x-\varepsilon\right)\right) = F_{\xi}\left(x\right) - F_{\xi}\left(x^{-}\right) \ .$$

Definition

A r.v. ξ is said to be:

- *absolutely continuous with respect to the Lebesgue measure (a.c.)*, if there exists a positive function f_{ξ}, called *probability density of* ξ, such that $\forall x \in \mathbb{R}$,

$$F_{\xi}\left(x\right) = \int_{-\infty}^{x} dy f_{\xi}\left(y\right) \ ;$$

- *discrete*, if F_{ξ} is a step function. In this case, denoting by $\mathcal{S}$ the set of jump points of F_{ξ} and by

$$p_s := \mathbb{P}\left\{\omega \in \Omega : \xi\left(\omega\right) = s\right\} \ , \ s \in \mathcal{S} \ ,$$

  we have

$$F_{\xi}\left(x\right) = \sum_{y \in \mathcal{S} \ : \ y \leq x} p_s \ .$$

Given a r.v. ξ, $\forall k \in \mathbb{N}$, the quantity

$$\mathbb{E}\left(\xi^k\right) := \begin{cases} \int_{\mathbb{R}} dx f_\xi\left(x\right) x^k & if\ \xi\ is\ a.c. \\ \sum_{x\in\mathcal{S}} x^k \mathbb{P}_\xi\left(x\right) & if\ \xi\ is\ discrete \end{cases} .$$

is called *moment of* ξ *of order* k. Moreover, the moment of ξ of order 1 is called *expectation value of* ξ, while the quantities

$$Var\left(\xi\right) := \mathbb{E}\left[\left(\xi - \mathbb{E}\left(\xi\right)\right)^2\right] = \mathbb{E}\left(\xi^2\right) - \mathbb{E}\left(\xi^2\right) \geq 0$$

and $\sqrt{Var\left(\xi\right)}$ are called respectively *variance of* ξ and *standard deviation of* ξ.

Example

(Gaussian and Normal r.v's) ξ is a Gaussian r.v. of parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ $\left(\xi \in N\left(\mu, \sigma\right)\right)$, if

$$f_\xi\left(x\right) = g\left(x|\mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

Notice that $\mathbb{E}\left(\xi\right) = \mu, Var\left(\xi\right) = \sigma^2$. If $\mu = 0$ and $\sigma = 1$, ξ is called *normal* r.v..

Example

(Bernoulli r.v's) ξ is a Bernoulli r.v. of parameter $p \in (0,1)$ $(\xi \in Ber(p))$ if it is discete and $|\mathcal{S}| = 2$. Therefore, setting $\mathcal{S} = \{s_1, s_2\}$, with $s_1 < s_2$, and $p_{s_2} = p$,

$$F_\xi(x) = (1-p)\,\mathbf{1}_{(s_1,s_2]}(x) + \mathbf{1}_{(s_2,+\infty)}(x)\ .$$

Notice that if $s_1 = 0, s_2 = 1,\ \mathbb{E}(\xi) = p$ and $Var(\xi) = p(1-p)$.

Example

(Poisson r.v's) ξ is a Poisson r.v. of parameter $\lambda > 0$ $(\xi \in Poi(\lambda))$ if

$$F_\xi(x) = e^{-\lambda} \sum_{k \in \mathbb{N}\ :\ 0 \leq k \leq x} \frac{\lambda^k}{k!}\ .$$

Notice that $\mathbb{E}(\xi) = Var(\xi) = \lambda$.

Remark

Given a r.v. ξ such that $\mathbb{E}(\xi), \mathbb{E}(\xi^2) < \infty$, the r.v. $\tilde{\xi} := \frac{\xi - \mathbb{E}(\xi)}{\sqrt{Var(\xi)}}$ has expectation value zero and unitary variance. Hence, in the Gaussian case, if $\xi \in N(\mu, \sigma)$ then $\tilde{\xi} \in N(0,1)$.

Definition

A map $\xi : \Omega \longrightarrow \mathbb{R}^n, \ n \geq 2$, such that $\forall B \in \mathcal{B}(\mathbb{R}^n) := \bigotimes_{i=1}^{n} \mathcal{B}(\mathbb{R}), \ \xi^{-1}(B) \in \mathcal{F}$, is called *random vector*. In other words, ξ is a vector in $\mathbb{R}^n$ whose components are the r.v's $\xi_1, .., \xi_n$. The function

$$\mathbb{R}^n \ni (x_1, .., x_n) \longmapsto F_\xi(x_1, .., x_n) := \mathbb{P}\{\omega \in \Omega : \xi_1(\omega) \leq x_1, .., \xi_1(\omega) \leq x_n\} \in [0, 1]$$

is called *distribution function of* ξ.

