# Inferenza

24 ottobre 2023

# Indice

# Capitolo 1

# Introduction to inference

*Osservazione importante* 1 (Statistics and machine learning). We have that:

- **statistics** was born in 1917, where the first contribution in classical inference (in modern sense) was due to Fisher. The general idea is that we have input data, we work on data (descriptive, inference). Most of time what we do is construct a model.
  The main interest of statistics is have good *interpretation* of data; understanding what data suggest us.
- **machine learning** (ML) on the other hand, speaks about algorithms; was born (from statistician and computer scientist) in the 80's because of computer availability.
  You have a kind of black box, you don't care about interpretation, don't know how it works: important is, as output, to have a good prediction (eg for image/audio recognition)

*Osservazione* 1. What is best? it's subjective, there are no closed boundaries, there is overlap between two; in stat we speak about models, in ml about algorithms.

## 1.1 Classical inference setup

*Osservazione importante* 2 (Setup). We observe a sample, subset of population, composed by $n$ (sample size) observation $(x_1, \dots, x_n)$ (denoted in *lowercase* letters).
We can view the sample as realization of $n$ random variables $(X_1, \dots X_n)$ (in *capital* letters): we assume that each rv is distributed according to a common an $F_X$ we don't know (the distribution function in the population). We want to infer characteristics of $F_X$ from the sample.

**Definizione 1.1.1** (Parametric inference)**.** It's when one assume that $F_X$ is a probabilistic model characterized by a parameter $\theta$ from a parameter space $\mathbf{\Theta}$:

$$F_X(\theta) = \{F(x;\theta) : \theta \in \mathbf{\Theta}\}$$

In this framework to make inference it's enough to estimate $\theta$ (eg for point estimation, interval estimation); if you know $\theta$ you know everything.

**Definizione 1.1.2** (Nonparametric)**.** The set of all possible distribution $F$ of interest is not restricted to belong to a probabilistic model, it's the complete set of all possible distribution function:

$$F = \{\text{all the CSF's}\}$$

So in this framework one doesn't have a probabilistic model and therefore a parameter $\theta$ to be estimated.

## 1.2 Parametric inference

*Osservazione importante* 3. Imagine we observed a sample of $n$ observation $(x_1, \ldots, x_n)$: we want to find the best guess for the parameter $\theta$ or a transformation of the parameter $T(\theta)$. In order to do our guess we're aware that when we make inference we can make mistakes; the idea however is to reduce occurrence by

- choosing best formula/estimator to do our guess
- reducing the variability and increase the precision of our guess (with sample size).

### 1.2.1 Point estimation

*Osservazione importante* 4 (Estimator, estimate). In the following we write:

- $T_n = T(X_1, \ldots, X_n)$ to mean our *estimator* for $\theta$, that is the procedure/statistics/transformation we apply on random variable to obtain an estimate; being a function of random variable, it's a random variable itself with an own distribution;
- $t_n = \hat{\theta} = T(x_1, \ldots, x_n)$ to mean our *estimate*, that is the result of applying the estimator to our sample.

### 1.2.2 Property of estimators

*Osservazione importante* 5 (On quality of estimators). When we have an estimate $t_n$ we don't know if it's good or not for $\theta$; our trust on $t_n$ is based on the behaviour of $T(X_1, \ldots, X_n)$ in the family of possible (infinite) samples, that is in the sample space $\Omega$.
Therefore it's crucial to search for estimators with good behaviour.

*Osservazione importante* 6 (Property of estimators). The desiderable properties for estimators are

1. unbiasedness
2. efficiency (comparative)
3. consistency

#### 1.2.2.1 Unbiasedness

**Definizione 1.2.1** (Unbiasedness)**.** $T_n = T(X_1, \ldots, X_n)$ is unbiased for $\theta$ if and only if $\mathbb{E}[T_n] = \theta$, $\forall \theta \in \boldsymbol{\Theta}$ .

*Osservazione* 2. We look at expectation because, as said, the estimator is a function of random variables, so it's a random variable itself.

**Definizione 1.2.2** (Bias)**.** It's the difference between the expected value and the real parameter to be estimated:

$$\text{Bias}(T_n) = \mathbb{E}[T_n] - \theta \tag{1.1}$$

*Osservazione importante* 7. One cannot compute the bias, because we don't know $\theta$.

**Esempio 1.2.1** (Sample mean)**.** Let $X_i$ be independent rvs with expected value $\mathbb{E}[X_i] = \mu$ and variance $\text{Var}[X_i] = \sigma^2$. The sample mean is defined as

$$T_n = \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \tag{1.2}$$

We have that

$$\mathbb{E}\left[\bar{X}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n}\sum_i \mathbb{E}[X_i] = \frac{1}{n} n\mu = \mu$$

$$\text{Var}\left[\bar{X}\right] = \text{Var}\left[\frac{\sum_i X_i}{m}\right] = \frac{1}{n^2}\text{Var}\left[\sum_i X_i\right] \overset{(\perp\!\!\!\perp)}{=} \frac{\sum_i \text{Var}[X_i]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

So the sample mean is a unbiased estimator of the mean $\mu$; its variance $\frac{\sigma^2}{n}$ is directly associated to population variance but it collapses on $\mu$ as $n \to \infty$.

**Esempio 1.2.2.** Let $X \sim \text{Bern}(p)$, our parameter of interest is $p$ and we have two estimators for it:

$$T_n^{(1)} = \frac{\sum_i X_i}{n}$$
$$T_n^{(2)} = \frac{\sum_i X_i^2}{n}$$

Checking for bias we have

$$\mathbb{E}\left[T_n^{(1)}\right] = \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] = \frac{\sum_i \mathbb{E}[X_i]}{n} \overset{(1)}{=} \frac{\sum_i p}{n} = \frac{np}{n} = p$$

$$\mathbb{E}\left[T_n^{(2)}\right] = \mathbb{E}\left[\frac{\sum_i X_i^2}{n}\right] = \frac{\sum_i \mathbb{E}\left[X_i^2\right]}{n} \overset{(1)}{=} \frac{\sum_i p}{n} = \frac{np}{n} = p$$

where in (1) since it's a bernoulli, and in (2) given that for the bernoulli distribution (and only for her) all the moments are equal (first moment = $p$, second = $p$ etc), or otherwise using computational formula for variance and obtaining $\mathbb{E}\left[X_i^2\right]$.
Therefore both estimators are unbiased: even the second estimator is so the mean is the not the only estimator. We could take the mean of power two three as well.

**Esempio 1.2.3** (Sample variance)**.** Given $X_i$ iid rvs with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, if our interest is in estimating $\sigma^2$. One estimator could be the sample variance, defined as follows using the sample mean:

$$T_n = \frac{\sum_i^n (X_i - \bar{X})^2}{n} \tag{1.3}$$

However this is a biased of the variance of population $\sigma^2$. Let's see why. First some results we have: since

$$\mathbb{E}\left[\bar{X}\right] = \mu, \quad \mathrm{Var}\left[\bar{X}\right] = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[X_i^2\right] = \sigma^2 + \mu^2, \quad \mathbb{E}\left[\bar{X}^2\right] = \frac{\sigma^2}{n} + \mu^2$$

with the first two as properties of sample mean, the latters by the calculation formula of the variance (applied to $X_i$ or $\bar{X}$).
Now we work algebrically the sample variance formula to calculate expectation easier:

$$\begin{aligned} T_n &= \frac{\sum_{i=1}^n (X_i^2 + \bar{X}^2 - 2\bar{X}X_i)}{n} = \frac{\sum_{i=1}^n X_i^2 + \sum_{i=1}^n \bar{X}^2 - 2\bar{X}\sum_{i=1}^n X_i}{n} \\ &= \frac{\left(\sum_{i=1}^n X_i^2\right) + n\bar{X}^2 - 2\bar{X}n\bar{X}}{n} = \frac{\left(\sum_{i=1}^n X_i^2\right) - n\bar{X}^2}{n} = \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \end{aligned}$$

Now take expectation

$$\begin{aligned} \mathbb{E}[T_n] &= \frac{\sum_i \mathbb{E}\left[X_i^2\right]}{n} - \mathbb{E}\left[\bar{X}^2\right] = \frac{n(\sigma^2 + \mu^2}{n} - \left[\frac{\sigma^2}{n} + \mu^2\right] = \sigma^2 + \mu^2 - \frac{\sigma^2}{\mu} - \mu^2 \\ &= \frac{n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2}}{n} = \frac{\sigma^2(n-1)}{n} = \frac{n-1}{n}\sigma^2 \end{aligned}$$

Being $\mathbb{E}[T_n] \neq \sigma^2$, sample variance is biased with bias

$$\mathrm{Bias}(T_n) = \mathbb{E}[T_n] - \sigma^2 = \frac{\cancel{n\sigma^2} - \sigma^2 - \cancel{n\sigma^2}}{n} = -\frac{\sigma^2}{n}$$

Two consideration:

- for a certain $n$, the estimator is biased; however it's asymptotically unbiased/correct, that is, when $n$ increases the biasing factor $(n-1)/n$ goes to 1 or, otherwise stated, if one computes $\lim_{n\to+\infty} \mathrm{Bias}(T_n) = 0$
- for a certain sample size, by putting $n-1$ to denominator in 1.3, the estimator becomes unbiased.

*Osservazione* 3. An *estimator* for correctness/unbiasedness of our estimators is mean squared error: it measures how much $T_n$ is concentrated around $\theta$.
If it's low the estimator is precise: it's a kind of proximity measure of the estimator around the parameter of interest. It's an expected value on all the sample we could observe before performing the experiment

**Definizione 1.2.3** (Mean squared error)**.** It's defined as:

$$\mathrm{MSE}(T_n) = \mathbb{E}\left[(T_n - \theta)^2\right] \tag{1.4}$$

*Osservazione* 4. MSE can be decomposed according to a famous decomposition

**Proposizione 1.2.1** (Decomposition of mse)**.**

$$\mathrm{MSE}\left(T_n\right) = \mathbb{E}\left[(T_n - \theta)^2\right] = \mathrm{Var}\left[T_n\right] + \mathrm{Bias}\left(T_n\right)^2$$

*Osservazione* 5. The decomposition highlights the source of estimates imprecision of our estimator:

1. variability of the estimator (with respect to its expectation): as the most spread the distribution of the estimator is, the most error we'll make using it to estimate $\theta$;

2. bias of the estimator.

*Dimostrazione.*

$$\begin{aligned}
\mathbb{E}\left[(T_n - \theta)^2\right] &\overset{(1)}{=} \mathbb{E}\left[(T_n - \theta + \mathbb{E}\left[T_n\right] - \mathbb{E}\left[T_n\right])^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{T_n - \mathbb{E}\left[T_n\right]}_{} + \underbrace{\mathbb{E}\left[T_n\right] - \theta}_{}\right)^2\right] \\
&\overset{(2)}{=} \mathbb{E}\left[\left(T_n - \mathbb{E}\left[T_n\right]\right)^2 + \left(\mathbb{E}\left[T_n\right] - \theta\right)^2 + 2(T_n - \mathbb{E}\left[T_n\right])(\mathbb{E}\left[T_n\right] - \theta)\right] \\
&\overset{(3)}{=} \mathbb{E}\left[\left(T_n - \bar{\theta_n}\right)^2 + \left(\bar{\theta_n} - \theta\right)^2 + 2\left(T_n - \bar{\theta_n}\right)\left(\bar{\theta_n} - \theta\right)\right] \\
&\overset{(4)}{=} \underbrace{\mathbb{E}\left[\left(T_n - \bar{\theta_n}\right)^2\right]}_{\mathrm{Var}[T_n]} + \underbrace{\left(\bar{\theta_n} - \theta\right)^2}_{\mathrm{Bias}\,(T_n)^2} + 2\left(\bar{\theta_n} - \theta\right) \underbrace{\mathbb{E}\left[T_n - \bar{\theta_n}\right]}_{=0} \\
&= \mathrm{Var}\left[T_n\right] + \mathrm{Bias}\left(T_n\right)^2
\end{aligned}$$

before expanding the square in (1) we use a trick; in (2) we expand the square of the grouped stuff where in (3) we called $\mathbb{E}\left[T_n\right] = \bar{\theta_n}$; finally in (4) the bias factor doesn't need expectation since it's a difference of constant (and its expectation is a constant) and the last factor is zero simply by applying expected value properties and remembering that $\bar{\theta_n} = \mathbb{E}\left[T_n\right]$. □

*Osservazione* 6 (Jargon). In inference (eg bootstrap) we speak standard error when speaking about estimator

**Definizione 1.2.4** (Standard error of an estimator)**.** It's the standard deviation of the estimator $T_n$

$$\mathrm{SE}\left(T_n\right) = \sqrt{\mathrm{Var}\left[T_n\right]}$$

**Esempio 1.2.4.** Let $X_1, \ldots, X_n \sim \mathrm{Unif}\left(0, \theta\right)$. We want to estimate $\theta$ and consider this estimator

$$T_n(X_1, \ldots, X_n) = X_{(n)} = \max\left\{X_1, \ldots, X_n\right\}$$

Let's find $\mathrm{Bias}\left(T_n\right)$ and $\mathrm{MSE}\left(T_n\right)$ of our estimator. This is a typical exercise; given the distribution, the parameter of interest and the estimator find the properties of the latter.

First being uniforms and considering the maximum (order statistics) we remember that

$$F_{X_i}(x) = \frac{x-0}{\theta-0} = \frac{x}{\theta}$$
$$F_{(n)}(x) = [F_X(x)]^n = \left(\frac{x}{\theta}\right)^n$$
$$f_{(n)}(x) = n \cdot \left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} = n \cdot \frac{x^{n-1}}{\theta^n}$$

To compute the bias we should compute expectation of the estimator

$$\mathbb{E}\left[X_{(n)}\right] = \int_0^\theta x \cdot f_{(n)}(x)\,\mathrm{d}x = \int_0^\theta x \cdot n \cdot \frac{x^{n-1}}{\theta^n}\,\mathrm{d}x = \frac{n}{\theta^n}\int_0^\theta x^n\,\mathrm{d}x$$
$$= \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{\theta^n}\cdot\frac{\theta^{n+1}}{n+1} = \frac{\theta\cdot n}{n+1}$$

Now we can answer the first question. The bias of our estimator is:

$$\mathrm{Bias}\,(T_n) = \frac{\theta\cdot n}{n+1} - \theta = \frac{n\theta-(n+1)\theta}{n+1} = -\frac{\theta}{n+1}$$

For the MSE we need only to compute the variance of the estimator, since $\mathrm{MSE}\,(T_N) = \mathrm{Var}\,[Tn] + \mathrm{Bias}\,(T_n)^2$. To compute the variance we can use the fact that $\mathrm{Var}\,[T_n] = \mathbb{E}\left[T_n^2\right] - \mathbb{E}\,[T_n]^2$; we have already the first moment, we need only the second moment:

$$\mathbb{E}\left[T_n^2\right] = \mathbb{E}\left[X_{(n)}^2\right] = \int_0^\theta x^2 f_{(n)}(x)\,\mathrm{d}x = \int_0^\theta x^2\cdot n\cdot\frac{x^{n-1}}{\theta^n}\,\mathrm{d}x$$
$$= \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,\mathrm{d}x = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n}{\theta^n}\left[\frac{\theta^{n+2}}{n+2}\right]$$
$$= \frac{\theta^2 n}{n+2}$$

And so the variance of $T_n$ is

$$\mathrm{Var}\,[T_n] = \mathbb{E}\left[T_n^2\right] - \mathbb{E}\,[T_n]^2 = \frac{\theta^2 n}{n+2} - \frac{\theta^2 n^2}{(n+1)^2}$$
$$= \frac{\theta^2 n(n+1)^2 - \theta^2 n^2(n+2)}{(n+2)(n+1)^2} = \ldots = \frac{\theta^2 n}{(n+2)(n+1)^2}$$

It cannot be decomposed more than above. Now for the MSE of the estimator

$$\mathrm{MSE}\,(T_n) = \frac{\theta^2 n}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} = \ldots = \frac{2\theta^2}{(n+1)(n+2)}$$

**Esempio 1.2.5.** Try to compute bias and mse of the minimum $X_{(1)}$ and check it will be not a good estimator: it will estimate the minimum of the interval of the uniform, that is 0, not $\theta$.

#### 1.2.2.2 Efficiency

**Definizione 1.2.5** (First definition)**.** Let $T_n^{(1)}$ and $T_n^{(2)}$ be two estimators for $\theta$. We say that $T_n^{(1)}$ is *more efficient* than $T_n^{(2)}$ if:

$$\text{Var}\left[T_n^{(1)}\right] \leq \text{Var}\left[T_n^{(2)}\right]$$

**Definizione 1.2.6** (Second definition)**.** Let $T_n^{(1)}$ and $T_n^{(2)}$ be two estimators for $\theta$. We say that $T_n^{(1)}$ is more efficent than $T_n^{(2)}$ if

$$\text{MSE}\left(T_n^{(1)}\right) \leq \text{MSE}\left(T_n^{(2)}\right)$$

*Osservazione* 7. Btw this second definition consider also the information about the bias, since the MSE can be decomposed in a part of variance and in one of bias of the estimator; if we know that the two estimators have the same bias then the two definitions are equivalent (since cancelling out the bias at the two members from the second gives the first).

*Osservazione* 8 (Relative efficiency). One could equivalently look at *relative efficiency* defined as ratio. Adopting the MSE definition we have:

$$e\left(T_n^{(1)}, T_n^{(1)}\right) = \frac{\text{MSE}\left(T_n^{(2)}\right)}{\text{MSE}\left(T_n^{(1)}\right)}$$

if $e\left(T_n^{(1)}, T_n^{(2)}\right) > 1 \implies T_n^{(1)}$ is preferable

#### 1.2.2.3 Consistency

*Osservazione importante* 8. It considers the behaviour of an estimator as $n$ (sample size) increases: if the estimator is consistent, as $n$ increases we have better results.
Consistency and unbiasedness are *indepent* properties/definitions: we can have an estimator which is consistent but is biased and viceversa.

**Definizione 1.2.7** (Simple (weak) consistency)**.** $T_n$ is weakly consistent for $\theta$ if it converges in probability $T_n \xrightarrow{p} \delta_\theta$, that is:

$$\lim_{n\to\infty} \mathbb{P}\left(|T_n - \theta| < \varepsilon\right) = 1, \quad \forall \theta \in \boldsymbol{\Theta}, \varepsilon > 0$$

*Osservazione importante* 9. Remember that we have the two sufficient condition that can be used to prove weak convergency:

$$\begin{cases} \lim_{n\to\infty} \mathbb{E}\left[T_n\right] = \theta \text{ (or } \lim \text{Bias}\left(T_n\right) = 0) \\ \lim_{n\to\infty} \text{Var}\left[T_n\right] = 0 \end{cases} \implies T_n \xrightarrow{p} \delta_\theta$$

Given the decomposition of the MSE (in variance and bias) the above result is equivalent to the following one

$$\lim_{n\to\infty} \text{MSE}\left(T_n\right) = 0 \implies T_n \xrightarrow{p} \delta_\theta$$

**Esempio 1.2.6.** In this example the difference between unbiasdness and consistency, with an estimator which is biased but consistent.
Let $X_n \sim \text{Bern}\left(\frac{1}{n}\right)n$. We have that:

$$\mathbb{E}[X_n] = 1 \cdot \frac{1}{n} \cdot n + 0 \cdot \left(1 - \frac{1}{n}\right) \cdot n = 1, \quad \forall n$$

but we have already shown that $X_n \xrightarrow{p} \delta_0$; so the expectation is 1 for every $n$ (we have bias), but when $n$ increases the estimator goes to 0 (which is the true value of the parameter)

**Esempio 1.2.7** (My take on previous example)**.** we have that $X_n \sim \text{Bern}\left(\frac{1}{n}\right)n$ and we want to estimate the parameter of the distribution $\theta = \frac{1}{n}$ (which is a moving value); we use as estimator the last observed value $T_n = X_n$. The estimator is biased since

$$\mathbb{E}[T_n] = \mathbb{E}[X_n] = 1 \cdot \frac{1}{n} \cdot n + 0 \cdot \left(1 - \frac{1}{n}\right) \cdot n = 1 \neq \theta = \frac{1}{n}$$

However it's consistent since $T_n = X_n \xrightarrow{p} \delta_0$ as proved previously; when $n$ increases the estimator goes to 0 (as the parameter $\theta = \frac{1}{n}$ does.)