Definition

Two r.v's ξ, η, are said to be *(stochastically) independent* if considering the random vector $\zeta = (\xi, \eta), \ \forall (x, y) \in \mathbb{R}^2$, we have

$$F_\zeta(x, y) = F_\xi(x) F_\eta(y) \ .$$

Given collection of r.v's $\{\xi_1, .., \xi_n\}$ this is said to be composed by (stochastically) independent elements if the vector $\xi = (\xi_1, .., \xi_n)$ has (stochastically) independent components.

Definition

A sequence $\{\xi_n\}_{n \in \mathbb{N}}$ of r.v's is said to *converge in distribution* to a r.v. ξ if the sequence of functions $\{F_n\}_{n \in \mathbb{N}}$, with $F_n := F_{\xi_n}$, converges to F_ξ at any point of continuity of F_ξ.

Outcome of an experiment as a random variable

We can model the outcome of our experiment as the possible realization of a random variable ξ^X. Hence a data sample $\mathcal{C}_N(X)$ represents a possible realization of the collection of N i.i.d.r.v's $\{\xi_i^X\}_{i=1}^N$.

Which is the probability distribution of ξ^X?

Usually this is a priori not known, anyway we can make use of the following general results:

Theorem

(Law of Large Numbers) Let $\{\xi_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d.r.v. such that $\mathbb{E}(\xi_1)<\infty$. Then,

$$\mathbb{P}\left\{\omega\in\Omega : \lim_{N\to\infty}\left|\frac{1}{N}\sum_{i=1}^{N}\xi_i(\omega)-\mathbb{E}(\xi_1)\right|\neq 0\right\}=0\ .$$

Theorem

(Central Limit Theorem) Let $\{\xi_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d.r.v. such that $\mathbb{E}(\xi_1)$ and $\mathbb{E}(\xi_1^2)<\infty$. The sequence of r.v. $\{\eta_N\}_{N\in\mathbb{N}}$, where $\eta_N:=\sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\xi}_i\right]$, converges in distribution to a standard normal r.v..

Let us set $\mathbb{P}_X := \mathbb{P}_{\xi^X}$. $\forall N \in \mathbb{N}$, let $\left\{a_k^N\right\}_{k\in\mathbb{Z}}$ to be an increasing sequence of real numbers such that

- $\forall k \in \mathbb{Z},\ a_{k+1}^{(N)} - a_k^{(N)} = \Delta_N$ is independent of k;
- $\mathbb{R} = \bigcup_{k\in\mathbb{Z}} (a_k^N, a_{k+1}^N]$;
- $\left\{\Delta_N\right\}_{N\in\mathbb{N}}$ is a decreasing sequence tending to 0.

$\forall k \in \mathbb{Z}$, we define the r.v.

$$\Omega \ni \omega \longmapsto \bar{\varphi}_k^{(N)}(X)(\omega) := \frac{\sum_{i=1}^{N} \mathbf{1}_{\left(a_k^{(N)}, a_{k+1}^{(N)}\right]}\left(\xi_i^X(\omega)\right)}{N\Delta_N} \in \mathbb{R}$$

representing, $\forall\omega \in \ \Omega$, the relative frequency of the associated realization of the data subsample $\mathcal{B}_N^k(X) \subset \mathcal{C}_N(X)$ divided by the diameter of the partition Δ_N. Moreover,

$$\mathbb{E}\bar{\varphi}_k^{(N)}(X) = \frac{\mathbb{P}_X\left(a_k^{(N)}, a_{k+1}^{(N)}\right]}{\Delta_N}\ .$$

Let us then consider the map

$$\Omega\times\mathbb{R}\ni(\omega,x)\longmapsto\bar{\phi}_X^{(N)}(x;\omega):=\sum_{k\in\mathbb{Z}}\bar{\varphi}_k^{(N)}(X)(\omega)\,\mathbf{1}_{\left(a_k^{(N)},a_{k+1}^{(N)}\right]}(x)\in\mathbb{R}\ ,$$

which:

- $\forall\omega\in\Omega,\ \bar{\phi}_X^{(N)}(\cdot;\omega)$ is a real function;
- $\forall x\in\mathbb{R},\ \bar{\phi}_X^{(N)}(x;\cdot)$ is a r.v.

and notice that the r.v. defined by the series

$$\Omega\ni\omega\longmapsto\sum_{k\in\mathbb{Z}}\bar{\varphi}_k^{(N)}(X)(\omega)\,\Delta_N\in\mathbb{R}$$

is nothing else but the integral over $\mathbb{R}$ of $\bar{\phi}_X^{(N)}$ whose expectation value is,by definition, equal to 1. Therefore, we can define the map

$$\Omega\times\mathbb{R}\ni(\omega,x)\longmapsto\bar{F}_X^{(N)}(x;\omega):=\int_{-\infty}^{x}dy\bar{\phi}_X^{(N)}(y;\omega)\in[0,1]\ \ ,$$

which is called *empirical partition function* and has the following properties:

- $\forall\omega\in\ \Omega,\ \bar{F}_X^{(N)}(\cdot;\omega)$ represents the relative frequency of the data of the associated realization of the sample $\mathcal{C}_N(X)$ falling in $(-\infty,x]$, i.e. $\bar{F}_X^{(N)}(\cdot;\omega)$ is the distribution function of a r.v. which we denote by $\xi_X^{(N)}$;
- $\forall x\in\mathbb{R},\ \bar{F}_X^{(N)}(x;\cdot)$ is a r.v..

Making use of the **Law of Large Numbers**, the sequence of functions $\{\bar{F}_X^{(N)}\}$ converges pointwise to the distribution function $F_X(x)$ of ξ_X. Therefore, we can assume that, if the size N of the data sample $\mathcal{C}_N(X)$ is very large, the relative frequency of the data falling into a given interval $[a,b] \subseteq \mathbb{R}$ of possible values of the measures of X, called *empirical relative frequency of* $[a,b]$, is

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[a,b]}\left(\xi_i^X\right) = \mathbb{P}_X(a,b] + \varepsilon_N \ ,$$

where the error in the aproximation ε_N is such that $\lim_{N\uparrow\infty} \varepsilon_N = 0$.
Now we know how to recover operationally the probability distribution of ξ^X, that is of the measured values of X, even if it is unknown. Unfortunately, usually we do not possess enough data to minimize the error we make in approximating such probability distribution with the empirical relative frequency.
On the other hand, by the **Central Limit Theorem**, we know that the sequence $\left\{\eta_N^X\right\}_{N\in\mathbb{N}}$, where $\eta_N^X = \sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\xi}_i^X\right]$, converges in distribution to a normal r.v. even if we have no knowledge of the distribution of ξ^X. Therefore, in the following, we stick to the Gaussian case in which the distribution of ξ^X is completely specified by the parameters μ and σ.

Statistics

In the framework of Inferential Statistics the word *statistic* refers to r.v's which are functions of the collection of i.i.d.r.v's $\{\xi_i^X\}_{i=1}^N$ whose outcomes are all the possible data sample $\mathcal{C}_N(X)$ we can produce performing our experiment.

1^{st} Consideration

In general, the distribution function of ξ^X, that is of the measured values of X, is unknown, but in some case we can guess its functional form as a function of some unknown parameters. Therefore, in this case, the problem of the identification of the probability distribution of the measures of X is reduced to the identification of the parameters specifying the distribution of ξ^X. For example, in the Gaussian case, the distribution is completely specified by the expectation value μ and by the variance σ.

2^{nd} Consideration

Let F_θ be the distibution function of ξ^X. Since we have a finite amount of data, we can estimate the values of these parameters in such a way that the large is the sample the more accurate is our estimation. To do this, we can define certain statistics $t\left(\xi_1^X, .., \xi_N^X\right)$, called *estimator* of the parameter θ, whose expectation value is exactly the value of the parameter θ we need to know to specify completely F_θ and whose probability distribution will concentrate on θ.

The Maximum Likelihood method

Suppose our data sample $\mathcal{C}_N(X)$ is the outcome of a collection of i.i.d. r.v's $N(\mu,\sigma)$. Then if $\mathcal{C}_N(X)=\{x_1,..,x_N\}$,

$$\mathbb{P}_X(\mathcal{C}_N(X))=\prod_{i=1}^{N} g(x_i|\mu,\sigma)\,dx_i\ .$$

Since μ and σ are not known, we can estimate them choosing those functions of the data that maximizing the quantity $\prod_{i=1}^{N} g(x_i|\mu,\sigma)$ called *likelyhood function of* (μ,σ), or, which is the same, minimizing its logarithm.

This method is particularly useful in the case we are given a bivariate data sample $\mathcal{C}_N(X,Y)$ such that the absolute value of the associated correlation coefficient is close to 1 and hence we want to estimate from the data of the sample the regression parameters linking the experimental quantities X and Y.

Given a bivariate data sample $\mathcal{C}_N(X,Y)$, without loss of generality, suppose X is modeld by a Gaussian r.v. of variance σ^2 and $Y=AX+B$. The likelyhood function is then

$$\prod_{i=1}^{N} g(y_i|\mu_i,\sigma)=\prod_{i=1}^{N} g(y_i|Ax_i+B,\sigma)=\frac{e^{-\sum_{i=1}^{N}\frac{[y_i-(Ax_i+B)]^2}{2\sigma^2}}}{(2\pi)^{\frac{N}{2}}\sigma^N}\ .$$

remark

Notice that maximizing $\prod_{i=1}^{N} g\left(y_i|\mu_i,\sigma\right)$ is equivalent to minimize

$$\sum_{i=1}^{N} \frac{\left[y_i-(Ax_i+B)\right]^2}{2\sigma^2} \ ,$$

this is why this analysis in the Gaussian case is also called *Least Squares Method.*

Therefore, the estimated values for A and B are

$$\begin{cases} \check{A}\left(x_1,..,x_N;y_1,..y_N\right)=\frac{s_{X,Y}}{s_X^2} \\ \check{B}\left(x_1,..,x_N;y_1,..y_N\right)=\bar{y}-\frac{s_{X,Y}}{s_X^2}\bar{x} \end{cases} \quad .$$

Exercise

Analyse the bivariate data sample

$$\begin{array}{cc} (51,74) & (68,70) \\ (97,93) & (55,67) \\ (95,99) & (74,73) \\ (20,33) & (91,91) \\ (74,80) & (80,86) \end{array} \quad .$$

The weighted mean

Suppose M research groups perform the same experiment producing each a data sample $\mathcal{C}_{N_i}(X)\,,\ i=1,..,M$ of Gaussian r.v's. It can be proven (see what stated in the following section) that the sample mean is the best estimator of the value of X. How to deal with the fact that we have M of such values? We can consider the collection of these sample mean $\{\bar{x}_1,..\bar{x}_M\}$ as a data sample $\mathcal{C}_M(X)$ and apply the Maximum Likelihood Method under the hypothesis that the variance of the Gaussian r.v. ξ_i modeling the outcome of the i-th data sample $\mathcal{C}_{N_i}(X)$ can be approximated by the sample variance of the $\mathcal{C}_{N_i}(X)$'s. Hence,

$$\prod_{i=1}^{M} g\left(\bar{x}_i|\mu,s_{X,i}\right)=\prod_{i=1}^{M}\frac{e^{-\frac{(\bar{x}_i-\mu)^2}{2s_{X,i}^2}}}{\sqrt{2\pi}s_{X,i}}=\frac{e^{-\sum_{i=1}^{M}\frac{(\bar{x}_i-\mu)^2}{2s_{X,i}^2}}}{(2\pi)^{\frac{M}{2}}\prod_{i=1}^{M}s_{X,i}}\ ,$$

which gives back as best estimate of X

$$\breve{x}=\frac{\sum_{i=1}^{M}\frac{1}{s_{X,i}^2}\bar{x}_i}{\sum_{i=1}^{M}\frac{1}{s_{X,i}^2}}\ .$$

Remark

Notice that the smaller is the sample variance of a data sample of our collection, i.e. the sharper is the probability distribution of the ξ_i's, the larger is the weight of the sample mean in the convex combination $\breve{x}$.

Introduction to parametric hypothesis tests

Suppose the outcome of the measurement process of an experimental quantity X is modeled by a r.v. ξ^X with distribution function F_θ.
We take into account, together with the two previous consideration,

3^{rd} Consideration

We can formulate an hypothesis on the value of the parameter θ which usually in Statistics is called *null Hypothesis* and denoted by H_0. Then, to test if our data sample verify or not the statement constituting the null hypothesis we will identify a subset of possible outcomes of $\{\xi_i^X\}_{i=1}^N$ called *critical region* such that such that if the data values fall in this region the hypothesis is rejected, otherwise is accepted.