**Definizione 1.2.8** (Strong consistency)**.** $T_n$ is strongly consistent for $\theta$ iff if $T_n \xrightarrow{L_2} \delta_\theta$:

$$\lim_{n\to\infty} \mathbb{E}\left[(T_n - \theta)^2\right] = 0, \quad \forall \theta \in \boldsymbol{\Theta}$$

*Osservazione importante* 10. In this case the definition using MSE is (both sufficient and necessary):

$$\lim_{n\to\infty} \text{MSE}(T_n) = 0 \iff T_n \xrightarrow{L_2} \delta_\theta$$

## 1.3 Methods for finding estimators

*Osservazione* 9. So far we discussed property of given estimators; but how to find them? Now we focus on this problem, with an historical approach.
The most important methods are least square, moments and likelihood, but not the only ones. Maximum likelihood is the most important/used one

### 1.3.1 Method of least squares

*Osservazione* 10. One of the first methods used/developed. We encounter this methods in linear models course, since there is very used.
Here we use $Y_i$ instead of $X_i$ since it's general notation (this's what she said).

**Definizione 1.3.1.** If $Y_1, \ldots, Y_n$ are independent rvs with same variance and higher moments, with $\mathbb{E}[Y_i] = \mathcal{T}(\theta)$[1], where $\mathcal{T}$ is a *linear* function, then the least square estimate for $\theta$ is obtained by minimizing

$$\sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i])^2$$

[1] This is general, but the function can identity as well, eg $\mathbb{E}[Y_i] = \theta$. Eg the transformation is used for glm.

*Osservazione importante* 11. Pros and cons:

- Pro: least square estimators are BLUE (best linear unbiased estimators; est since the most efficient one)
- Cons: limited applicability (we cannot use this square in many problems because we don't have a regression model)

### 1.3.2 Method of minimum distance

*Osservazione* 11. Not very used/known but we cite it. First some prereqs.

**Definizione 1.3.2** (Empirical distribution function)**.** It's defined as the distribution function we can construct with our sample without assuming any probabilistic distributional form:

$$F_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

where $I(A)$ is the indicator function

**TODO**: mettere a posto l'indicatrice nelle sintesi latex

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

**Esempio 1.3.1.** Let $(x_1, \ldots, x_5) = (2, 3, 5, 5, 7)$; then we have $F_n(0) = 0$, $F_n(3) = \frac{2}{5}$, $F_n(5) = \frac{4}{5}$ and $F_n(10) = 1$

*Osservazione importante* 12. It can be shown that the empirical distribution function is a good estimator for the true distribution function, that is $F_n$ is a good estimator for $F_X$.

**Definizione 1.3.3** (Method of minimum distance)**.** Let $(x_1, \ldots, x_n)$ be a random sample from $F_X(x; \theta)$ with $\theta$ our object of interest. Now let $F_n(x)$ be the *empirical* distribution function (not the maximum, without tonde). An estimate for $\theta$ can be obtained by minimizing (by $\theta$) the *distance* between the empirical distribution function and the theorical/true distribution function:

$$\min_{\theta} d\left[F_n(x), F_X(x; \theta)\right]$$

*Osservazione importante* 13. The distance can be any distance function: eg euclidean, manhattan or the maximum distance. The latter is defined as the maximum value of the the absolute difference

$$\sup_{x \in D_X} |F_n(x) - F_X(x; \theta)|$$

Different choice about distance imply different results

*Osservazione importante* 14. Pros and cons:

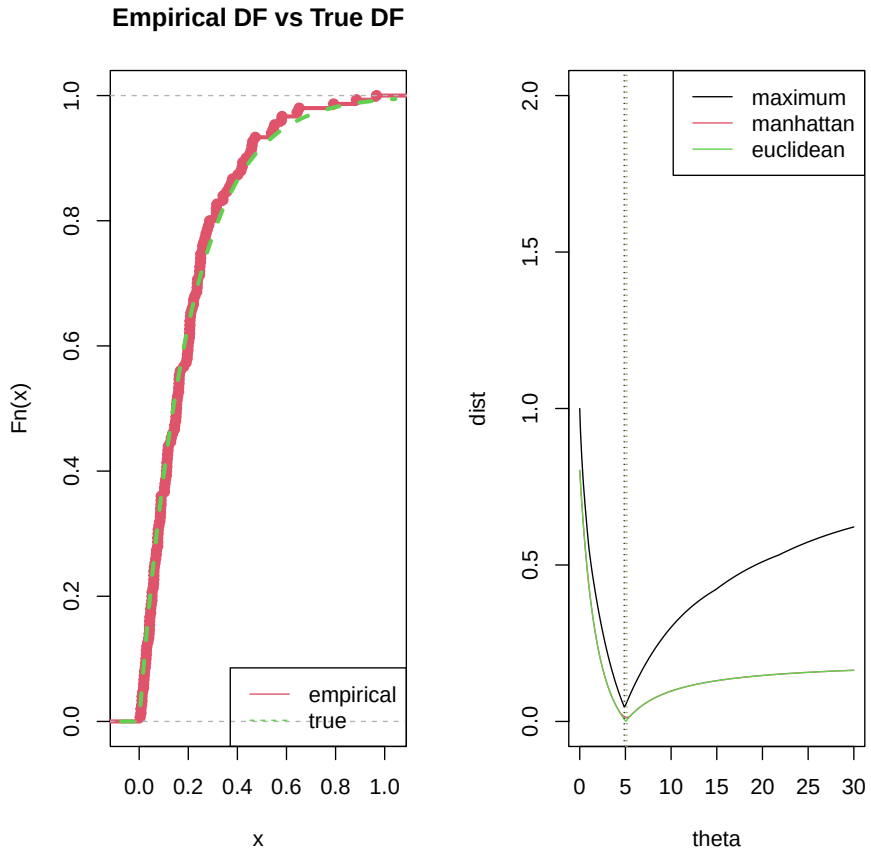
1. pros: very large applicability. Can be used to estimate one or more parameters with sophisticadted method of optimization. It needs assumption on theoretical distribution for my data only;

2. cons: we have just the result of the optimization problem, is difficult to get an analytical function for the estimator $\hat{\theta_n}$ so we cannot study the theoretical properties of the estimator (it's similar to neural networks where one specifies cost function and minimize it)

**Esempio 1.3.2** (Example of method of minimum distance)**.** we have

```
set.seed(1)

# our sample
theta = 5
n = 150
x = rexp(n, theta)

## its empirical cdf
(Fn = ecdf(x))

## Empirical CDF
## Call: ecdf(x)
##  x[1:150] = 0.0040501, 0.0074305, 0.0074537,  ..., 0.88479, 0.96656

## Note that after estimating the Fn we can use it as a normal
## function! eg Fn(x)
Fn(0.5)

## [1] 0.9333333

## ------------------------------------------------------------------------
## 1) Ecdf is a good approximation of real df for n to infty
## ------------------------------------------------------------------------
## this full comparison plot can't be done in reality because we don't
## know theta. However as n go to infty (here is 150) the empirical
## distribution function is a good estimator of population cdf (if we
## do the same with n = 1000 its almost perfectly overlapping)

par(mfrow = c(1,2))
plot(Fn, lwd = 3, col = 2, main = "Empirical DF vs True DF")
curve(pexp(x, theta), add = TRUE, lwd = 3, lty = 2, col = 3)
legend("bottomright", legend = c('empirical', "true"),
       col = c(2,3), lty = c(1,2))

## -------------------------
## 2) minimum distance method
## -------------------------
## imagine that we know the population distribution is exponential but
## we don't know the parameter theta (here = 5). How do we find theta =
## 5 by using this method?

### distance

## To optimize for theta we need to write a function depending only on
```

```
## it. Therefore we define some instrumental stuff

## points on F, F_n codomain where the distance are evaluated
xx = seq(0, 1, length.out = 1000)

## we use as distance the following
d1 = function(theta) max(abs(Fn(xx) - pexp(xx, theta))) # maximum
d2 = function(theta) mean(abs(Fn(xx) - pexp(xx, theta))) # manhattan
d3 = function(theta) sqrt(mean(Fn(xx) - pexp(xx, theta))^2) # euclidean

# minimizing for theta, the didactic way
# --------------------------------------
## theta searched for in maximization
theta.val = seq(0, 30, length.out = 1000)
## vectors of distance (for each theta) between Fn and F
out1 = out2 = out3 = 0
for (i in 1:1000) {
    out1[i] = d1(theta.val[i])
    out2[i] = d2(theta.val[i])
    out3[i] = d3(theta.val[i])
}
## plot of distances for several thetas
plot(theta.val,  out1, type='l', col=1,
     ylim=c(0,2), xlab = 'theta', ylab = 'dist')
lines(theta.val, out2, type='l', col=2)
lines(theta.val, out3, type='l', col=3)
legend('topright', legend = c("maximum", "manhattan", "euclidean"),
       col = 1:3, lty=1)
## find where the distance is minimal
theta.val[which.min(out1)]

## [1] 4.864865

theta.val[which.min(out2)]

## [1] 5.165165

theta.val[which.min(out3)]

## [1] 5.045045

abline(v = theta.val[which.min(out1)], col = 1, lty='dotted')
abline(v = theta.val[which.min(out2)], col = 2, lty='dotted')
abline(v = theta.val[which.min(out3)], col = 3, lty='dotted')
```




```
## all the estimates are not on 5; they come from an estimator and
## depends on the sample

## we don't have a superiority of a distance over another one, it
## depends on the sample (so look at every)

## minimizing for theta, the automatic way
## ---------------------------------------
## with optim you do the same in a quick manner.
## optim minimizes by default so if one has to optimize
## change sign
## - the first is the initial value for the parameters
##   to be optimized over (it doesn't matter)
## - the second is the function to optimize
## - the third is optimization method
## $par returns the estimated parameter
optim(3, d1, method = "BFGS")$par
```

```
## [1] 4.873426
```

```
optim(3, d2, method = "BFGS")$par
```

```
## [1] 5.172686

optim(3, d3, method = "BFGS")$par

## [1] 5.042332

## if we below increase n (eg 1000) we expect a better estimates,
## nearer to 5) at least on the poor man way
```

### 1.3.3 Method of moments (MM)

*Osservazione* 12. Originates with Karl Pearson in 1894, one of the first method invented; at the days there were no computers and calculation were done by hand. The idea behind this method is smart.

Let $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_k \end{bmatrix}$ be unknown quantities/parameters of interest. So we start esplicitly in a multivariate situation (good, extension of minimum distance on multivariate can be complicate).
Now we define $\mu_j = \mathbb{E}\left[X^j\right]$ the $j$-th *moment* of our variable $X$

**Esempio 1.3.3.** Imagine we have a gaussian $\mathrm{N}\left(\mu, \sigma^2\right)$ and we are interested in estimating the couple of its parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \quad k = 2$$

then we define a system where we have the first two moments, and we make an equivalence between these two moments and the moments of the population (respectively $\mu$ and given the computation variance formula $\sigma^2 + \mu^2$)

$$\begin{cases} \mu_1 = \mathbb{E}\left[X^1\right] = \mu \\ \mu_2 = \mathbb{E}\left[X^2\right] = \sigma^2 + \mu^2 \end{cases}$$

Define the *j-sample* moment (not the population moment) as:

$$M_j = \frac{\sum_{i=1}^{n} x_i^j}{n}$$

Finally the method of moments (MM) define the estimator $\hat{\boldsymbol{\theta}}_n$ such that we can construct a system where we equal sample moments (which we can calculate) and population moments, which are functions of the population parameters

$$\begin{cases} M_1 = \mu_1 \\ M_2 = \mu_2 \\ \dots \\ M_k = \mu_k \end{cases}$$

In case where we have a gaussian $\mathrm{N}\left(\mu, \sigma^2\right)$ and we want to estimate $\mu, \sigma^2$ we set up the following equation system

$$\begin{cases} M_1 = \frac{\sum_{i=1}^n x_i}{n} = \bar{x} = \mu \\ M_2 = \frac{\sum_{i=1}^n x_i^2}{n} = \sigma^2 + \mu^2 \end{cases}$$

According to this system, the estimators of $\mu$ and $\sigma^2$ becomes:

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \frac{\sum_i x_i^2}{n} - \hat{\mu}^2 = \frac{\sum_i x_i^2}{n} - \bar{x}^2 = \frac{\sum_i (x_i - \bar{x})^2}{n} \end{cases}$$

Here we find the sample variance which we know is a biased estimator (only asymptotically unbiased).

**Esempio 1.3.4.** Let $(x_1, \ldots, x_n)$ be a sample of $X$ with pdf

$$f(x;\theta) = \begin{cases} \frac{2}{\theta^2}(\theta - x) & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

This not a known density; use MM to get an estimator of $\theta$, $\hat{\theta}$.
This is a univariate problem; to solve it we equate the first sample moment equal to the first population moment. First of all we compute the first population moment

$$\begin{aligned} \mathbb{E}[X] = \mu_1 &= \int_0^\theta x \frac{2}{\theta^2}(\theta - x)\, \mathrm{d}x = \int_0^\theta \frac{2x}{\theta^2}\theta\, \mathrm{d}x - \int_0^\theta \frac{2x^2}{\theta^2}\, \mathrm{d}x \\ &= \frac{2}{\theta}\left[\frac{x^2}{2}\right]_0^\theta - \frac{2}{\theta}\left[\frac{x^3}{3}\right]_0^\theta = \frac{2}{\theta}\frac{\theta^2}{2} - \frac{2}{\theta^2}3 = \theta - \frac{2}{3}\theta \\ &= \frac{\theta}{3} \end{aligned}$$

Therefore to get the formula for the estimator we equate

$$\frac{\theta}{3} = M_1 = \bar{x}$$

in order to obtain the estimator

$$\hat{\theta} = 3\bar{x}$$

## 1.4 Inference: direct and inverse problem

*Osservazione importante* 15 (General framework based on probabilistic models). The probabilistic framework

1. $X$ rv that describes a feature of interest on the population;
2. one set a **probabilistic model**: we make an assumption on the distribution of $X$: the distribution of $X$ is indexed by a parameter $\theta \in \mathbf{\Theta}$ (or a vector of parameters), so $f(x;\theta)$

3. we observe a random sample $(x_1, \ldots, x_n)$ according to a proper **sampling model**.
   If the sampling model is the *simple random sampling* with replacement (most common and what we assume here), then $X_1, \ldots, X_n$ are iid rvs distributed according to $f(x; \theta)$.
   Simple sampling scheme is not true in practice: eg we don't have replacement (when you observe/sample people you observe a person just one time and don't repeat the observation), but if the population is large enough, with/without replacement becomes equivalent (the probability of observing the same person is very rare/impossible so, even if we replace, taking the second time the same person is very rare) , so we assume it's with replacement even it's not true in practice.
   In other courses other sampling scheme are studied as well (eg stratified, clustering and so on). According to different sampling you develop different theory in inference.

4. the assumption of both *probabilistic model* and *sampling model* are equal to our the **statistical model**, from which we derive our inference

How to make inference on $\theta$? In the general probabilistic framework there are two other methods for finding estimators: maximum likelihood and bayesian estimator.
At the beginning of history people were divided in these two methods; todays situation is different especially in advanced statistics were methods are mixed but in the foundational aspect of statistics were in contraddiction

1. frequentist (classic) framework: we work in a direct problem with a likelihood function $f(\mathbf{x}|\theta)$. The likelihood function is not what we want to know; it's the probability given the parameter of observing our sample

2. bayesian framework: we work in a inverse problem with the posterior function $f(\theta|\mathbf{x})$. The posterior what we want to know/do: it is the probability of the possible values of theta given our sample

It's seems contrary what is direct and inverse: Fisher was inventor of likelihood function and justified the idea of likelihood function which is strange. The idea is: i want to find the parameter that better justify what i have observed. if the probability of my sample given theta = 5 is low, then i will discard it. I try different theta and keep what according to which what i've observed have maximum

Both functions are probability: likelihood function is a probability for the data (even if we already observed them), the posterior function is a probability for the values of the parameter.

Both methods are very used in statistics; we have advantages and disadvantages for both of them.

In most complicated models bayesian stuff allows us to have an inference, while solution based on likelihood could be very difficult. In most simple models likelihood is the most used method

EM algorithm is likelihood based; stochastic EM algorithm is a mixed frequentist-bayesian

**Likelihood: frequentist (classic) framework** It arise under simple sampling scheme *only* (otherwise it's not likelihood). Under simple sampling scheme *with replacement* $X_1, \ldots, X_n$ are iid; then

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

the joint density (left: which is the density of our observation) is the product of marginal density (according to independence of observation).
$f(\mathbf{x}|\theta)$ is the probability function of my data $\mathbf{x}$ given $\theta$. Why do we use it? choose $\theta$ that makes most likely what we observed. Some notation:

$$\begin{aligned} \mathrm{L}(\theta) &= f(\mathbf{x}|\theta) \qquad \text{Likelihood function} \\ \ell(\theta) &= \log(\mathrm{L}(\theta)) \qquad \text{Log-likelihood function} \end{aligned}$$

Likelihood means the level of agreement between what we have observed and the possible level of $\theta$ that generated it. $\hat{\theta}_n$ can be estimated by maximizing $\mathrm{L}(\theta)$ or $\ell(\theta)$; log-likelihood is used since maximization happen for the same value of $\theta$ (the maximum of a positive function is preserved if we transform it according to a logarithm) and most of time it simplifies the derivation/optimization just from a math standpoint.

**posterior function: bayesian framework** On the contrary $f(\theta|\mathbf{x})$ is a probability function of $\theta$ given the data we have observed. This is more logical.
We can rewrite the posterior function according to the bayesian rule/theorem:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\,\mathbb{P}(B|A)}{\mathbb{P}(B)}$$

as follows:

$$f(\theta|\mathbf{x}) = \frac{f(\theta, \mathbf{x})}{f(\mathbf{x})} = \frac{f(\theta) \cdot f(\mathbf{x}|\theta)}{f(\mathbf{x})} \propto f(\theta) \cdot f(\mathbf{x}|\theta)$$

The last used "proportional to" because we can ignore the denominator.

- $f(\mathbf{x})$ at denominator does not depend on the model, eg we don't depend on $\theta$. So the denominator is the probability of observing the data without any specific probabilistic model (according to any possible probabilistic model); btw it can be rewritten as the integral of all possible models/values of theta (think law of total probabilities)

  $$f(\mathbf{x}) = \int_{\boldsymbol{\Theta}} f(\theta) f(\mathbf{x}|\theta)\, \mathrm{d}\theta$$

  is the marginal distribution of $X$.

- $f(\theta)$ is called prior on $\theta$: it's the probability without observing data (before the experiment);

- $f(\mathbf{x}|\theta)$ is the likelihood so the probability of data given $\theta$

So we say the posterior function is proportional to the product between the *prior* and *likelihood*.
The two approaches (frequentist and bayesian) are linked (at the numerator): in one approach one have a prior, in the other not, this is the only difference.

**Esempio 1.4.1.** Imagine we have some observation from a Bernoulli of unknown parameter, $X_1, \dots X_n \sim \text{Bern}(\theta)$ iid; imagine we have a *prior distribution* for the parameter, that is $\theta \sim \text{Beta}(a, b)$, with $a, b$ known. Find the Bayes *estimator* for $\theta$.
We want to find the posterior distribution that is

$$\begin{aligned} f(\theta|\mathbf{x}) = \frac{f(\theta) \cdot f(\mathbf{x}|\theta)}{f(\mathbf{x})} &\overset{(1)}{\propto} f(\theta) f(\mathbf{x}|\theta) \\ &\overset{(2)}{=} \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \theta^{a-1} \cdot (1-\theta)^{b-1}}_{\text{prior}} \cdot \underbrace{\prod_{i=1}^{n} \theta^{x_i} \cdot (1-\theta)^{1-x_i}}_{\text{likelihood}} \\ &= c \cdot \theta^{a-1} \cdot (1-\theta)^{b-1} \cdot \theta^{\sum x_i} \cdot (1-\theta)^{n-\sum x_i} \\ &\overset{(4)}{=} c \cdot \underbrace{\theta^{\sum x_i + a - 1} \cdot (1-\theta)^{n-\sum x_i + b - 1}}_{\text{kernel of a Beta}(\cdot)} \end{aligned}$$

where

- (1) since we want to maximize $f(\theta|\mathbf{x})$ we can ignore $f(\mathbf{x})$ (at its denominator) and all the terms that do not depend on $\theta$.
- (2) we substituted the density for theta (prior) from a beta distribution with parameters $a$ and $b$ (having $\theta$ as our $x$) and likelihood assuming iid observation (therefore the product) of bernoulli distribution with a given $\theta$ fixed
- (3) we rewrite the normalization constant of the gamma $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = c$ (because it does not depend on $\theta$) and sums of $x_i$ due to the product
- in (4) after putting terms together, we recognize the kernel of a Beta, especially $\text{Beta}(a + \sum x_i, b + n - \sum x_i)$. The *posterior is proportional to a kernel of a Beta so the posterior is a Beta*: this because we know the posterior is a proper density (constructed according to the Bayes rule) and is equal to the kernel of a beta times normalization constant not considered (we don't care about normalization constant, it doesn't alter the distribution, only set his integral to 1). Therefore the posterior is a beta. So In bayesian statistics our aim is to identify the kernel of the posterior; that is enough. therefore we conclude that $\theta|\mathbf{x} \sim \text{Beta}(a + \sum x_i, b + n - \sum x_i)$: here in bayesian approach we have a full distribution for the parameter estimator.

Three important facts :

1. when the prior $f(\theta)$ and the posterior $f(\theta|\mathbf{x})$ belong to the same probabilistic model (as in the case above, both have Beta), we say they are **conjugate**

2. in the posterior parameters ($a$, $b$, $\sum x_i$, $n-\sum x_i$) we have the contribution of both prior ($a$,$b$) and likelihood ($\sum x_i$, $n - \sum x_i$)

3. we have a full probability distribution function for $\theta$, that is $f(\theta|\mathbf{x})$. Then what is the best representative value/ our *estimate* for $\theta$? In the literature we have two solutions:

   (a) mode: this is a good choice because it maximizes $f(\theta|\mathbf{x})$

   (b) mean: $\mathbb{E}[\theta|\mathbf{x}] = \hat{\theta}$ is used if the posterior distribution is asymmetric (as was in this example: having the two parameters of the beta become very different the beta becomes asymmetric); in our example

$$\hat{\theta} = \frac{a + \sum x_i}{a + \sum x_i + b + n - \sum x_i} = \frac{a + \sum x_i}{a + b + n}$$

   Nowadays mean is more used than mode: computing mean is easier than computing a mode and it's handy for both symmetric and asymmetric distributions.

**Esempio 1.4.2.** What happens to the previous problem if we take as prior a lognormal density, for instance $\theta \sim \text{LogN}(\log(0.5), 0.1)$? The density of the lognormal theta (*prior*) is reported below (only the kernel *without* the normalization constant because in a bayesian framework we forget about it).

$$f(\theta) \propto \frac{1}{\theta} \exp\left(-\frac{1}{2}\frac{(\log\theta - \log 0.5)^2}{0.1}\right)$$

In this case the posterior will be

$$f(\theta|\mathbf{x}) \propto \frac{1}{\theta} \exp\left(-\frac{1}{2}\frac{(\log\theta - \log 0.5)^2}{0.1}\right) \cdot \theta^{\sum x_i} \cdot (1-\theta)^{n-\sum x_i}$$

In this case we cannot do further semplification, so $f(\theta|\mathbf{x})$ is known from an analytical point of view but we are not able to draw values from it: it's a kernel of a distribution we don't recognize, it's complicate.
To reconstruct the distribution (to obtain the estimate) we have to sample from it. We can do it by accept-reject algorithm (or one could do sampling-resampling, we didn't). Once we have generated many values from the target we have reconstructed the distribution of the target, we can take the mean, and it is the bayesian estimator.
How can we do it: remember that for accept-reject algorithm we should have a target $\pi(x)$ such as $\pi(x) \leq M \cdot$ proposal: in our case the target is the posterior $f(\theta|\mathbf{x})$ and we want to generate many values from it. It's simple, look at the step by step explanation below:

$$f(\theta|\mathbf{x}) \overset{(1)}{\propto} f(\theta) \cdot \mathrm{L}(\mathbf{x}|\theta) \overset{(2)}{\leq} \underbrace{f(\theta)}_{\text{proposal}} \cdot \underbrace{\mathrm{L}\left(\mathbf{x}|\hat{\theta}\right)}_{\text{maximum likelihood } M}$$

- in (1) the target $f(\theta|\mathbf{x})$ is proportional to prior $f(\theta)$ times the likelihood $\mathrm{L}(\mathbf{x}|\theta)$ as usual;

- in (2) it is lower or equal to the same prior times the likelihood evaluated at it maximum point for $\theta$, $\mathrm{L}\left(\mathbf{x}|\hat{\theta}\right)$ (that is the likelihood of our maximum likelihood estimate). At this point, then, the idea is one can take the maximum likelihood as the constant $M$ of the accept reject method, and the prior as the proposal.

So to generate value from the posterior:

**TODO**: da rivedere

1. we first solve the maximum likelihood problem to obtain $\mathrm{L}\left(\mathbf{x}|\hat{\theta}\right)$ for our ml estimate; so for a Bernoulli distribution we should find theta which maximizes the likelihood of a bernoulli distribution (and for the bernoulli distribution its the mean, look below) and calculate the likelihood of the sample with it. The found values constitutes $M$, a number

2. then we start generating values from a proposal (eg Lognormal)

3. compute the ratio between the target evaluated at the value we have drawn, divided by $M$ times the probability of the proposal with that value, and we accept or reject according to this ratio. Accept reject algorithm can be used even if we don't know the normalization constant (it works) and this is the case

**Maximum likelihood estimation** Under simple sampling with replacement $X_1, \ldots, X_n$ are iid; then joint density is equal to product of marginal density that is:

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

**Esempio 1.4.3.** Let $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta)$ be iid rvs then

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i} \cdot (1-\theta)^{n-\sum x_i}$$

Here for this distribution, we don't have normalization constant, only the kernel.
Now for maximization we set first derivative with respect to $\theta = 0$; maximizing directly the likelihood (find its derivative) could be complicate so in most problem work with log likelihood simplify from a math standpoint. Therefore taking logs of both members we get:

$$\ell(\theta) = \sum x_i \log(\theta) + \left(n - \sum x_i\right) \log(1-\theta)$$

And finally

$$0 = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{\sum x_i}{\theta} + \frac{n - \sum x_i}{1-\theta}(-1)$$

so solve for $\theta$

$$\frac{\sum x_i}{\theta} = \frac{n - \sum x_i}{1-\theta}$$
$$(1-\theta)\sum x_i = \theta(n - \sum x_i)$$
$$\sum x_i - \theta \sum x_i = n\theta - \theta \sum x_i$$
$$\sum x_i = n\theta$$

therefore

$$\hat{\theta} = T_n(\theta) = \frac{\sum x_i}{n} = \bar{x}$$

*Osservazione* 13. We have seen bayesian estimation and likelihood estimation of a bernoulli distribution with two different priors

**Teorema 1.4.1** (Relation between direct and inverse problems (frequentist and Bayes approach))**.** *When $n \to +\infty$, likelihood and Bayesian estimation become equivalent*

*Dimostrazione.* When $n$ increases the importance of the prior becomes negligible so likelihood and posterior become similar □

**Esempio 1.4.4.** Es see the parameters of the posterior of the previous examples $\theta|\mathbf{x} \sim \text{Beta}\,(a + \sum x_i, b + n - \sum x_i)$. When $n \to +\infty$ $\sum x_i$ increases a lot, as well as $n - \sum x_i$ increases: the effect of the prior (characterized by $a$ and $b$ become small compared to the contribution of the likelihood). This happen all the times.

*Osservazione importante* 16. So we can argue likelihood bayesina, but when $n$ increases it doesnt matter, they go in the same direction.

## 1.5 Property of maximum likelihood estimators

- invariance
- efficiency

### 1.5.1 Invariance

**Definizione 1.5.1** (Invariance)**.** If $T_n$ is a maximum likelihood estimator for $\theta$, then *any* function $\mathcal{T}$ of the estimator $T_n$, $\mathcal{T}(T_n)$, is the maximum likelihood estimator for $\mathcal{T}(\theta)$.

**Esempio 1.5.1.** $\log \bar{x}$ is the mle of $\log \theta$ in the previous example.

### 1.5.2 Efficiency

#### 1.5.2.1 Fisher information

*Osservazione importante* 17. In order to talk about efficiency we need to talk about an important quantity we have in statistics which is the **Fisher information**. In different book there are different notations for this concept.
There are *three* different definitions.

**Definizione 1.5.2** (Simple Fisher information)**.** It's the second derivative with respect to $\theta$ of the loglikelihood, with changed sign, that can be observed in my sample

$$i(\theta) = i_n(\theta) = -l^{''}(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta)$$

it's also denoted as $i_n(\theta)$ to emphasize we have a (it's calculated on a) sample size of dimensionality $n$.

**Definizione 1.5.3** (Expected Fisher information)**.** Its the expected value for the previous one:

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$$

It's a teoretical quantity that could be observed having all the sample.
If the observation are iid it's denoted also in some book as

$$I_n(\theta) = nI_1(\theta)$$

here again to stress the dimensionality of sample size we would compute the expected value; the equation above state that if obs are iid, the information coming from a sample of dimensionality $n$ is $n$ time the information coming from a sample of dimensionality 1

*Osservazione importante* 18. It's possible to prove that $I(\theta)$ is also equal to the expectation of the first derivative of the log likelihood squared:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2\right]$$

*Osservazione importante* 19. Therefore it's possible to prove that the square of the first derivative of the log likelihood is always equal to minus the second derivative of the log likelihood

$$\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2 = -\frac{\partial^2}{\partial^2\theta}\ell(\theta) \tag{1.5}$$

**Definizione 1.5.4** (Observed Fisher information)**.** Differently from previous which are function of theta and we don't explicit the value of theta, here is a value/evaluated : it minus the second derivative of the log likelihood evaluated at the estimate of theta:

$$i(\hat{\theta}_n) = -l''(\theta)\big|_{\theta=\hat{\theta}_n}$$

We take the first information and we evaluate it for a value, which is the estimate of theta and so we have a value.

**Esempio 1.5.2.** Consider $X_i \sim \mathrm{Exp}\,(\theta)$ with iid obs, $i = 1, \ldots, n$. We want to write the likelihood and then the loglikelihood (to compute the second derivatibe):

$$\mathrm{L}\,(\theta) = \prod_{i=1}^{n} \theta \cdot e^{-\theta x_i} = \theta^n \cdot e^{-\theta \sum_i x_i}$$

$$\ell(\theta) = n \cdot \log\theta - \theta \cdot \sum_{i=1}^{n} x_i$$

Going with the derivatives

$$\frac{\partial^1}{\partial\theta}(\ell(\theta)) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$
$$\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = -\frac{n}{\theta^2}$$

Starting from the first derivative one can obtain the maximum likelihood estimator of $\theta$ equating it to 0:

$$0 = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$
$$\theta \sum_{i=1}^{n} x_i = n$$
$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Now we start deriving the simple Fisher information:

$$i(\theta) = -\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = \frac{n}{\theta^2}$$

It's a function of $\theta$: if we change theta we have a different values for the information. The information can also be algebrically rewritten as product of $i_1(\theta)$ as follows:

$$i_1(\theta) = \frac{1}{\theta^2}$$
$$i(\theta) = n \cdot i_1(\theta)$$

Now, for the second definition:

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[\frac{n}{\theta^2}\right] \overset{(1)}{=} \frac{n}{\theta^2}$$

in (1) expectation of a constant since we don't have $x$ (otherwise we should compute the expectation). So in this case definition 1 is equivalent to definition 2, that is $i(\theta) = I(\theta)$ (this is not always the case).
Before the third definition we check that we get the same $I(\theta) = \frac{n}{\theta^2}$ by using the definition

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}(\ell(\theta))\right)^2\right]$$

Since we have already computed the first derivative:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}(\ell(\theta))\right)^2\right] = \mathbb{E}\left[\left(\frac{n}{\theta} - \sum x_i\right)^2\right]$$

Now we assume that $n = 1$ and then we will compute $I(\theta) = n \cdot I_1(\theta)$. To obtain $I_1(\theta)$ the information of theta from a single random variable $X$ (of the above

$n$) we adapt the information formula using first derivative:

$$I_1(\theta) = \mathbb{E}\left[\left(\frac{1}{\theta} - X\right)^2\right] = \mathbb{E}\left[\frac{1}{\theta^2} + X^2 - 2\frac{X}{\theta}\right] = \frac{1}{\theta^2} + \mathbb{E}\left[X^2\right] - \frac{2}{\theta}\mathbb{E}\left[X\right]$$
$$\overset{(1)}{=} \frac{1}{\theta^2} + \frac{2}{\theta^2} - \frac{2}{\theta^2} = \frac{1}{\theta^2}$$

where in (1) we remembered the first moment of an $X \sim \text{Exp}(\theta)$ is $\mathbb{E}[X] = \frac{1}{\theta}$ while the second moment $\mathbb{E}\left[X^2\right] = \frac{2}{\theta^2}$.
So even using this first derivative based formula we get the same result as before:

$$nI_1(\theta) = \frac{n}{\theta^2} = I(\theta)$$

Finally the observed information (evaluated at the mle estimate $\hat{\theta} = \frac{1}{\bar{x}}$ which was derived previously) is:

$$i(\hat{\theta}) = \frac{n}{\hat{\theta}^2} = \frac{n}{\left(\frac{1}{\bar{x}}\right)^2} = n\bar{x}^2$$

**Esempio 1.5.3** (Bernoulli distribution)**.** Let $X_i \sim \text{Bern}(\theta)$ iid rvs; the loglikelihood of the sample is

$$\ell(\theta) = \sum_i x_i \cdot (\log\theta) + (n - \sum x_i) \cdot \log(1-\theta)$$

The maximum likelihood estimator $\hat{\theta}_n = \bar{x}$. The first and second derivatives of loglikelihood are:

$$\frac{\partial^1}{\partial\theta}(\ell(\theta)) = \frac{\sum x_i}{\theta} + \frac{n - \sum x_i}{1-\theta}(-1)$$
$$\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = -\frac{\sum x_i}{\theta^2} - \frac{n - \sum x_i}{(1-\theta)^2}$$

The simple information is:

$$i(\theta) = -\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = \frac{\sum x_i}{\theta^2} + \frac{n - \sum x_i}{(1-\theta)^2}$$

The expected information is:

$$I(\theta) = n \cdot I_1(\theta) = n \cdot \mathbb{E}\left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}\right]$$
$$= n \cdot \left[\frac{\theta}{\theta^2} + \frac{1}{(1-\theta)^2}(1-\theta)\right] = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$$

The observed information is:

$$i(\hat{\theta}_n) = \frac{\sum x_i}{\hat{\theta}^2} + \frac{n - \sum x_i}{(1-\hat{\theta})^2} = \frac{n\bar{x}}{\bar{x}^2} + \frac{n - n\bar{x}}{(1-\bar{x})^2} = \ldots = \frac{n}{\bar{x}(1-\bar{x})}$$

#### 1.5.2.2 Rao-Cramer theorem and efficiency

*Osservazione importante* 20. Back to efficiency, we needed fisher information for one of the most important theorem regarding efficiency of maximum likelihood estimators. This theorem gives a lower bound for the variance of an estimator $\mathrm{Var}[T_n]$ under regularity conditions.

**Teorema 1.5.1** (Rao-Cramer Theorem)**.** *Assume $T_n$ is an unbiased estimator for $\theta$ (or for a transformation $\tau(\theta)$: it doesn't matter because of invariance property), so $\mathbb{E}[T_n] = \tau\theta$.*
*Supposing that (*regularity conditions*):*

1. *the domain or support $D_X$ (set of values for $X$), does not depend on $\theta$;*
2. *the likelihood $\mathrm{L}(\theta)$ have first and second derivatives (not infinite etc);*
3. *there should be exchangability between integral and derivative*

$$\frac{\partial}{\partial\theta}\int f(\mathbf{x}|\theta)\,\mathrm{d}x = \int \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)\,\mathrm{d}x$$

   *This is regularity condition needed for math development of the theorem (this situation holds most of the time)*

4. *we should have that the expected information is positive but finite that is, using the formula based on the first derivative:*

$$0 \leq \mathbb{E}\left[\left(\frac{\partial^1}{\partial\theta}(\ell(\theta))\right)^2\right] < +\infty, \quad \forall\theta \in \boldsymbol{\Theta}$$

*Under these condition, not only the maximum likelihood estimator is unbiased (that is $\mathbb{E}[T_n] = \tau(\theta)$) but there is a lower bound for its variance equal to the square of first derivative of $\tau$ over the expected Fisher information:*

$$\mathrm{Var}[T_n] \geq \frac{[\tau'(\theta)]^2}{I(\theta)} = \frac{[\tau'(\theta)]^2}{\mathbb{E}[-l''(\theta)]} = \frac{[\tau'(\theta)]^2}{\mathbb{E}[l'(\theta]^2} \tag{1.6}$$

*Furthermore in 1.6 the equality holds (so the variance is equal to the ratio) if and only if we have that the first derivative of the loglikelihood can be written in the product of two parts, that is:*

$$\frac{\partial^1}{\partial\theta}(\ell(\theta)) = k(\theta)\cdot(T_n - \tau(\theta))$$

*with $k(\theta)$ just a generic function of $\theta$ times the difference between the estimator and the quantity to be estimated.*
*In this case we say the estimator $T_n$ is* UMVUE*, uniformly, miniumum variance, unbiased estimator and we say it's* fully efficient*.*

**Esempio 1.5.4.** If $T_n$ is

- an unbiased estimator for $\theta$ (that is $\tau$ is the identity function) then $\tau(\theta) = \theta$ and $\tau'(\theta) = 1$, so at the numerator one gets 1, so in this case one have that

$$\mathrm{Var}[T_n] \geq \frac{1}{I(\theta)}$$

- an unbiased estimator for $\log\theta$ then $\tau(\theta) = \log\theta$ and $\tau'(\theta) = \frac{1}{\theta}$ and at the numerator one has $\frac{1}{\theta^2}$

**Esempio 1.5.5.** A situation where condition 1 does not hold is $Unif 0, \theta$ with $\theta$ the parameter of interest; $\theta$ is the domain upper bound.