Remark

It is important to note when developing a procedure for testing a given null hypothesis that two different types of errors can result. The first of these, called a *type I error*, is said to result if the test incorrectly calls for rejecting H_0 when it is indeed correct. The second, called a *type II error*, results if the test calls for accepting H_0 when it is false.

Since the goal of the statistical testing is not to explicitly determine whether or not H_0 is true but rather to determine if the data sample is consistent with its validity, the classical approach to testing the null hypothesis is to fix a value α, called *significance level*, and then require that the test have the property that the probability of a type I error occurring can never be greater than α. Therefore, the critical region to the test the null hypothesis $H_0 : \theta \in W \subset \mathbb{R}$ will be then identified with the subset of possible outcomes of $\left\{\xi_i^X\right\}_{i=1}^N$ for which the probability of the resulting value of an estimator of θ to fall outside W is smaller than or equal to the significance level α.

The Gaussian case

To deal with Gaussian samples we need to introduce other two probability distributions.

Definition

Let $\{\xi_i\}_{i=1}^{n}$ be a collection of $N(0,1)$ i.i.d. r.v's. The probability distribution of the r.v.

$$\zeta := \sum_{i=1}^{n} \xi_n^2$$

is called χ^2*-distribution with* n *degrees of freedom* $\left(\zeta \in \chi_n^2\right)$.

Definition

Let $\zeta \in \chi_n^2$ and $\xi \in N(0,1)$ be stochastically independent r.v's. The probability distribution of the r.v.

$$\eta := \frac{\xi}{\sqrt{\zeta}}$$

is called *(Student)* t*-distribution with* n *degrees of freedom* $(\eta \in T_n)$.

Let our data sample $\mathcal{C}_N(X)$ be the outcome of a collection $\{\xi_i^X\}_{i=1}^N$ of $N(\mu,\sigma)$ i.i.d.r.v's. The statistics we will consider are:

- the *empirical mean*

$$\bar{\xi}_N^X := \frac{1}{N}\sum_{i=1}^N \xi_i^X$$

  whose possible outome is the sample mean and whose probability distribution is Gaussian with parameter μ and $\frac{\sigma}{\sqrt{N}}$;

- the *empirical variance*

$$\begin{aligned}\left(S_N^X\right)^2 &:= \frac{1}{N-1}\sum_{i=1}^N \left(\xi_i^X - \bar{\xi}_N^X\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^N \left[\left(\xi_i^X\right)^2 - \left(\bar{\xi}_N^X\right)^2\right]\end{aligned}$$

  whose possible outcome is the sample variance and whose expectation value is σ^2. Furthermore the probability distribution of the r.v. $\frac{(N-1)}{\sigma^2}\left(S_N^X\right)^2$ is a χ^2-distribution with $N-1$ degrees of freedom.

Remark

We remark that the expectation value of the empirical mean and of the empirical variance is always respectively the expectation value and the variance of ξ^X. Moreover, in the Gaussian case, these two r.v's are stochastically independent.

The Pearson χ^2 test

How can we understand if the data sample we have produced is composed by the outcomes of a collection of Gaussian i.i.d.r.v's?
Here is a criterion to establish if the data sample $\mathcal{C}_N(X)$ issued from our experiment is or not compatible with the hypothesis of being an outcome of a collection of N i.i.d.r.v's $N(\bar{x}, s_X)$.
Let us partiton $\mathbb{R}$ in $2(K+1)$ disjoint intervals setting

- $\mathbb{R} = (-\infty, \bar{x}] \cup (\bar{x}, +\infty)$;
- $\forall k = 0, .., K-1,$

$$\begin{aligned} A_{k+1} &:= (\bar{x} + k s_X, \bar{x} + (k+1) s_X] \ , \\ A_{-(k+1)} &:= (\bar{x} - (k+1) s_X, \bar{x} - k s_X] \ ; \end{aligned}$$

- $A_{K+1} := (\bar{x} + K s_X, +\infty) \ , \ A_{-(K+1)} := (-\infty, \bar{x} - K s_X) \ .$

Hence,

$$(\bar{x}, +\infty) = \bigcup_{k=0}^{K-1} A_{k+1} \bigcup A_{K+1} \ , \ (-\infty, \bar{x}] = \bigcup_{k=0}^{K-1} A_{-(k+1)} \bigcup A_{-(K+1)} \ .$$

The Pearson χ^2 test (continued)