**Definizione 1.5.5** (Full efficiency)**.** When $\text{Var}[T_n] = \frac{\left[\tau'(\theta)\right]^2}{I(\theta)}$, we say $T_n$ is fully efficient.

#### 1.5.2.3 Examples

**Esempio 1.5.6.** An example where regularity conditions are satisfied, here everything is regular.
We want to show that $\bar{x}$ (sample mean) is fully efficient for $\mu$ if the sample $(x_1, \ldots, x_n)$ comes from a gaussian distribution $\text{N}\left(\mu, \sigma^2\right)$.
We start by computing the expected information:

$$I(\theta) = nI_1(\theta)$$

The loglikelihood function, for a sample of $n = 1$ considered $I_1$, is logarithm of the density of a normal (we don't have anymore the product). So given the density:

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

we have the loglikelihood for $n = 1$ being:

$$\log f(x|\mu, \sigma^2) = \log\frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2} = -\log(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}$$

And its first derivative

$$\frac{\partial}{\partial\mu}\log f(x|\mu, \sigma^2) = -2\frac{(x-\mu)}{2\sigma^2}(-1) = \frac{(x-\mu)}{\sigma^2}$$

The expected fisher information (for $n = 1$) is

$$I_1(\mu) = \mathbb{E}\left[\left(\frac{\partial}{\partial\mu}\log f(x|\mu, \sigma^2)\right)^2\right] = \mathbb{E}\left[\frac{(x-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^4}\mathbb{E}\left[(x-\mu)^2\right] \overset{(1)}{=} \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

where in (1), we have $\mathbb{E}\left[(x-\mu)^2\right] = \sigma^2$ by definition for a gaussian. Finally the expected information for $n$ observation is:

$$I(\mu) = n \cdot I_1(\mu) = \frac{n}{\sigma^2}$$

Who is the lower bound of the variance for our estimator? Here we us $\bar{x}$ to estimate $\mu$ so $\tau(\mu) = \mu$ (identity function)

$$\text{Var}[\bar{x}] \geq \frac{1}{I(\mu)} = \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n}$$

Now we are interested in the full efficiency for the estimator on $\mu$ not on $\sigma^2$. We have computed the variance of sample mean for iid sample previously and it's

$$\operatorname{Var}[\bar{x}] = \operatorname{Var}\left[\frac{\sum X_i}{n}\right] = \frac{1}{n^2}\operatorname{Var}\left[\sum X_i\right] = \frac{\sigma^2}{n}$$

Since in this case the variance is exactly equal to the lower possible bound, therefore $\bar{x}$ is fully efficient for $\mu$.

**Esempio 1.5.7.** In this example one regularity condition is not satisfied and things doesn't work as expected.
Suppose $X \sim \text{Unif}(0, \theta)$ with $\theta > 0$ and we are interested in estimating $\theta$. The density of $X$ for uniform

$$f(x, \theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$$

Let's find the expected information for the uniform distribution. Again it's convenient specify $n = 1$ and go for $I(\theta) = nI_1(\theta)$. If $n = 1$, the loglikelihood is $\log f(x, \theta) = -\log\theta$ therefore

$$I_1(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] = \mathbb{E}\left[-\frac{1}{\theta}^2\right] = \frac{1}{\theta^2}$$

so the expected information is

$$I(\theta) = n \cdot \frac{1}{\theta^2} = \frac{n}{\theta^2}$$

Imagine now we have a sample of $n$ observation we know have a likelihood with:

$$\mathrm{L}(\theta) = \prod_{i=1}^{n}\frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n$$

and according to RC theorem, the variance of an unbiased estimator for $\theta$ shoud have

$$\operatorname{Var}[T_n] \geq \frac{1}{I(\theta)} = \frac{\theta^2}{n}$$

Now we take any unbiased estimator for $\theta$; the estimator is

$$T_n = X_{(n)} \cdot \frac{n+1}{n}$$

this is unbiased: let's check. Remembering that the density of the maximum formula and applying it to the uniform:

$$f_{(n)}(x) = n \cdot F(x)^{n-1} \cdot f(x) = n\Big(\frac{x}{\theta}\Big)^{n-1}\frac{1}{\theta}$$

Now the expected value of the maximum is

$$\begin{aligned}\mathbb{E}\left[X_{(n)}\right] &= \int_0^\theta x \cdot f_{(n)}(x)\,\mathrm{d}x = \int_0^\theta x \cdot n \cdot \Big(\frac{x}{\theta}\Big)^{n-1} \cdot \frac{1}{\theta}\,\mathrm{d}x = \int_0^\theta n\Big(\frac{x}{\theta}\Big)^n\,\mathrm{d}x \\ &= \frac{n}{\theta^n} \cdot \left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \theta \cdot \frac{n}{n+1}\end{aligned}$$

if this is expectation of the maximum, its clear that our estimator $T_n$ is unbiased, because it fix the $n/(n+1)$ ratio:

$$\mathbb{E}[T_n] = \mathbb{E}\left[\frac{(n+1)}{n} X_{(n)}\right] = \frac{(n+1)}{n}\mathbb{E}\left[X_{(n)}\right] = \frac{n+1}{n}\frac{n}{n+1}\theta = \theta$$

So we have an unbiased estimator; according to Rao Cramer the variance lower bound of any unbiased estimator is $\geq \frac{\theta^2}{n}$. Now we want to compute the variance to check if it's equal to the lower bound and to say the estimator is fully efficient. We can compute the second moment of the maximum

$$\begin{aligned}\mathbb{E}\left[X_{(n)}^2\right] &= \int_0^\theta x^2 \cdot n \cdot \left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta}\,\mathrm{d}x = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,\mathrm{d}x = \frac{n}{\theta^2}\cdot\left[\frac{x^{n+2}}{n+2}\right]_0^\theta \\ &= \frac{n}{n+2}\theta^2\end{aligned}$$

To the variance of the maximum is

$$\mathrm{Var}\left[X_{(n)}\right] = \mathbb{E}\left[X_{(n)}^2\right] - \mathbb{E}\left[X_{(n)}\right]^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\right)^2\theta^2 = \ldots = \frac{n\theta^2}{(n+1)^2(n+2)}$$

and the variance of our estimator is

$$\begin{aligned}\mathrm{Var}[T_n] &= \mathrm{Var}\left[\frac{n+1}{n}X_{(n)}\right] = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}\left[X_{(n)}\right] \\ &= \left(\frac{n+1}{n}\right)^2 \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}\end{aligned}$$

So we have an estimator which is unbiased, which has a variance $\frac{\theta^2}{n(n+2)}$ that is lower than the lower bound $\frac{\theta^2}{n}$. This is called *super efficient* or *more than efficient*.
So if not all the regularity conditions are satisfied (here the domain $[0,\theta]$ depends on the parameter of interest) we could find an estimator with variance lower than the lower bound given by Rao-Cramer (if equal the estimator is *fully efficient*). Otherwise if all the regularity cond are satisfied the lower bounds *holds*.

### 1.5.3 Properties of ML estimators

**Teorema 1.5.2.** *1. invariance;*

2. *if an unbiased and fully efficient (variance equal to the lower bound) for $\theta$ exists, then it can be found by maximum likelihood*

3. *under non restrictive conditions maximum likelihood estimators are:*

   (a) *asympototically unbiased: we are not sure to find an unbiased estimator, but we are sure they are asympototically unbiased (eg estimator of the variance for the gaussian distribution is biased having n instead of $(n-1)$ but when n increases this small difference is negligible)*

   (b) *asymptotic efficient: they reach the lower RC bound for variance when $n \to +\infty$*

(c) *weakly consistent* $T_n \xrightarrow{p} \theta$

(d) *they are asymptotically gaussian: if one take* $T_n$ *and rewrite it in the canonical form for the central limit theorem*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2(\theta)\right)$$

*where* $\sigma^2(\theta) = \frac{1}{I(\theta)}$ *so it's the lower bound (this because of the first two)*

(e) $T_n$ *is BAN estimators: best asymptotically normal estimators.*

*Dimostrazione.* Here we proof only that if an unbiased and fully efficient for $\theta$ exists, then it can be found by maximum likelihood.
In Rao Cramer the equality to the lower bound holds when

$$\frac{\partial}{\partial\theta} \log f(x;\theta) = k(\theta) \cdot [T_n - \theta]$$

At the same time in M.L. we solve

$$\frac{\partial}{\partial\theta} \log f(x;\theta) = 0$$

so

$$k(\theta) \cdot [T_n - \theta] = 0$$

implies $T_n$ is M.L. estimator. □

# Capitolo 2

# Computational stat part

*Osservazione* 14. We start to study methods/computational tools to find maximum likelihood, from classical numerical techniquest to the EM algorithm (one of the most used algorithm in statistics).

## 2.1 Optimization techniques for maximum likelihood

**Esempio 2.1.1** (Motivating example)**.** In some cases M.L. estimators cannot be written in closed form. Consider as example the Gamma distribution $X \sim \text{Gamma}\,(()\alpha, \beta)$ with $\alpha > 0$ (shape) and $\beta > 0$ (rate); the density is

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

where

- $\frac{\beta^\alpha}{\Gamma(\alpha)}$ is a normalization constant
- $x^{\alpha-1}e^{-\beta x}$ is the kernel
- $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}\,\mathrm{d}x$
- our parameter of interest are $\boldsymbol{\theta} = \big[\alpha, \beta\big]$

In this case the likelihood and the logklikelihood functions (under iid obs) are:

$$\mathrm{L}\,(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \cdot \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \cdot e^{-\beta \sum_i x_i}$$

$$\ell(\mathbf{x}|\boldsymbol{\theta}) = n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\log x_i - \beta\sum_{i=1}^{n} x_i$$

Now

- the maximum likelihood estimators $\beta$ can be found by setting the first derivative with respect to $\beta$ (also called *score*) equal to zero:

$$\frac{\partial \ell(\mathbf{x}|\boldsymbol{\theta})}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0$$

and this leads, resolving for $\beta$, to a closed-form solution for $\beta$ as function of $\alpha$ (and we don't know $\alpha$)

$$\hat{\beta} = T_n = \frac{n\alpha}{\sum_{i=1}^{n} x_i} = \frac{\alpha}{\bar{x}}$$

- when it comes to $\alpha$ we have several possibilities:

  1. we do the same procedure and hope it don't come out an estimator of $\alpha$ as function of $\beta$; in that case one has the estimator for $\alpha$ and put the estimator for $\alpha$ in the solution for $\beta$ above
  2. a second possibility is that in the process the estimator of $\alpha$ is a function of $\beta$ (as above in switched roles). In this case we could either:
     - set up a *linear equation system*
     - a third possibility is to use/get the *profile loglikelihood*, that is: we substitute the estimator of $\beta$ in the loglikelihood derived above, instead of $\beta$ itself. Then one could compute the first derivative with respect to $\alpha$ (the only parameter remaining)

  For this example we do the same procedure as before computing the derivative of the original loglikelihood with respect to $\alpha$: unfortunately here the ML estimator cannot be obtained in closed form since:

$$\frac{\ell(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha} = n \log \beta - n \frac{1}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha} + \sum_{i=1}^{n} \log x_i = 0$$

  in the loglikelihood we have three terms which depends on $\alpha$ and the middle one $-n \log \Gamma(\alpha)$ when deriving become the complex $-n \frac{1}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha}$. So we don't have an explicit solution for the parameter/we can't isolate alpha: it's inside an integral (the $\Gamma$) and inside a derivative of the $\Gamma$

## 2.1.1 Newton-Raphson algorithm

### 2.1.1.1 The algorithm

This is the simplest solution: the idea is optimize a function, in our case the loglikelihood. The idea is: if one cannot solve/maximize the $\ell(\theta)$ directly one can try to approximate locally by taking a quadratic function. We approximate the loglikelihood by a very good quadratic function of it

The assumptions are that:

1. $\ell(\theta)$ is differentiable (have first derivative)
2. the third derivative of the loglik should be not infinite: $\ell'''(\theta) < \infty$

3. the second derivative should not be null: $\ell''(\theta) \neq 0$

Considering a point $\theta_0 \in \mathbf{\Theta}$: we want to approximate the loglikelihood by a quadratic in this point. Developing the quadratic using Taylor expansion we have:

$$\begin{aligned}\ell(\theta) &\overset{(1)}{=} \ell(\theta_0) + (\theta - \theta_0)\ell'/\theta_0) + \frac{1}{2}(\theta - \theta_0)^2\ell''(\theta_0) + r_2(\theta, \theta_0) \\ &\overset{(2)}{=} \ell(\theta_0) + (\theta - \theta_0)S(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 H(\theta_0) + r_2(\theta, \theta_0)\end{aligned}$$

where in (1) we compacted the rest of the expansion

$$r_2(\theta, \theta_0) = \frac{1}{3!}\frac{\partial^3 \ell(\theta)}{\partial \theta^3}\Big|_{\theta=\theta_0}(\theta - \theta_0)^3 + \dots$$

and in (2) we merely replaced naming by:

- indicating $\ell'(\theta_0)$, the first derivative of the loglikelihood evaluated in $\theta_0$ with $S(\theta_0)$ as *score function* ( first derivative)
- indicating $\ell''(\theta_0)$, the second derivative, as $H(\theta_0)$ meaning *Hessian function*

Ignoring the remainder term $r_2(\theta, \theta_0)$ we have that the function at the point $\theta_0$ is approximated by the following quadratics

$$\ell(\theta) \approx \ell(\theta_0) + (\theta - \theta_0)S(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 H(\theta_0) \tag{2.1}$$

which represents a quadratic approximation of $\ell(\theta)$ at $\theta_0$. Taking the first derivative of with respect to $\theta$ and putting it equal to 0 for maximization (that is we're maximizing an approximation in $\theta_0$) we get

$$\ell'(\theta) \approx S(\theta_0) + (\theta - \theta_0)H(\theta_0)$$

So equating to 0 and solving for $\theta$ leads to a solution which maximizes the log likelihood (its approximation)

$$S(\theta_0) + \theta H(\theta_0) - \theta_0 H(\theta_0) = 0 \implies \theta_1 = \theta_0 - \frac{S(\theta_0)}{H(\theta_0)}$$

from which we derive $\theta_1$ which is a local approximation of $\hat{\theta}$.
If we apply the idea recursively we obtain the general iterative rule

$$\theta_{t+1} = \theta_t - \frac{S(\theta_t)}{H(\theta_t)} \tag{2.2}$$

from which we derive a sequence of values $\{\theta_t\}_{t\in\mathbb{N}}$.
The idea is approximate a function at a point, obtain a solution, use the solution as a new starting point and continue up to a stop.

#### 2.1.1.2 Stopping criteria

We need a criteria to stop the sequence and obtain our estimate of $\theta$. Ideally we want to stop where the point we get $\theta_t$ is near enough to the maximum likelihood point $\hat{\theta}$, that is $\left|\theta_t - \hat{\theta}\right| \leq \varepsilon$ with very small $\varepsilon$ (eg $\varepsilon = 0.0001$). In practice we don't know $\hat{\theta}$ so we have several alternative stopping criteria (not involving it):

- we stop when the absolute difference between estimates at step is lower than a bound $|\theta_{t+1} - \theta_t| \leq \varepsilon$
- we stop with relative difference $\frac{|\theta_{t+1} - \theta_t|}{|\theta_t|} \leq \varepsilon$ which is better than the previous one since its robust for theta scale: if theta is small or high they are handled appropriately as well by going on relative variations
- we stop with low change of loglikelihood $|\ell(\theta_{t+1}) - \ell(\theta_t)| \leq \varepsilon$
- we stop with low score function $|S(\theta_{t+1})| \leq \varepsilon$ since at maximum point should be equal to 0

The last three are better (using one or the other depends on the coding), the first one is to be avoided.

#### 2.1.1.3 Conditions for convergence

We have a problem: by Newton-Raphson one is not perfectly sure that we converge/end the algorithm, and therefore NR algorithm is not guaranteed to reach the global maximum.
Convergence might depend on the starting point $\theta_0$ and if there are several local maximum: if we start near a local maximum that isn't a global maximum, the algorithm could converge to the local maximum not the global maximum. Other times one could diverge to $+\infty$ or $-\infty$ or go outside the support of the parameter etc.

We should have therefore some math conditions to be sure that NR converges. These are given by a theorem

**Teorema 2.1.1.** *If*

- $f(\theta)$ *has the first two derivatives*
- $f$ *it's concave, that is* $H(\theta) < 0, \forall \theta \in \mathbf{\Theta}$

*then we are sure that, by NR algorithm, we'll have that:*

$$\lim_{t \to \infty} \theta_t = \hat{\theta}, \forall \theta_0 \in \mathbf{\Theta}$$

*that is the algorithm converges from all the starting points.*

### 2.1.2 Quasi-Newton algorithms

NR algorithm is the most popular but it is a special case of the *family* of methods (called quasi-newton algorithm) which implement an iterative solution searching following the step prescribed by the equation below

$$\theta_{t+1} = \theta_t - \alpha \frac{S(\theta_t)}{m(\theta_t)}$$

where

- $\alpha$ is a number you fix;
- $m$ is another function/quantity one fixes;
- $S$ is the score function

In this family it is possible to prove that the loglikelihood is increasing step after step, that is $\ell(\theta_{t+1}) > \ell(\theta_t)$, provided that $m(\theta_t) < 0$ and $\alpha$ is sufficiently small. Some cases of the QN algorithms are the following:

- if $\alpha = 1$ and $m(\theta_t) = H(\theta_t)$ we have the *Newton-Raphson algorithm*;
- if $\alpha > 0$ (but better small) and $m(\theta_t) = -1$ we have the *gradient descent algorithm*

  $$\theta_{t+1} = \theta_t + \alpha S(\theta_t)$$

  It's very used in neural networks because it's very simple, but its slower than NR: not using info provided by the hessian of the loglikelihood i don't need to compute the hessian all the time but use less information and it takes more iteration to converge.
  In the univariate case we have two directions: right step is the score is positive or left step is the score is negative.
- if $\alpha = 1$ and

  $$m(\theta_t) = \frac{S(\theta_t) - S(\theta_{t-1})}{\theta_t - \theta_{t-1}}$$

  we have the *secant method*. This formula for $m$ resemble the definition of the derivative (pendenza della retta passante tra $(\theta_{t-1}, S(\theta_{t-1}))$ e $(\theta_t, S(\theta_t))$. So the secant method is basically the NR when i approximate the hessian (second derivative) with the slope of the first derivative (which is conceptually close).
  So the secant method is used when one cannot compute the hessian but want an approximation nonetheless, which is better than $-1$ implemented in gradient descent: so we expect that the secant method will be faster than the gradient method (it uses more information)
- if $\alpha > 0$ (but better small) and if $m(\theta_t) = -I(\theta_t)$ (minus expected Fisher information), we have the *Fisher scoring method*, very used by statistician:

  $$\theta_{t+1} = \theta_t + \alpha \frac{S(\theta_t)}{I(\theta_t)}$$

**Convergence order/speed**

*Osservazione* 15. It represents the 'average' speed of convergence of a method to the optimal point.

**Definizione 2.1.1** (Convergence order)**.** An algorithm has order of convergence $\beta$ if

$$\lim_{t\to\infty} \frac{\left|\theta_{t+1} - \hat{\theta}\right|}{\left|\theta_t - \hat{\theta}\right|^{\beta}} = c$$

where $c \neq 0$ is a constant and $\beta > 0$.

*Osservazione* 16. The higher $\beta$, the faster the convergence.

*Osservazione importante* 21. Regarding convergence order we have that for

- NR algorithm, $\beta = 2$ (quadratic order)
- Gradient descent algorithm, $\beta = 1$ (linear order)
- Secant method, $\beta = 1.62$ (superlinear order)
- Fisher scoring, $\beta = 2$ (quadratic order)

**Extension to the $p$-dimensional case** Most of the QN algorithms (not all) can be generalized to the case in which $\theta$ is a $p$-dimensional vector and we want to estimate several parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_p \end{bmatrix}$$

In this case the score is a vector containing the first $p$ partial derivatives

$$S(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \dots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}$$

The Hessian is $p \times p$ matrix of the partial second derivatives:

$$H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1^2} & \cdots & \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \cdots & \cdots & \cdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p^2} \end{bmatrix}$$

**Esempio 2.1.2** (Multivariate problem with gaussian)**.** Let $x_1, \dots x_n$ be a iid sample from the gaussian density $\mathrm{N}(\theta_1, \theta_2)$. We are interested in estimating $\theta_1$ and $\theta_2$. We are interested in computing the score vector and the hessian matrix. The density function is

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}\frac{(x_i-\theta_1)^2}{\theta_2}}$$

The loglikelihood $\ell(\theta_1, \theta_2)$ is:

$$\ell(\theta_1, \theta_2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

**DO**: check

The score (vector of partial first derivative of the loglikelihood):

$$S(\theta_1, \theta_2) = \begin{bmatrix} \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2} \\ \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2} \end{bmatrix}$$

The Hessian:

**TODO**: check

$$H(\theta_1, \theta_2) = \begin{bmatrix} -\frac{n}{\theta_2} & -\frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} \\ -\frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} & -\frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^3} - \frac{n}{2\theta_2^2} \end{bmatrix}$$

The expected information (used in fisher-scoring) which is the expected value of minus the second-derivative/Hessian in the multivariate case:

$$I(\theta_1, \theta_2) = \mathbb{E}\left[-H(\theta_1, \theta_2)\right] = \mathbb{E}\left[\begin{bmatrix} \frac{n}{\theta_2} & \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} \\ \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} & \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^3} + \frac{n}{2\theta_2^2} \end{bmatrix}\right]$$
$$\overset{(1)}{=} \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{\theta_2^2} - \frac{n}{2\theta_2^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{bmatrix}$$

where in (1):

- in position (1,1) we have the expected value of a constant
- in (1,2) and (2,1) the expected value is 0 because the expected value of the sum (numerator) is the sum of expected values and it semplifies with $\mathbb{E}[x_i] - \mathbb{E}[\theta_1] = \theta_1 - \theta_1$ (since $\theta_1$ is the mean).
- in $(2,2)$ the expected value $\mathbb{E}\left[(x_i - \theta_1)^2\right]$ at the numerator is by definition $\sigma^2$ that in this notation is $\theta_2$, so considering iid its the expected value of the sum is $n$ times $\theta_2$; the second term again is a constant

In this case we have both the Hessian and the expected information; so we could use newton raphson (having the hessian) the fisher scoring (having the expected information) or the gradient descent as well (having the score).
The only method that can't be used here is the secant: in a multivariate problem, the approximation of the hessian cannot be done in this manner (because it's multivariate, it's a matrix not a function).
We have seen that the inverse of the expected information is the lower bound for rao cramer. Here in the multivariate setup rao cramer works as well and is the inverse of the information matrix gives us the lower bound of variances according to RC theorem

$$I^{-1}(\theta_1, \theta_2) = \begin{bmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{bmatrix}$$

Here having a diagonal matrix as information matrix, the inverse is easily obtained by inverting the diagonal terms.
Therefore:

- the lower bound for the variance of an estimator for $\theta_1$ is $\theta_2/n$ (actually $\bar{x}$ has this variance and therefore it is fully efficient)
- the lower bound for the variance of an estimator for $\theta_2$ is $2\theta_2^2/n$
- the ml estimator for $\theta_2$ is the uncorrected variance $s(x) = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n}$ but it is biased (as shown previously). So we here cannot apply the lower bound given by rao cramer theorem (being the estimator biased)
- the sample variance $s(x) = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n-1}$ is unbiased but it has variance $\frac{2\theta_2^2}{n-1}$ which is larger than the lower bound (so it's not fully efficient)
- so we can conclude that an unbiased and fully efficient estimator for $\theta_2$ does not exist

*Osservazione importante* 22 (Extension to the $p$-dimensional case). Most algorithm can be generalized in this way. The extention of the secant algorithm is instead challenging and problematic.
For these situations other proposals exist. In particular, the BFCS algorithm works well in the p-dimensional context. BFGS stands for Broyden, Fletcher, Goldfarb and Shannon (the authors). In R it is implemented in the command `optim`.

### 2.1.3 Exercises oilspills

*Osservazione* 17. In this section two exercises on same dataset. The data regards crude oil spills of at least 1000 barrels from tankers in US waters during 1974-1999. Columns are:

- `year` considered (1974-1999)
- the count of `spills` (denoted by y)
- `importexport` (x) , the estimated amount of oil shipped through US waters as part of US import/export operations (adjusted for spillage in international or foreign waters)
- `domestic` (z), the amount of oil shipped through US waters during domestic shipments

Oil shipments are measured in billions of barrels of oil (Bbbl).

```
##   year spills importexport domestic
## 1 1974      2        0.720     0.22
## 2 1975      5        0.850     0.17
## 3 1976      3        1.120     0.15
## 4 1977      3        1.345     0.20
## 5 1978      1        1.290     0.59
## 6 1979      5        1.260     0.64
```

**Esempio 2.1.3** (First part - Univariate case)**.** Here we assume that the variable spills, $y_i$, follows a Poisson process with parameter $\lambda_i = \theta_1 \cdot x_i$ (in this case lambda is not fixed, but the number of accidents can depend on the number of operations) where $i = 1, \ldots, 26$ . We are interested in estimating $\theta_1$:

1. write the likelihood function and the log-likelihood function:

$$\begin{aligned}
\mathrm{L}(\theta_1) &= \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} = \prod_{i=1}^{n} \frac{(\theta_1 x_i)^{y_i} e^{-\theta_1 x_i}}{y_i!} \\
\ell(\theta_1) &= \sum_{i=1}^{n} y_i \log(\theta_1 x_i) - \theta_1 x_i - \log(y_i!) \\
\mathbb{E}[Y_i] &= \lambda_i = \theta_1 x_1
\end{aligned}$$

2. compute the score and find the maximum likelihood estimator for $\theta_1$.

$$\frac{\partial}{\partial \theta_1} \ell(\theta_1) = S(\theta_1) = \sum_{i=1}^{n} \left[ y_i \frac{1}{\theta_1 x_i} x_i - x_i \right] = \sum_{i=1}^{n} \frac{y_i}{\theta_1} - \sum_{i=1}^{n} x_i$$

therefore our estimator can be obtaining equating above to 0 and solving for $\theta$ that is

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

so our estimate is computed as follows

```
y = oilspills$spills
x = oilspills$importexport
z = oilspills$domestic

### maximum likelihood estimation
(hattheta1 = sum(y) / sum(x))

## [1] 1.715778
```

in this case it would be not necessary for us to use numerical methods because a close formula for the estimator exists. However in the following we pretend that this is not the case and try to get the same estimate by the methods presented before

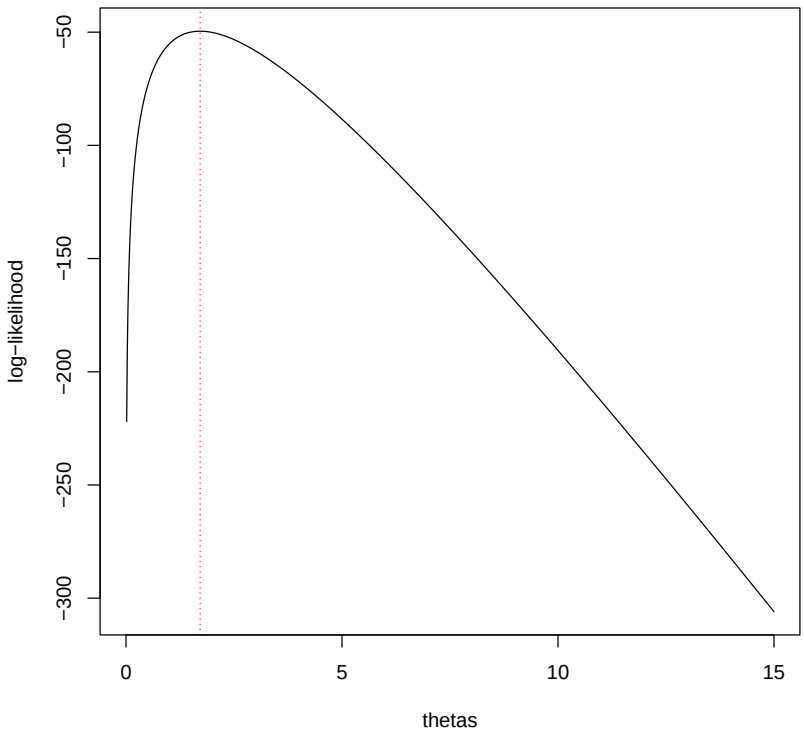
3. compute the Hessian. We have

$$H(\theta_1) = \frac{\partial}{\partial \theta_1} S(\theta_1) = \frac{\partial}{\partial \theta_1} \left( \sum_{i=1}^{n} \frac{y_i}{\theta_1} - x_i \right) = -\sum_{i=1}^{n} \frac{y_i}{\theta_1^2}$$

4. apply the Newton-Raphson algorithm and with starting value 0.6 and then with starting value 4. Compare the two solutions

```
## loglikelihood function, score and hessian as derived above
## loglik <- function(theta1, y, x){
loglik <- function(theta1){
    a = sum(y * log(theta1 * x))
    b = theta1 * sum(x)
```

```
    c = sum(log(factorial(y)))
    return(a-b-c)
}
S <- function(theta1) sum(y)/theta1 - sum(x)
H <- function(theta1) -sum(y)/(theta1^2)

## loglikelihood plotting for several possible values of theta
thetas <- seq(0, 15, length.out = 1000)
l = 0  # computed loglikelihood for the values of theta
for (i in 1:1000){
    l[i] = loglik(thetas[i])
}
plot(thetas, l,
     ylab = 'log-likelihood',
     type = "l", xlim = c(0,15))
## add the mle for check
abline(v = hattheta1, col = 'red', lty = 'dotted')
```




```
## the loglikelihood is regular/concave, with 1 mode, so newton
## raphson should work
nr <- function(theta_0){
```

```
    delta = 100       # value for stopping rules
    epsilon = 0.0001 # tolerance level
    it = 0            # iteration counter
    theta.all = theta_0 # vector with all thetas estimates
    theta = theta_0     # current value of theta considered
    while (delta > epsilon) {
        it = it + 1
        theta = theta - S(theta) / H(theta) # new theta estimate
        theta.all = c(theta.all, theta)  # save it
        delta = abs(S(theta)) # score criteria implemented
        print(paste("it=", it, " theta=", theta))
    }
    return(invisible(list(theta = theta.all, it = it)))
}

## starting from 0.6, in 5 iteration we have basically the same
## estimate
out06 = nr(theta_0 = 0.6)

## [1] "it= 1  theta= 0.990182608695652"
## [1] "it= 2  theta= 1.4089266204895"
## [1] "it= 3  theta= 1.6609001998968"
## [1] "it= 4  theta= 1.71402249113269"
## [1] "it= 5  theta= 1.71577589935419"

## instead here, it doesnt converge it diverges, we don't have
## solution
## out10 = nr(theta_0 = 4)
```

So not all the starting point for $\theta_0$ are good, some starting point can be bad. Why we have this result: from the math point of view there's no reason, the likelihood plot is very nice. here the problem is related to the domain space of the poisson distribution (the parameter lambda of a poisson should be positive, but in some iteration the theta computed using the iterative rule is negative).
We are not sure that it will work

```
## while with theta =3 its ok
out06 = nr(theta_0 = 3)

## [1] "it= 1  theta= 0.754565217391304"
## [1] "it= 2  theta= 1.17728752238637"
## [1] "it= 3  theta= 1.54677464353514"
## [1] "it= 4  theta= 1.69913099791267"
## [1] "it= 5  theta= 1.71561618648414"
## [1] "it= 6  theta= 1.71577767968697"
```

So before givin up saying i made mistakes try with different starting values for $\theta_0$.

**Esempio 2.1.4** (Second part – Multivariate case)**.** In this case we assume that both variables ($x$ and $z$) affect the outcome in this way

$$\lambda_i = \theta_1 x_i + \theta_2 z_i, \quad i = 1, \dots, 26$$

1. Write the likelihood and log-likelihood function. We have:

$$\begin{aligned} \mathrm{L}(\theta_1, \theta_2) &= \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \\ \ell(\theta_1, \theta_2) &= \sum_{i=1}^{n} (y_i \log \lambda_i - \lambda_i - \log y_i!) \\ &= \sum_{i=1}^{n} (y_i \log(\theta_1 x_i + \theta_2 z_i) - (\theta_1 x_i + \theta_2 z_i) - \log y_i!) \end{aligned}$$

   so $\mathbb{E}[Y_i] = \lambda_i = \theta_1 x_i + \theta_2 z_i$.

```
loglik <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    sum(y * log(lambda) - lambda - log(factorial(y)))
}

## check
loglik(c(1,1))

## [1] -48.06826
```

2. Write the score vector. The components are

$$\mathbf{S}(\theta_1, \theta_2) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ell(\theta_1, \theta_2) \\ \frac{\partial}{\partial \theta_2} \ell(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \frac{y_i x_i}{\theta_1 x_i + \theta_2 z_i} - \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} \frac{y_i z_i}{\theta_1 x_i + \theta_2 z_i} - \sum_{i=1}^{n} z_i \end{bmatrix}$$

   So

```
S <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    c(sum(y*x/lambda) - sum(x),
      sum(y*z/lambda) - sum(z))
}

# Score for starting point
S(c(1,1))

## [1] 1.1128229 0.3871771

# not so close to 0 for the starting thetas c(1,1), especially for the
# first: so probably the first theta will change more than the second
```

Observe that this time if we equate the two partial derivatives to zero we don't get a close solution for the two parameters. The only part of the score vector elemets where $\theta_1$ and $\theta_2$ are included is the denominator; in order to simiplify it we should multiply by the denominator, but the denominator depends on sum. So there is no way to construct a system or substitute or whatever. So there are no close form for the estimator and simple value for maximum likelihood estimates; therefore approximation and numerical methods comes very handy here.

$$\begin{cases} \frac{\partial}{\partial \theta_1}\ell(\theta_1,\theta_2) = 0 \\ \frac{\partial}{\partial \theta_2}\ell(\theta_1,\theta_2) = 0 \end{cases} \not\Longrightarrow \text{closed form solution for } \theta_1, \theta_2$$

3. Write the Hessian matrix:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

with $h_{12} = h_{21}$ we have

$$h_{11} = \frac{\partial^2}{\partial \theta_1^2}\ell(\theta_1,\theta_2) = -\sum_{i=1}^{n} \frac{y_i x_i^2}{(\theta_1 x_i + \theta_2 z_i)^2}$$

$$h_{22} = \frac{\partial^2}{\partial \theta_2^2}\ell(\theta_1,\theta_2) = -\sum_{i=1}^{n} \frac{y_i z_i^2}{(\theta_1 x_i + \theta_2 z_i)^2}$$

$$h_{12} = h_{21} = \frac{\partial^2}{\partial \theta_1 \partial \theta_2}\ell(\theta_1,\theta_2) = -\sum_{i=1}^{n} \frac{y_i x_i z_i}{(\theta_1 x_i + \theta_2 z_i)^2}$$

So

```
H <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    rval = matrix(0, nrow = 2, ncol = 2)
    rval[1,1] = -sum(y * x^2 / lambda^2)
    rval[2,2] = -sum(y * z^2 / lambda^2)
    rval[1,2] = rval[2,1] = -sum(y*x*z/lambda^2)
    rval
}

## Check with the (1,1) starting point
H(c(1,1))

##             [,1]      [,2]
## [1,] -18.315084 -9.607739
## [2,]  -9.607739 -8.469438
```