Then,$\forall l \in \{-(K+1),..,-1\} \cup \{1,..,K+1\}$, let

$$O_l := |\{x \in \mathcal{C}_N(X) : x \in A_l\}| = \sum_{i=1}^{N} \mathbf{1}_{A_l}(x_i)$$

be the observed frequencies of the data subsamples of $\mathcal{C}_N(X)$ associated to the elements of the partition $\{A_k\}_{k\in\{-(K+1),..,-1\}\cup\{1,..,K+1\}}$ of $\mathbb{R}$. Let also, $\forall k \in \{-(K+1),..,-1\} \cup \{1,..,K+1\}$,

$$E_k := N \int_{A_k} dx g(x|\bar{x}, s_X) \ .$$

remark

By the symmetry of the Gaussian probability density function

$$g(x-\bar{x}|\bar{x}, s_X) = g(-(x-\bar{x})|\bar{x}, s_X),$$

$\forall k = 1,..,K+1, E_k = E_{-k}$ so it sufficient to compute just the E_k's.

The Pearson χ^2 test (continued)

If we choose K such that:

1. $K \geq 1$ and is smaller than N;
2. the O_k are always positive;

we can define the quantity

$$\bar{\chi}^2 := \frac{1}{2(K+1)-3} \sum_{k=1}^{K} \left[\frac{(O_k - E_k)^2}{E_k} + \frac{(O_{-k} - E_{-k})^2}{E_{-k}} \right]$$

called *normalized* χ^2 which is the outcome of the statistic $[2(K+1)-3]\,\mathbf{X}^2$ which, for large size of the sample N is distributed as χ^2 with $2(K+1)-3$ degrees of freedom r.v.. Hence, if the deviation of O_k from E_k is small enough $\bar{\chi}^2$ will be close to 0 and the hypothesis of $\mathcal{C}_N(X)$ being generated by the outcome of a $N(\bar{x}, s_X)$ r.v. will be plausible.

TEST Given a data sample of size N, the hypothesis that the data are the outcomes of a r.v. $N(\bar{x}, s_X)$ is accepted if $\bar{\chi}^2 \leq 1$ otherwise is rejected.

Exercise

Test the hypothesis that the following data sample is $N(\bar{x}, s_X)$.

16	20	21	23	34	38	22	33	23	18
17	34	49	55	60	49	45	45	44	41
44	34	38	35	20	23	22	33	46	22
29	24	38	18	35	26	20	36	28	39
17	31	26	32	16	40	32	27	28	17

The t-test

We want to test the hypothesis that our data sample $\mathcal{C}_N\left(X\right)$ represents the outcomes of a Gaussian r.v. $N\left(\mu,\sigma\right)$ where μ is equal to a given value μ_0. Hence,

$$H_0 : \mu = \mu_0 \ .$$

Let t denote the observed value of the test statistic $T = \frac{\sqrt{N}(\bar{\xi}_N^X - \mu_0)}{S_N^X} \in T_{N-1}$. Then compute the probability that $|T|$ would exceed $|t|$. This is called p-*value* of the test. The test then calls for rejection at all significance levels α higher than the p-value and acceptance at all lower significance levels.

Exercise

Given the data sample

31	33	35	39	41	43	45	47	51	53
55	57	61	63	65	67	69	71	75	77
79	81	87	89	91	97	29	59	73	49

test the hypothesis $H_0 : \mu = 62$.

Test on the variance of a Gaussian sample

We want to test the hypothesis that our data sample $\mathcal{C}_N(X)$ represents the outcomes of a Gaussian r.v. $N(\mu,\sigma)$ where σ is equal to a given value σ_0. Hence,

$$H_0 : \sigma^2 = \sigma_0^2 \ .$$

Let s denote the observed value of the test statistic $S = (N-1)\frac{(S_N^X)^2}{\sigma_0^2} \in \chi^2_{N-1}$. Then compute the probabiliy that S is smaller than s. The quantity

$$p = 2\left(\mathbb{P}\{S < s\} \wedge 1 - \mathbb{P}\{S < s\}\right)$$

is called p-*value* of the test. The test then calls for rejection at all significance levels α higher than the p-value and acceptance at all lower significance levels.

Exercise

Given the data sample of the previous exercise, test the hypothesis $H_0 : \sigma^2 = 361$.

S. Ross *Introduction to Probability and Statistics for Engineers and Scientists, III edition* Elsevier Academic Press (2004).