In this multivariate case Newton Raphson will be applied in this manner

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}^{-1}\mathbf{S}$$

4. Implement the gradient descent algorithm with $\alpha = 0.1$, $\alpha = 0.01$ and $\boldsymbol{\theta}_0 = (1, 1)$ and comment the results.

```
## gradient descent
gd <- function(thetas_0, alpha, tol = 10^-5){
    it = 1                                          # iteration id
    p = length(thetas_0)                            # n of params
    thetas = matrix(thetas_0, nrow = 1, ncol = p)  # matrix with saved res (it x p
    l.it = loglik(thetas_0)                          # loglik of the considered par
    print(c(it, l.it[it], thetas_0))                 #
    delta = 100
    while (delta > tol){
        S.it = S(thetas[it,])                       # score: is only needed for gra
        theta = thetas[it, ] + alpha * S.it         # new values of theta estimated
        thetas = rbind(thetas, theta)               # save them
        l.it = c(l.it, loglik(theta))               # compute the loglik and add to
        delta = abs((l.it[it+1]-l.it[it])/l.it[it])  # compute the check with re
        it = it+1
        print(c(it, l.it[it], theta))               # iteration results
    }
    out = list(likelihood = l.it, theta = thetas)
    return(invisible(out))
}

## with alpha = 0.1 it doesnt work

## gd(c(1,1), alpha = 0.1)

## loglik moves between  -49.7352 and -52.286
## and estimated parameters between
## 0.8088     0.7291
## 1.702     1.304
## the estimates of the parameters oscillates and dont converge. the
## jumps are due to the fact that alpha is too much high and the
## change in the theta after each step is too much over, so it bounces
## from one side to another of the maximum


## diminishing alpha it converges to 1.052 and 1.003 after 12 iteration
## (loglikelihood of the last couple is -48)
gd(c(1,1), alpha = 0.01)

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.000000 -48.055982   1.011128   1.003872
## [1]   3.000000 -48.048641   1.019867   1.006357
## [1]   4.000000 -48.044136   1.026811   1.007815
## [1]   5.000000 -48.041283   1.032393   1.008507
## [1]   6.000000 -48.039403   1.036935   1.008628
## [1]   7.000000 -48.038105   1.040677   1.008322
## [1]   8.000000 -48.037158   1.043800   1.007699
```

```
## [1]   9.000000 -48.036428   1.046443   1.006842
## [1]  10.000000 -48.035836   1.048711   1.005814
## [1]  11.000000 -48.035332   1.050684   1.004664
## [1]  12.000000 -48.034888   1.052425   1.003427

## using lower alpha again it takes much iteration but onverges to
## other values..
gd(c(1,1), alpha = 0.001)

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.000000 -48.066890   1.001113   1.000387
## [1]   3.000000 -48.065581   1.002202   1.000760
## [1]   4.000000 -48.064332   1.003267   1.001120
## [1]   5.000000 -48.063139   1.004309   1.001466
## [1]   6.000000 -48.062000   1.005329   1.001800
## [1]   7.000000 -48.060913   1.006328   1.002121
## [1]   8.000000 -48.059874   1.007305   1.002430
## [1]   9.000000 -48.058882   1.008262   1.002727
## [1]  10.000000 -48.057935   1.009198   1.003012
## [1]  11.000000 -48.057030   1.010115   1.003286
## [1]  12.000000 -48.056164   1.011013   1.003549
## [1]  13.000000 -48.055337   1.011892   1.003801
## [1]  14.000000 -48.054546   1.012753   1.004043
## [1]  15.000000 -48.053790   1.013596   1.004275
## [1]  16.000000 -48.053067   1.014421   1.004497
## [1]  17.000000 -48.052375   1.015230   1.004709
## [1]  18.000000 -48.051714   1.016023   1.004912
## [1]  19.000000 -48.051080   1.016799   1.005106
## [1]  20.000000 -48.050474   1.017560   1.005291
## [1]  21.000000 -48.049894   1.018305   1.005467
## [1]  22.000000 -48.049338   1.019036   1.005635
## [1]  23.000000 -48.048806   1.019752   1.005794
## [1]  24.000000 -48.048296   1.020453   1.005946
## [1]  25.000000 -48.047807   1.021141   1.006089
## [1]  26.000000 -48.047339   1.021815   1.006226
```

5. Implement the Newton-Raphson algorithm and comment the results. Try also with the starting points $\boldsymbol{\theta}_0 = (100, 1)$

```
# newton raphson
nr <- function(thetas_0, tol = 10^-5){
    it = 1
    p = length(thetas_0)
    thetas = matrix(thetas_0, nrow = 1, ncol =p)
    l.it = loglik(thetas_0)
    print(c(it, l.it[it], thetas_0))
    delta = 100
    while (delta>tol)  {
        S.it = S(thetas[it,])
```

```
        H.it = H(thetas[it,])
        theta = c(thetas[it,] - solve(H.it) %*% S.it)
        thetas = rbind(thetas, theta)
        l.it = c(l.it, loglik(theta))
        delta = abs((l.it[it+1] - l.it[it])/l.it[it])
        it = it+1
        print(c(it, l.it[it], theta))
    }
    out=list(likelihood = l.it, theta = thetas)
    return(invisible(out))
}

# same as befor the convergence is faster: in 3 iterations we have our solution
nr(c(1,1))

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.0000000 -48.0273087   1.0908311   0.9426757
## [1]   3.0000000 -48.0271623   1.0971283   0.9375748

## starting from a different point: manda in crash probabilmente per
## il valore della poisson con lambda ancora negativo
nr(c(100,1))

## [1]     1.000 -2514.898   100.000     1.000

## Warning in log(lambda):  Si è prodotto un NaN

## [1]     2.000       NaN -7920.792  2815.468

## Error in while (delta > tol) {:  valore mancante dove è richiesto
TRUE/FALSE
```

6. Derive the Fisher expected Information.

$$I(\theta_1, \theta_2) = -\mathbb{E}\left[H(\theta_1, \theta_2)\right] = \begin{bmatrix} \sum_{i=1}^{n} \frac{x_i^2}{\theta_1 x_i + \theta_2 z_i} & \sum_{i=1}^{n} \frac{x_i z_i}{\theta_1 x_i + \theta_2 z_i} \\ \sum_{i=1}^{n} \frac{x_i z_i}{\theta_1 x_i + \theta_2 z_i} & \sum_{i=1}^{n} \frac{z_i^2}{\theta_1 x_i + \theta_2 z_i} \end{bmatrix}$$

because $\mathbb{E}\left[Y_i\right] = \lambda_i = \theta_1 x_i + \theta_2 z_i$. So

```
I <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    rval = matrix(0,2,2)
    rval[1,1] = sum(x^2/lambda)
    rval[1,2] = rval[2,1] = sum(z*x/lambda)
    rval[2,2] = sum(z^2/lambda)
    rval
}
```

```
## test
I(c(1,1))

##           [,1]     [,2]
## [1,] 17.140991 9.669009
## [2,]  9.669009 8.020991
```

7. Implement Fisher-Scoring algorithm with $\alpha = 0.1$, $\theta = (1, 1)$ and $\theta = (100, 1)$.

```
## Fisher scoring
fs <- function(thetas_0, alpha, tol=10^-5){
    it = 1
    p = length(thetas_0)
    thetas = matrix(thetas_0, nrow = 1, ncol = p)
    l.it = loglik(thetas_0)
    print(c(it, l.it[it], thetas_0))
    delta = 100
    while (delta>tol)  {
        S.it = S(thetas[it,])  # score
        I.it = I(thetas[it,])  # expected info
        theta = c(thetas[it,] + alpha * solve(I.it) %*% S.it) # solve(I.it) is inverse of
        thetas = rbind(thetas, theta)
        l.it = c(l.it, loglik(theta))
        delta=abs((l.it[it+1]-l.it[it])/l.it[it])
        it = it+1
        print(c(it, l.it[it], theta))
    }
    out=list(likelihood=l.it,theta=thetas)
    return(invisible(out))
}

output <- fs(c(1,1), alpha = 0.1)

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.0000000 -48.0593625   1.0117785   0.9906284
## [1]   3.000000 -48.052445   1.022090   0.982632
## [1]   4.0000000 -48.0470535   1.0311241   0.9758091
## [1]   5.0000000 -48.0428410   1.0390436   0.9699881
## [1]   6.0000000 -48.0395421   1.0459904   0.9650233
## [1]   7.0000000 -48.0369530   1.0520873   0.9607901
## [1]   8.0000000 -48.0349170   1.0574411   0.9571825
## [1]   9.0000000 -48.0333131   1.0621448   0.9541095
## [1]  10.0000000 -48.0320473   1.0662791   0.9514938
## [1]  11.0000000 -48.0310468   1.0699148   0.9492689
## [1]  12.0000000 -48.0302547   1.0731132   0.9473781
## [1]  13.0000000 -48.0296268   1.0759282   0.9457727
## [1]  14.0000000 -48.0291284   1.0784068   0.9444112
## [1]  15.0000000 -48.0287323   1.0805900   0.9432578
```

```r
fs(c(1,1), alpha = 0.2) # increasing alpha it is faster
```

```
## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.0000000 -48.0516085   1.0235571   0.9812569
## [1]   3.0000000 -48.0418533   1.0412553   0.9680014
## [1]   4.000000 -48.036068   1.054596   0.958636
## [1]   5.0000000 -48.0326017   1.0646818   0.9520339
## [1]   6.0000000 -48.0305060   1.0723257   0.9473956
## [1]   7.0000000 -48.0292292   1.0781324   0.9441523
## [1]   8.0000000 -48.0284461   1.0825532   0.9418981
## [1]   9.0000000 -48.0279630   1.0859257   0.9403434
## [1]  10.0000000 -48.0276636   1.0885035   0.9392818
```

```r
## starting here it takes longer because very different from the solution
## fs(c(100,1), alpha = 0.1)
```

8. Compute the standard errors of the parameter estimates.

```r
## standard errors
final_theta = output$theta[nrow(output$theta), ]
I(final_theta)
```

```
##           [,1]     [,2]
## [1,] 16.55713 9.455030
## [2,]  9.45503 7.922531
```

```r
diag(solve(I(final_theta)))^0.5
```

```
## [1] 0.4354762 0.6295424
```

9. Solve the same problem with the function `optim` of R.

```r
## To optimize the loglikelihood we use optim givin it the loglikelihood functio

## bfgs is an optimization method
## c(1,1) is the starting parameter values for loglik function
## at the end "control=list(fnscale=-1)" is to say to R to maximize instead of m
optim(c(1,1), fn = loglik, method = "BFGS", control = list(fnscale = -1))
```

```
## $par
## [1] 1.0971528 0.9375544
##
## $value
## [1] -48.02716
##
## $counts
## function gradient
```

```
##       18        5
##
## $convergence
## [1] 0
##
## $message
## NULL

## the maximum log likelihood is -48 and is obtained when the thetas are 1.0972 0.9376
```

# Parte I

# old shit

# Capitolo 3

# Campioni casuali e distribuzioni campionarie

## 3.1 Introduzione all'inferenza

Nella statistica descrittiva si affronta la descrizione di una popolazione nota rispetto a una o più variabili.
Nel calcolo delle probabilità si definisce una prova sulla popolazione di interesse creando una corrispondenza aleatoria tra i valori della variabile ed i risultati della prova; la prova genera eventi con una certa probabilità e cio conferisce un ordinamento oggettivo in termini di plausibilità. Anche in tal caso, grazie alla teoria delle vc e delle loro distribuzioni, si individuano pochi elementi (i parametri) per caratterizzare il fenomeno in esame.

Nell'**inferenza** l'ottica si capovolge perchè si suppone di conoscere solo i risultati della prova e non la popolazione da cui provengono. L'approccio dell'inferenza è quello di utilizzare gli eventi osservati, in un approccio induttivo, per giungere alla conoscenza della popolazione che verosimilmente li ha generati.

Ogni inferenza si basa sulla specificaizone accurata dei seguenti elementi:

1. popolazione di riferimento
2. procedura di raccolta e selezione delle informazioni
3. tecnica inferenziale per giungere dal risultato parziale alla popolazione
4. validità statistica della procedura utilizzata

La **popolazione** è l'insieme delle informaizoni statistiche che esauriscono il problema oggetto dello studio. Nel seguito, popolazione X sarà sinonimo di vc X e la conoscenza della popolazione X coinciderà strettaente con la conoscenza della funzione di ripartizione $F(x, \theta)$ della vc X, con $\theta$ un vettore di $m \geq 1$ parametri che caratterizza la v.c. X all'interno di una prefissata famiglia parametrica.
L'insieme di valori ce i parametri itemize $\theta$ possono assumere sarà definito *spazio parametrico* e indicato con $\Omega(\theta)$.

Dalla popolazione $X$ viene estratto un sottoinsieme di $n$ unità statistiche e la procedura di selezione (assimilabile ad una prova nel calcolo della probabilità, poichè soggetta ad incertezza) genera una $n$-pladi vc $(X_1, X_2, \ldots, X_n)$,

che portano dopo lo svolgimento degli esperimenti ad una n-pla di numeri reali $(x_1, x_2, \ldots, x_n)$ detto *campione osservato*. Ogni numero reale $x_i$ è la relizzazione di una vc $X_i$ detta vc della i-esima estraizone.
È essenziale notare che la vc $X_i$ ha la stessa distribuzione della vc $X$ (della popolazione), ovvero le due distribuzioni sono somiglianti. Le $X_i$ sono in tal senso identicamente distribuite; quando le $X_i$ sono fra loro anche indipendenti, poichè derivano ad esempio da un campionamento bernoulliano (ovvero effettuato con ripetizione) si parla di *campione casuale*.
La procedura di scelta delle unità statistiche che vanno a costituire il campione forma oggetto della *teoria dei campioni*

La parte centrale dell'inferenza è costituita dalle tecniche mediante le quali l'informazione del campione viene rapportata a quella della popolazione; le procedure inferenziali della statistica sono sostanzialmente tre e tra loro interconnesse[1]:

- *stima di un parametro*: nella teoria della stima si cerca di determinare un valore numerico per il vettore di parametri $\theta$ che caratterizza la popolazione $X \sim f(x; \theta$ sulla base delle informazioni campionarie desumibili dal campione osservato $(x_1, x_2, \ldots, x_n)$

- *test di ipotesi*: in questa si controlla quale tra le due affermazioni complementari e in conflitto, dette ipotesi, possa esser ritenuta maggiormente verosimile sulla base delle informazioni campionarie

- costruzione di un *intervallo di confidenza*: qui si cerca di determinare, sulla base dei dati campionari, un intervallo di valori reali (o una regione, nel caso di un vettore di parametri) in cui riporre una prefissata fiducia per il parametro $\theta$

Tutte le procedure inferenziali si articoleranno in due momenti sucessivi:

- stabilire cosa si intende per procedura ottimale

- individuare metodi statistici che producano procedure ottimali

Nelle sezioni seguenti si sviluppa l'inferenza classica (o frequentista/campionaria), maggiormente consolidata e diffusa, la quale si basa sul **principio del campionamento ripetuto**, ovvero: le conclusioni inferenziali basate sull'unico campione osservato devono esser giudicate sulla base della distribuzione di probabilità dei possibili campioni che potevano essere generati.

## 3.2 Campioni casuali e distribuzioni campionarie

Si consideri una vc popolazione $X \sim f(x; \theta)$.

Se si effettuano n prove ripetute di estrazione nelle medesime condizioni (procedendo ad estrazioni bernoulliane, con reinserimento), si realizzerà una v.c. n-pla $\underline{X} = (X_1, X_2, \ldots, X_n)$, in cui le componenti $X_i$, denominate *variabili campionarie*, sono vc indipendenti (dall'estrazione) e identicamente distribuite

[1] Anche se è possibile una loro unificaizone nell'ambito della teoria delle decsioni statistiche

| Statistica | Simbologia | Definizione |
|---|---|---|
| Media campionaria | $\underline{X_n}$ | $\frac{1}{n}\sum_{i=1}^{n} X_i$ |
| Mediana campionaria | | |
| Varianza campionaria | | |
| Varianza campionaria corretta | $S_n^2$ | $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ |
| SQM campionario | | |
| SQM campionario corretto | | |
| Coefficiente di correlazione camp. | | |
| Minimo campionario | | |
| Massimo campionario | | |
| Campio di variazione camp. | | |

Tabella 3.1: Statistiche Campionarie

esattamente come la vc $X \sim f(x;\theta)$. In tal caso si ha un **campione casuale**[2]. Il **campione osservato** $\underline{x} = (x_1, x_2, \ldots, x_n)$ costituisce la n-pla di numeri reali che sono realizzazioni delle corrispondenti vc componenti; $n$ è la *numerosità campionaria*.

Ne deriva che la distribuzione confiunta del campione casuale $\underline{X}$ è

$$f(\underline{x};\theta) = f(x_1, x_2, \ldots, x_n;\theta) = f(x_1;\theta)\cdot f(x_2;\theta)\cdot \ldots \cdot f(x_n;\theta) \tag{3.1}$$

La seconda uguaglianza è resa possibile dall'indipendenza delle vc componenti.

Si è spesso interessati alla distribuzione di funzioni del campione casuale $(X_1, X_2, \ldots, X_n)$ che a loro volta sono vc con una distribuzione di densità/probabilità. Tra tutte queste funzioni sono fondamentali quelle definite *statistiche*, anche dette *stimatori*. Definiamo statistica $T_n = T(X_1, X_2, \ldots, X_n)$ qualunque funzione a volari reali del campione casuale $(X_1, X_2, \ldots, X_n)$ che *non* dipende da quantità incognite. Essendo basata su variabili casuali, anche la $T_n$ sarà una variabile casuale: la distribuzione di probabilità della statistica $T_n$ calcolata sullo spazio di probabilità del campione casuale, si chiama *distribuzione campionaria di* $T_n$
Il valore della statistica/stimatore $T_n$ calcolata sul campione osservato $(x_1, x_2, \ldots, x_n)$ si definisce *statistica calcolata*, o *stima* e si esprime mediante $t_n = T(x_1, x_2, \ldots, x_n)$.

Ogni stimatore è pertanto una sintesi delle vc campionarie. Gli stimatori più comuni sono riportati in tabella 3.1.

Lo studio della distribuzione campionaria di $T_n$ non è sempre agevole per cui va affrontata con approcci differenti a seconda del caso:

- nei casi più semplici, nota la vc $X \sim f(x;\theta)$ si ottiene la distribuzione esatta delle statistice tramite la tecnica delle trasformaizoni di variabili

[2]Come si vedrà in seguito un campione estratto senza ripetizione da una popolazione "finita" di N elementi viene chiamato **campione casuale semplice** ed esso, per un N finito non coincide necessariamente con il campione casuale. Tuttavia se la dimensione campionaria n è molto piccola rispetto alla dimensione finita N della popolazione, si realizza anche per estrazioni senza ripetizione da una popolazione finita, una sorta di quasi-indipendenza. Il che autorizza con qualche approssimazione l'applicazione della teoria del campionamento con ripetizione, altrimenti detto del campionamento da popolazioni infinite, anche a tali situazioni

- in alcuni casi può esser opportuno riferirsi alle funzioni caratteristice o alle funzioni generatrici di momenti di $T_n$ è immediatamente riconoscibile
- si possono calcolare i momenti caratteristici della $T_n$ a partire dalla conoscenza dei momenti di X (in modo esatto o approssimato) verificando poi le proprietà della statistica
- si deriva una distribuzione approssimata di tipo asintotico per la vc $T_n$, sfruttando i teorimi limiti del calcolo delle probabilità
- si deriva per la statistica $T_n$ una distribuzione approssimata di tipo numerico, cioè si calcola la sua funzione di ripartizione per ogni punto di interesse mediante tecniche di simulazione

### 3.2.1 Media e varianza campionaria

Non tutte le sintesi sono di interesse per le decisioni statistiche per cui, nello sviluppo della disciplina, alcune di esse sono divenute piu rilevanti, perchè presenti in numerose applicazioni e problematiche di interesse comune. Tra queste, un ruolo dilevante assumono due statistiche: la media campionaria e la varianza campioniaria

#### 3.2.1.1 Media campionaria

Sia data una vc $X \sim f(x;\theta)$ dalla quale con procedura di selezione casuale estraiamo un campione casuale $\underline{X} = (X_1, X_2, \ldots, X_n)$.
La statistica media campionaria indicata con $\overline{X}_n$ è definita come la somma delle v.c. che definiscono il campione casuale divisa per la numerosità campionaria:

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.2}$$

I valori ce essa può assumere sono determinati attraverso la sintesi appena descritto, dall'universo di tutti i possibiili campioni inclusi nello spazio campionario.

Qualunque sia la distribzuione della vc X se si conoscnono il suo valore medio $E(X) = \theta$ e la sua varianza $Var(X) = \sigma^2 < +\infty$, possiamo dedurre valore medio e varianza della vc $\overline{X}_n$. Infatti le cv componenti del campioni sono indipendenti e somiglianti, per cui grazie alle proprietà del valore medio e della varianza per vc indipendenti si avrà:

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{1}{n} n\theta = \theta \tag{3.3}$$

e

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \tag{3.4}$$

Come si vede a prescindere dalla distribuzione di partenza, la statistica media campionaria ha una distribuzione il cui valore medio $\theta$ coincide con qullo della vc X ed una varianza $\sigma^2/n$, ridotta di un fattore dovuto al campione rispetto alla varianza originaria della distribuzione della popolazione. Quindi in ogni caso

la distribuzione della media campionaria sarà più accentrata attorno al valore medio $\theta$ rispetto a quella della vc X perchè la sua varianza è più piccola e, per $n \to a0$

Se sappiamo che la popolazione è distribuita normalmente $X \sim N(\mu, \sigma^2)$ , allora ciascuna delle variabili componenti lo sarà, ed anche una loro somma/combinazione lneare lo sarà grazie alla proprietà riproduttiva delle v.c. Normali e indipendenti. Quindi la relativa media campionaria sarà:

$$\overline{X_n} \sim N(\mu, \frac{\sigma^2}{n}) \tag{3.5}$$

Si noti che:

- questo è un risultato esatto di estrema importanza per l'inferenza, e vale per qualsiasi dimensione campionaria $n$.
- continua a valere approssimativamente per qualsiasi v.c che abbia varianza finita grazie al Teorema del limite centrale, poichè la media è una trasformazione lineare (somma di vc, divisa per una costante che non cambia la distribuzione). Tale approssimazione migliora se $n \to \infty$

Questo consente di fare affermazioni probabilistiche sulla v.c. media campionaria in modo esatto (se si puo supporre che la popolazione di partenza X sia normale), infatti grazie ad una standardizzazione della vc media campionaria si ha:

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma\sqrt{n}} \sim N(0,1) \tag{3.6}$$

ovvero la media campionaria standardizzata si distribuisce come una $N(0,1)$.
Alternativamente si può sempre effettuare delle affermazioni probabilistiche, se pur in modo approssimato, per qualsiasi campione casuale derivato da vc che abbia varianza finita (come è ragionevole assumere nella gran parte dei casi reali) se vale il TLC.

Infine si può dimostrare che la vc media campionaria ha i seguenti ulteriori momenti caratteristici:

$$Asym(\overline{X_n}) = \frac{Asym(X)}{\sqrt{n}}$$
$$Kurt(\overline{X_n}) = 3(1 - \frac{1}{n}) + \frac{Kurt(X)}{n}$$

Come si vede, per qualsiasi vc la media campionaria è centrata sul valore medio della vc della popolazione. AL crescere di n, lavarianza tende a zero, la simmetria cresce e la forma della distribuzione si avvicina alla normale (asimmetria tende a 0, la curtosi a 3).

#### 3.2.1.2 Varianza campionaria corretta

Sia data una vc $X \sim f(x;\theta)$ dalla quale con procedura di selezione casuale estraiamo un campione casuale $\underline{X} = (X_1, X_2, \ldots, X_n)$.
Definimo la statistica varianza campionaria corretta:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 \tag{3.7}$$

La divisione per (n-1) anzichè per n è opportuna nei piccoli campioni perchè fa si che il valore medio della v.c. $S_n^2$ coincida con la varianza $\sigma^2$. Si può dimostrare infatti che:

$$E(S_n^2) = \sigma^2 \tag{3.8}$$

Se la popolazione X è distribuia normalmente, si può dimostrare che, a meno di una costante, la statistica varianza campionaria corretta è distribuita in modo proporzionale ad una vc Chi-quadrato con $g = n - 1$ gradi di libertà. Più precisamente:

$$S_n^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2 \tag{3.9}$$

Ricordando i momenti caratteristici della Chi quadrato si derivano agevolmente quelli della varianza cmapionaria corretta:

$$E(S_n^2) = \sigma^2 \tag{3.10}$$

$$Var(S_n^2) = 2\frac{\sigma^4}{n-1} \tag{3.11}$$

$$Asym(S_n^2) = \sqrt{\frac{8}{n-1}} \tag{3.12}$$

$$Kurt(S_n^2) = 3 + \frac{12}{n-1} \tag{3.13}$$

Anche questa vc per $n \to 0$ ha una distribuzione che converge alla normale.

Anche se X non è distribuita normalmente, escludendo casi estremi, la varianza campionaria corretta tende ad avere una distribuzione che per $n \to \infty$ può eser approssimata dalla normale, anche se tale convergenza è più lenta. Infatti anche alla varianza campionaria si può applicare il Teorema del Limite Centrale.

#### 3.2.1.3 Indipendenza di media e varianza campionaria corretta

In merito al rapporto tra media campionaria e varianza corretta, va citato il seguente importante risultato riguardante *campioni casuali estratti da vc Normali*: se il campione casuale $\underline{X}$ è generato da una vc normale, allora le statistiche $\overline{X_n}$ e $S_n^2$ *sono vc indipendenti e viceversa* (ovvero se sono indipendenti, allora vuol dire che il campione è stato generato da una popolazione normale).

### 3.2.2 Altre distribuzioni campionarie notevoli

Alcune statistiche derivabili ca campioni causali di vc normali presentano distribuzioni notevoli (nel senso che sono ricondcibili a particolari modelli di vc connesse alla normale.

Se $\underline{X} = (X_1, X_2, \ldots, X_n)$è un campione casuale generato da $X \sim N(\mu, \sigma^2)$, allora la statistica definita da:

$$T_n = \frac{\overline{X_n} - \mu}{S_n\sqrt{n}} \sim T_{n-1} \tag{3.14}$$

si distribuisce come una t di student con $g = n - 1$ gradi di libertà.
L'aspetto notevole di questo risultato, e la sua importanza inferenziale, è che non dipende da parametri incogniti del campione ($\sigma^2$), ad esempio a contrario della

media campionaria standardizzata, per cui è possibile calcolare la probabilità di eventi che riguardano tale statistica senza ipotizzare la conoscenza dei parametri della popolazione.

Il risultato di sopra si può generalizzare alla **differenza delle vc medie campionarie** definite su due campioni casuali tra loro *indipendenti*. Se $\underline{X} = (X_1, X_2, X_n)$ e $\underline{Y} = (Y_1, Y_2, Y_n)$ sono due campioni casuali indipendenti generati da $X \sim N(\mu_x, \sigma^2)$ e $Y \sim N(\mu_y, \sigma^2)$ (ovvero si richiede che le varianze delle due popolazioni coincidano), allora la statistica $T_{n,m}$ definita da

$$T_{n,m} = \frac{\overline{X_n} - \overline{Y_n} - (\mu_x - \mu_y)}{\sqrt{(n-1)S^2_{x,n} + (m-1)S^2_{y,n}}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim T_{n+m-2} \qquad (3.15)$$

si distribuisce come una T con $g = n + m - 2$ gradi di libertà

Infine se $\underline{X} = (X_1, X_2, X_n)$ e $\underline{Y} = (Y_1, Y_2, Y_n)$ sono due campioni casuali indipendenti generati da $X \sim N(\mu_x, \sigma^2_x)$ e $Y \sim N(\mu_y, \sigma^2_y)$ (ovvero ove le due varianze non necessariamente coincidono, allora la statistica

$$T_{n,m} = \frac{S^2_{x,n}}{S^2_{y,n}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n}^2)}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y_n}^2)} \sim \frac{\sigma^2_x}{\sigma^2_y} \cdot F_{(n-1,m-1)} \qquad (3.16)$$

ovvero il rapporto tra le varianze corrette si distribuisce come una v.c. F con gradi di libertà $g_1 = n-1$ e $g_2 = m-1$. L'aspetto notevole di tale risultato è che se le due varianze delle due vc coincidono, allora la distribuzione della statistica non di pende da nessuno dei parametri che caratterizzano le distribuzioni delle popolazioni.

## 3.3 La funzione di verosimiglianza

La funzione di verosimiglianza ha un ruolo centrale per l'inferenza statistica; la sua importanza deriva dal fatto che in un certo modo, essa racchiude tutte le informazioni statisticamente rilevanti riguardanti la popolazione e derivabile dal campione.
Come detto, la distribuzione congiunta del campione casuale $\underline{X}$ è

$$f(\underline{x}; \theta) = f(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta) \qquad (3.17)$$

Ora, se si conosce la vc $X \sim f(x; \theta)$ prima di effettuare l'esperimento campionario, l'espressione $f(\underline{x}; \theta)$ rappresenta la funzione di probabilità/densità che si verifichi il vettore numerico $\underline{x}$. Pertanto una prima possibile utilizzazione della funzione è di natura probabilistica, se ci troviamo in un setting pre estrazione del campione.

Alternativamente se il campione $\underline{x}$ è già stato osservato e se $\theta$è incognito, allora $f(\underline{x}; \theta)$ è funzione solo del parametro $\theta \in \Omega(\theta)$ e viene definita funzione di verosimiglianza, ovvero

$$\mathcal{L}(\theta; \underline{x}) = f(\underline{x}; \theta) \qquad (3.18)$$

essendo in funzione del solo parametro viene anche indicata con $\mathcal{L}(\theta)$.
In questo contesto la funzione di verosimiglianza esprime la probabilità/densità di aver ottenuto quel campione casuale le cui determinazioni numeriche si sono

effettivamente realizzate. Sembra pertanto ragionevole inferire che, se il parametro è sconosciuto, si possa giungere alla sua determinazione massimizzando la funzione di verosimiglianza; la stima di $\theta$ determinata in tal modo verrà indicata con $\hat{\theta}$.
In altre parole se $\mathcal{L}(\theta_1; \underline{x}) > \mathcal{L}(\theta_2; \underline{x})$ sembra ragionevole preferire il valore $\theta = \theta_1$ anzichè $\theta = \theta_2$ perchè a parità di ogni altra circostanza, al primo valore corrisponde una più elevata probabilità di ottenere proprio quel campione.
Si dice che, dato $\underline{x}$, il valore $\theta_1$ è più plausibile del valore $\theta_2$, per cui *la funzione di verosimiglianza è una misura di plausibilità dei differenti valori $\theta$ dello spazio parametrico* $\Omega(\theta)$, una volta osservato $\underline{x}$.
Nell'inferenza ci si pone tipicamente in questo setting per l'utilizzo della verosimiglianza.

Quando $\mathcal{L}(\theta; \underline{x}) > 0$ è convenienente trattare con il suo logaritmo neperiamo, che chiameremo funzione di log-verosimiglianza, la quale è definita da:

$$\begin{aligned} \log \mathcal{L}(\theta; \underline{x}) &= \log(f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta) \\ &= \sum_{i=1}^{n} \log f(x_i; \theta) \end{aligned}$$

La log verosimiglianza è una funzione monotona della verosimiglianza per cui ha gli stessi punti estremanti (massimi e minimi) ma risulta spesso piu agevole da trattare perchè molte distribuzioni di vc sono espresse tramite le funzioni esponenziali o il prodotto di funzioni elementari.

La funzione di verosimiglanza sintetizza le informazioni presenti nel campione con rifereimento ad uno specifico modello probabilistica per la vc $X \sim F(x; \theta)$ e l'inferenza derivata da tale funzione è la conseguenza di principi generali che brevemente riassumiamo:

**Principio debole di verosimiglianza** : con riferimento ad un certo modello $X \sim F(x; \theta)$, se due campioni $\underline{x}, \underline{y}$ sono tali che $\mathcal{L}(\theta; \underline{x}) \propto \mathcal{L}(\theta; \underline{y})$, allora i due campioni debono condurre alle stesse inferenze su $\theta$

**Principio forte di verosimiglianza** se $\underline{x}$ è un risultato campionario relativo alla vc $X \sim F(x; \theta)$ con funzione di verosimiglianza $\mathcal{L}_F(\theta)$, e se $\underline{y}$ è un risultato campionario relativo alla vc $X \sim G(y; \theta)$ con funzione di verosimiglianza $\mathcal{L}_G(\theta)$, allora se: $\mathcal{L}_F(\theta) \propto \mathcal{L}_G(\theta)$ le conclusioni inferenziali riguardanti $\theta$ devono essere le stesse per i due campioni

Mentre il proncipio debole conduce alle stesse inferenze a parità di modello, nel principio forte le conseguenze sono le medesime anche con modelli differenti. La gran parte dell'inferenza statistica aderisce almeno al principio debole della verosimiglianza

La *variabilità della stima derivante dalla funzione di verosimiglianza* dipende dalla curvatura attorno al punto di massimo (in corrispondenza della stima ottenute $\hat{\theta}$): se la funzione non è particolarmente appuntita attorno al suo massimo, allora valori di $\theta$ anche alquanto distanti da $\hat{\theta}$ avranno una verosimiglianza abbastanza simile a quella massima. A contrario la stima di $\hat{\theta}$ sarà più sicura se un movimento da essa comporti una notevole perdita di verosimiglianza.
Ne consegue che la curvatura della funzione di verosimiglianza nell'introno del suo punto di massimo è una misura dell'accurattezza della valutazione numerica del parametro, deditta dallo specifico campione osservato

Infine anticipiamo che la funzione di verosimiglianza interviene in tutte le decisioni inferenziali mediante la seguente logica:

- nella teoria della stima, tra più valori ammissibili per il parametro $\theta$ si cerca quello che sia piùverosimile compatibilmente con il campione osservato
- nel test delle ipotesi, tra due ipotesi statistiche si rifiuta quella meno verosimilmente compatibilmente con il campione osservato
- negli intervalli di confidenza, si può costruire sullo spazio parametrico ammissibile per $\theta$, un intervallo di valori reali tale che i valori di $\theta$ qui contenuti siano tutti adeguatamente verosimili entro una prefissata soglia

# Capitolo 4

# Teoria e metodi di costruzione degli stimatori

Abbiamo detto che inferire significa utilizzare le informazioni campionarie per prendere decisioni riugardandi l'intera popolazione. In qusta sezione ci occuperemo di quella particlare inferenza che concerne la determinazione numerica di un parametro $\theta$ , incognito ma fisso che caratterizza la popolazione $X \sim f(x;\theta)$.

Ricordando quanto già esaminato in precedenza possiamo dire che lo *stimatore* è una statistica (ovvero una sintesi) definita sul campione casuale estratto dalla popolazione mentre la *stima* è il suo corrispondente valore numerico calcolato sul campione osservato.
Essendo lo stimatore del tipo $T_n = T(X_1, X_2, \ldots, X_n)$ esso è una variabile casuale che fornisce stime differenti a seconda del campione estratto; per questo motivo anche esso è una variabile casuale. Come sarà presto evidente occorre pervenire alla consoscenza della distribuzione campionaria dello stimatore $T_n$, sia per valutare la bontà di una particolare procedura di stima ce per confrontare tra loro stimatori alternativi.

In generale per individuare lo stimatore $T_n$ per un parametro $\theta$ bisogna tenere presente due aspetti differenti e sequenziali:

1. bisogna stabilire cosa si intenda per **bontà di uno stimatore**, enucleando tra i molteplici aspetti della distribuzione di $T_n$ quelli ritenuti importanti per un'accurata determinazione numerica di $\theta$. Spesso ciò si sostanzia nel confrontare la vc $T_n$ in rapporto al parametro $\theta$; saremo spesso interessati a discutere della differenza $T_n - \theta$, che è una vc che misura l'errore che inevitabilmente si commette quando al posto di $\theta$ sconosciuto si è costretti ad utilizzare $T_n$ derivante dal campione. Esistendo molti modi per definire la *vicinanza* tra il numero $\theta$ e la vc $T_n$ e quindi esistono molti modi per parlare di bontà, accuratezza, affidabilità di uno stimatore

2. successivamente, occorre individuare **metodi di costruzione degli stimatori** che possiedano alcune o molte delle proprietà indicate come desiderabili per uno stimatore

In generale la bontà di uno stimatore scaturisce dalla validità delle premesse, dalla correttezza delle derivazioni e anche dal rigore con cui viene utilizzato nelle applicaizoni: ciò implica che una superficiale verifica delle ipotesi di base e/o

una raccolta dei dati molto approssimativa posono distruggere l'utilità di uno stimatore, anche se esso è stato ottenuto in modo accurato sul piano formale.

## 4.1 Proprietà degli stimatori

### 4.1.1 Sufficienza di uno stimatore

Sul piano intuitivo, uno stimatore $T_n$ è **sufficiente** se rachiude ed esaurisce tutte le informazioni riguardanti $\theta$ e contenute nel campione casuale

Sia $\underline{X} = (X_1, X_2, \ldots, X_N)$ un campione casuale generato dalla vc $X \sim f(x : \theta)$ dove $\theta \in \Omega(\theta)$ è il parametro oggetto di stima. Diremo che $T_n$ è uno **stimatore sufficiente** per $\theta$ se la distribuzione condizionata del campione osservato $(X_1, X_2, \ldots, X_n)$ dato che $T_n$ abbia assunto un qualsiasi valore $t_0$ :

$$\varphi_X(x_1, x_2, \ldots, x_n | T_n = t_o) = \frac{h(x_1, x_2, \ldots, x_n, T_n = t_0; \theta}{g(T_n = t_0; \theta)} \tag{4.1}$$

non dipende dal parametro $\theta$. O in altre parole, una volta conosciuto il valore dello stimatore non vi è informazione aggiuntiva necessaria per conoscere la probabilità di verificarsi di quello specifico campione.
In tal modo, tutte le informazioni riguardanti $\theta$ vengono integralmente trasferite nello stimatore; infatti una volta osservato uno specifico valore $t_0$, non vi è più alcuna in formazione riguardante il parametro $\theta$ nella distribuzione condizionata del campione casuale

Nel caso multivariato, se $\theta$ è un vettore composto da m>1 parametri della vc $X \sim f(x; \theta)$, dalla quale è generato un campione casuale $(X_1, X_2, \ldots, X_n)$, allora lo *stimatore vettore* $\mathbf{T_n}$ è *congiuntamente sufficiente* (joint sufficient) per il vettore di parametri $\theta$ se la distribuzione $\varphi(X_1, X_2, \ldots, X_n)$ non dipende da $\theta$.

Si osservi che la sufficienza si definisce in rapporto ad un parametro $\theta$ all'interno di una bene specificata vc $X \sim f(x; \theta)$; in altri termini si può parlare di sufficienza solo dopo aver definito qual'è la famiglia di vc che ha generato il campione casuale. Quindi uno stimatore $T_n$ può esser sufficiente per il parametro $\theta$ di una vc ma non per quello di un'altra vc appartenente ad una famiglia diversa

Operativamente la definizione di sufficienza è complessa per esser utilizzata quando si tratta di stabilire se uno stimatore è sufficiente (bisogna derivare tutte le funzioni di densità del campione per ogni valore $T_n = t_0$); a tal fine si usa il teorema di fattorizzazione che esplicita le condizioni necessarie e sufficienti affinché un determinato stimatore sia sufficiente per $\theta$. TODO Piccolo 2000 pag 541).

In generale:

- se $T_n$ è uno stimatore sufficiente per $\theta$ lo sarà anche qualsiasi funzione $g(T_n)$ che non contenga valori incogniti; infatti tutte le informazioni riguardanti $\theta$ e contenute in $T_n$ saranno ancora presenti in $g(T_n)$

- da questo deriva che la sufficienza individua una classe molto ampia di funzioni delle vc campionarie, ma occorre qualche criterio aggiuntivo per individuare fra gli infiniti stimatori sufficienti quello migliore

### 4.1.2 Proprietà finite di uno stimatore

È necessario introdurre altre proprietà di uno stimatore, al fine di poter effettuare delle scelte, e tra queste qui discutiamo quelle che sono definite per una dimensione campionaria $n$ prefissata. Nell'ordine parliamo di *non distorsione*, *efficienza* ed *efficienza relativa* all'interno di una particolare classe di stimatori

#### 4.1.2.1 Non distorsione

Uno stimatore $T_n$si dice non distorto (*unbiased*), o corretto, per $\theta$ se:

$$E(T_n) = \theta \tag{4.2}$$

La distorsione (*bias*) di uno stimatore $T_n$ è definita da

$$b(T_n) = E(T_n) - \theta \tag{4.3}$$

e sarà nulla nel caso di uno stimatore corretto.
La non distorsione è una proprietà desiderabile di uno stimatore, perchè pur non asserendo alcunche sulla singola stima, richiede che la procedura inferenziale prescelta per la stima *non produca deviazioni sistematiche* rispetto al parametro $\theta$. Ovvero la distribuzione campionaria dello stimatore è centrata proprio sul parametro $\theta$.

#### 4.1.2.2 Efficienza

D'altra parte il valore atteso di una vc è tanto piu rappresentativo quanto più la varianza è piccola. La varianza dello stimatore misura la dispersione delle stime fornite attorno al valore medio; se lo stimatore è distorto e il valore medio non coincide con il parametro, la varianza dello stimatore non può essere un indicatore di bontà dello stimatore (si è piu o meno precisi ma su un valore sbagliato).
Per pervenire ad un criterio utile per valutare sia stimatori non distorti che distorti, in relazione alla vicinanza rispetto alla quale forniscono stime del parametro vero $\theta$, si introduce l'**errore quadratico medio** come:

$$MSE(T_n) = E(T_n - \theta)^2 \tag{4.4}$$

Esso tiene conto sia della varianza che della distorsione di uno stimatore; dopo alcuni passaggi algebrici, si può dimostrare che:

$$MSE(T_n) = Var(T_n) + [b(T_n)]^2 \tag{4.5}$$

ovvero è uguale alla varianza dello stimatore più la distorsione al quadrato; il MSE di uno stmatore non distorto coincide con la varianza dello stimatore. Si noti ce MSE è funzione di $\theta$, anche se non l'abbiamo esplicitato per semplicità di notazione.

L'importanza di confrontare il MSE di due o più stimatori come criterio di vicinanza reltiva rispetto al parametro $\theta$ preferendo quello che, al variare di $\theta$, possiede MSE inferiore può esser formalizzato introducendo **l'efficienza relativa** di uno stimatore: uno stimatore $T_{1n}$ si dice più efficiente di uno stimatore

$T_{2,n}$ per lo stesso parametro $\theta$ se $MSE(T_{1n}) < MSE(T_{2n})$, o equivalentemente se

$$eff(T_{1n}|T_{2n}) = \frac{MSE(T_{2n})}{MSE(T_{1n})} > 1 \tag{4.6}$$

Evidentemente, se entrambi gli stimatori sono non distorti per $\theta$, allora l'efficienza relativa $eff(T_{1n}|T_{2n})$ sarà il rapporto tra le varianze.
Vanno preferiti stimatori più efficienti.

L'efficienza relativa si può concretamente interpretare in termini di costi di campionamento necessari per assumere informazioni su un certo parametro, a parità di precisione (misurata dall'inverso dell'MSE).
Come sarà formalizzato in seguito la gran parte degli stipatori possiede MSE che si riducono al crescere di n (per la riduzione della varianza dello stesso), per cui l'unico modo per aumentare l'efficienza di uno stimatore (e quindi diminuire il suo MSE) è quello di aumentare la sua dimensione campionaria. Ora nel confronto tra due stimatori risulterà preferibile quello che a parità di dimensione campionaria possiede un MSE più basso.
Cfr piccolo 2000 pag 556-557.

Il concetto di efficienza relativa risolve il problema del confronto tra due stimatori, ma non esclude che possano esistere altri stimatori la cui variabilità sia minore di quella dei due considerati. Occorre allora chiedersi se esiste un *limite inferiore per la variabilità di uno stimatore di un certo parametro.*
Se eiste uno stimatore $T_n$ non distorto per $\theta$ che, fra tutti gli stimatori non distorti, possiede la varianza più piccola - cioè è il piu efficiente - allora $T_n$ è chiamato lo **stimatore non distorto con varianza uniformemente minima**.
Per risolvere il problema occorre rifarsi alla disuguaglianza di Cramer e Rao che, per una prefissata famiglia di vc individua, sotto condizione di regolarità, il limite inferiore della varianza di uno stimatore di un parametro. Allora se uno stimatore non distorto ha varianza pari a quella del limite fissato dalla disuguaglianza di Cramer e Rao, esso è uno stimatore **efficiente in assoluto**.
Nella terminologia statistica, la dizione "efficiente" senza alcuna specificazione ulteriore significa "efficiente in assoluto".

Per la famiglia delle vc esponenziale si dimostra che ci sia uno stimatore efficiente.

Infine, poichè una stima che possa variare molto da campione a campione segnala imprecisione e inaffidabilità, è sempre auspicabile accompagnare una stima con una misura della sua variabilità.
L'efficienza dello stimatore costituisce un punto di partenza, ma desiderando esprimere la variabilità dello steso nella sua unità di misura è preferibile al radice quadrata $\sqrt{MSE(T_n)}$ o, se lo stimatore è non distorto, la radice quadrata di $\sqrt{Var(T_n)}$.
Purtroppo pero tali quantità quasi sempre dipendono dagli stessi parametri incogniti oggetto di stima e la soluzione più ovvia è quella di sostituire nelle formule dalla varianza degli stimatori, al posto dei parametri incogniti le corrispondenti stime, giungendo ad ottenere quantità universalmente note come **errore standard della stima**

### 4.1.3 Proprietà asintotiche di uno stimatore

È ragionevole richiedere ce le proprietà statistiche di uno stimatore migliorino al crescere della numerosità campionaria e, in questa parte tale aspetto verrà formalizzato mediante l'introduzione di alcune proprietà, definite *asintotiche*, perchè valide quando $n \to \infty$.

Uno stimatore $T_n$ si dice **asintoticamente non distorto** per $\theta$ se:

$$\lim_{n\to\infty} E(T_n) = \theta \iff lim_{n\to\infty} b(T_n) = 0 \tag{4.7}$$

Quindi uno stimatore asintoticamente non distorto è uno stimatore eventualmente distorto per $n$ finito, ma la cui distorsione tende a zero al crescere della numerosità campionaria.
Di maggior rilevanza sono le proprietà asintotiche collegate alla **consistenza**[1] (in *media quadratica*, in *probabilità*, *quasi certa*).

Uno stimatore $T_n$ si dice **consistente in media quadratica** per $\theta$ se:

$$\lim_{n\to\infty} MSE(T_n) = \lim_{n\to\infty} E(T_n - \theta)^2 = 0 \tag{4.8}$$

ovvero se il suo MSE tende a 0 all'aumentare della dimensione campionaria.
Poichè il MSE di uno stimatore è la somma di due quantità non negative, $Var(T_n)$ e $[b(T_n)]^2$ la definizione precedente equivale al contemporaneo verificarsi delle due seguenti condizioni:

$$\lim_{n\to\infty} Var(T_n) = 0 \tag{4.9}$$

$$\lim_{n\to\infty} [b(T_n)]^2 = 0 \tag{4.10}$$

In particolare, se uno stimatore è non distorto (o asintoticamente non distorto), allora è consistente in media quadratica se la varianza dello stimatore tende a 0 al crescere della numerosità campionaria[2].
Essendo l'MSE una misura della variabilità media dello stimatore $T_n$ attorno al parametro $\theta$, la consistenza in media quadratica garantisce che al crescere di n, la distribuzione dello stimatore $T_n$ tende a concentrare tutta la massa di probabilità in un intervallo infinitesimo attorno al parametro $\theta$, fornendo via via stime più precise, e per $n \to \infty$ assumerà il solo valore $T_n = \theta$ con probabilità 1.

Uno stimatore si dice **consistente in probabilità** per $\theta$ se per ogni $\epsilon > 0$ fissato piccolo a piacere,

$$\lim_{n\to\infty} P\left(|T_n - \theta| < \epsilon\right) = 1 \tag{4.11}$$

In generale la consistenza in media quadratica di uno stimatore $T_n$ per il parametro $\theta$ implica la consistenza in probabilità.

[1] Il termine consistenza deriva dall'errata traduzione dall'inglese dei vocaboli consistency/consistent che in realtà significherebbe coerenza/coerente. La traduzione erronea è però la piu impiegata in Italia.

[2] Vale ance viceversa, cioè se uno stimatore è consistente in media quadratica, esso è asintoticamente non distorto poiche la somma di due quantità non negative può tendere a zero se e solo se entrambi gli addenti tendono a zero.
Quindi la consistenza in media quadratica implica la non distorsione asintotica.

Uno stimatore $T_n$ per il parametro $\theta$ si dice **asintoticamente Normale** se:

$$\frac{T_n - E(T_n)}{\sqrt{Var(T_n)}} \xrightarrow{d} Z \sim N(0,1) \tag{4.12}$$

cioè se al crescere della numerosità campionaria, la distribuzione della *vc stimatore standardizzato* tende alla distribuzione della vc Normale standardizzata.
Molti stimatori di uso comune in statistica sono asintoticamente normale perchè sono esprimibili come successioni di funzioni campionarie per le quali è possibile applicare il teorema del limite centrale

## 4.2 Metodi di costruzione degli stimatori

In precedenza presupponevamo che si conoscesse già la formula di $T_n$; a questo punto bisogna introdurre i metodi per la costruzione di uno stimatore.

I metodi di costruzione principali sono:

- metodo dei momenti
- metodo dei minimi quadrati
- metodo del minimo Chi-quadrato
- metodo della massima verosimiglianza

### 4.2.1 Metodo della massima verosimiglianza

Esso deriva da un principio elementare: tra i possibili valori del parametro $\theta$ si preferisce quello che corrisponde alla massima probabilità di generare i dati osservati.

Sia $(X_1, X_2, X_n)$ un campione casuale generato da $X \sim f(x;\theta)$ ove $\theta \in \Omega(\theta)$. Il metodo della massima verosimiglianza propone come stima per $\theta$ il valore $t = T(x_1, x_2, \ldots, x_n)$ per il quale la funzione di verosimiglianza $\mathcal{L}(\theta; \underline{x})$ è massima.
Poichè la funzione logaritmica è una funzione monotona, è spesso conveniente utilizzare il logaritmo della funzione di verosimiglianza, cioè la log-verosimiglianza.

Pertanto nei casi regolari si può definire la stima ML come quel valore di $\theta$ per cui si verificano:

$$V_n^{'}(\theta) = \frac{\delta \log \mathcal{L}(\theta; \underline{x})}{\delta\theta} = 0 \tag{4.13}$$

$$V_n^{''}(\theta) = \frac{\delta^2 \log \mathcal{L}(\theta; \underline{x})}{\delta\theta^2} < 0 \tag{4.14}$$

$$\tag{4.15}$$

In qualche caso la soluzione è esplicita e si riesce, mediante la soluzione delle equazioni a giungere ad una formula; spesso invece la soluzione non è esplicitabile ed è necessario ricorrere a metodi numerici (metodo della sostituzione, metodo di Newton e Raphson, metodo dello scoring ecc).
Quando non è possibile effettuare le derivate rispetto al parametro (perchè non sussistono le condizioni di regolarità) occorre esaminare attentamente la funzione di verosimiglianza verificando per quali valori di theta essa possa raggiungere il suo massimo.

#### 4.2.1.1 Proprietà

Il metodo ML trova giustificazioni logiche nel principio di verosimiglianza e costituisce un riferimento obbligato per le analisi inferenziali grazie ad una serie di risultati teorici, che possono esser così riassunti: *sotto condizioni di regolarità, soddisfatte nelle situazioni più comuni, gli stimatori ML possiedono tutte le proprietà considerate ottimali se non al finito, almeno asintoticamente* . Vedi per approfondimenti.

# Capitolo 5

# Test d'ipotesi

Nel test delle ipotesi statistiche si evidenzia con estrema chiarezza il ruolo dells statistica come scienza delle decisioni in condizioni di incertezza.
Nel seguito dopo una discussione sulla logica e sulle caratteristiche di un test, elencheremo alcuni test per verificare specifiche ipotesi statistiche.

Analogamente al problema di stima di un parametro esistono due questioni da affrontare:

- quando un test è ottimale rispetto ad un obiettivo? La risposta riguarda le proprietà di un test delle ipotesi
- come si ottengono test ottimali per una data ipotesi statistica? La risposta concerne i metodi di costruzione di un test delle ipotesi

Fondamentale è qui il fatto che la natura delle conoscenze presupporte determina le metodologie statistiche che si adotteranno per le decisioni; in particolare si può

- derivare un test supponendo nota la distribuzione di una vc X della popolazione; l'inferenza in questo caso riguarderà i parametri che specificano la distribuzione (*test parametrici*)
- derivare un test senza fare assunzioni stringenti sulla forma della X, per cui l'inferenza riguarda sia la forma della distribuzione che i suoi parametri (*test non parametrici*)

## 5.1 Teoria dei test

### 5.1.1 Logica e caratteristiche fondamentali

Il test delle ipotesi statistiche può esser deifnito come una regola istituita sullo spazio campionario mediante la quale, in funzione del campione osservato, si decide se rifiutare o meno una ipotesi statistica $H_0$ riferita alla popolazione, detta *ipotesi nulla*.
Gli elementi che caratterizzanoun test sono:

- l'ipotesi statistica

- il campione casuale
- la regola di decisione

#### 5.1.1.1 L'ipotesi statistica

Per giungere ad un test delle ipotesi bisogna tradurre un'affermazione del mondo reale in un'affermazione riguardante la distribuzione di probabilità di una vc, la cui veridicità si controlla sulla base delle eidenze campionarie.
Indichiamo con $\Omega(\theta)$ lo spazio parametrico della vc $X \sim f(x;\theta)$ (non abbiamo utilizzato il grassetto per $\theta$, ma comunque può essere ad $m \geq 1$ dimensioni, date da m parametri per avere una variabile ben definita), e con $\omega_0 \subset \Omega(\theta)$ l'insieme dei valori specificati dall'ipotesi nulla $H_0$. Se $H_0$ non è vera, ed è quindi vera l'ipotesi alternativa $H_1$ allora $\theta \notin \omega_0$.
Il test delle ipotesi cerca di verificare quale delle due seguenti affermazioni

$$H_0 : \theta \in \omega_0 \tag{5.1}$$
$$H_1 : \theta \notin \omega_0 \tag{5.2}$$

non sia contraddetta dai risultati campionari $(x_1, x_2, \ldots, x_n)$.
Le ipotesi $H_0$ e $H_1$ sono esaustive e disgiunte: o vale l'una o vale l'altra
Ora si può distinguere tra ipotesi semplici e composite:

- se $H_0$è una ipotesi semplice allora $\omega_0$ consiste di un suolo punto $\theta_0$ (consistente in un numero se m=1, una coppia di numeri se m=2 ecc). In tali casi l'ipotesi nulla è anche indicata con $H_0 : \theta = \theta_0$. Si dice ipotesi semplice poichè vi è una specificazione completa dei parametri di riferimento, che così facendo individuano solo una vc. X generatrice possibile.

- negli altri casi (quando $\omega_0$ si compone di più valori possibili) si ha una $H_0$ come **ipotesi composita**. Se l'ipotesi statistica composita riguardante il parametro $\theta$ include valori reali in una sola direzione (es maggiori e uguali ad un certo $\theta_0$) parleremo di ipotesi statistica unidirezionale, altrimenti di ipotesi statistica bidirezionale.
  La dizione ipotesi "composita" deriva dal fatto che non individuando una vc completamente, bensì individuando una famiglia di vc, è come se fossero una composizione di ipotesi semplici, ottenute individuando un particolare elemento della famiglia

Per ipotesi nulla $H_0$ si intende l'ipotesi che sussiste fino a prova contraria, mentre l'ipotesi alternativa è la situazione complementare rispetto ad $H_0$: le ipotesi non sono mai equivalenti ai fini della decisione nel senso che, come si vedrà, il test non è mai conclusivo circa $H_1$ (se mai lo può essere per $H_0$). Infine si anticipa che la decisione di rifiutare $H_0$ ha una conseguenza univoca, poichè attesta la non validità di una affermazione precisa e ben specificata. Al contrario la decisione di non rifiutare $H_0$ non implica un sostegno a favore di $H_0$, ma solo l'impossibilità di evindenziare dal campione osservato elementi tali da indurci a rifiutarla.

#### 5.1.1.2 Il campione casuale

Il test è una regola basata sullo spazio campionario che si concretizza in una funzione definita sull'insieme di tutti i possibili campioni generati da X, cioè su

un sottospazio di $\mathbb{R}^n$.
Ovvero esisteranno valori $(x_1, x_2, \ldots, x_n) \in R_0 \subset \mathbb{R}^n$ per i quali la regola impone di rifiutare $H_0$ (la regione $R_0$ è detta *regione di rifiuto*, o *regione critica* per $H_0$, abbreviata ad RC) e altri valori $(x_1, x_2, \ldots, x_n) \notin R_0$ per i quali la regola dettta impone di non rifiutare $H_0$.
Per poter applicare un test è necessario e sufficiente conoscere la RC.

#### 5.1.1.3 La regola di decisione

La regola di decisione che individua una RC crea una bipartizione dello spazio campionario; essa non è agevole da applicare perchè richiederebbe di elencare (o almeno individuare) tutte le n-ple realizzazioni campionarie generate dal campione casale che appartengono ad $R_0$.
Si pone pertanto il problema di ridurre lo spazio delle informazioni campionarie a quello effettivamente necessario ed efficace per decidere sulla popolazione. Tale sintesi, se ci poniamo nell'obiettivo di testare ipotesi su un dato parametro, è la statistica sufficiente per il parametro $\theta$, la quale nell'ambito della teoria del test prende il nome di *statistica-test sufficiente.*

In generale per ipotesi concernenti un vettore di parametri $\theta$ di dimensione $m$, la riduzione di $(X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ avviene mediante la statistica $T_n = T(X_1, X_2, \ldots, X_n) \in \mathbb{R}^m$ ove $m < n$. In altre parole la decisione effettuata in base al campione (in uno spazio ad $n$ dimensioni) potrà esser resa equivalente ad una decisione effettuata sulla base di una statistica di dimensione inferiore.
La regola decisionale sarà tradotta in una nuova regola, fondata ora su $T_n$:

$$T_n \in C_0 \quad \longrightarrow \quad \text{si rifiuta } H_0 \tag{5.3}$$
$$T_n \notin C_0 \quad \longrightarrow \quad \text{non si rifiuta } H_0 \tag{5.4}$$

ove abbiamo indicato con $C_0$ la RC, inclusa nello spazio $\mathbb{R}^m$ ottenuta dalla trasformazione di quella originaria, ovvero

$$\{(X_1, X_2, \ldots, X_n) \in R_0\} \iff \{T_n \in C_0\} \Rightarrow \text{ si rifiuta } H_0 \tag{5.5}$$

Quando $m = 1$ le RC individuano si semplicemente tramite intervalli: in particolare si definiscono regioni critiche unidirezionali se sono del tipo maggiore/minore di ($T_n \geq c$, $T_n \leq c$, con $c$ da specificare), o regioni critiche bidirezionali ($c_1 \leq T_n \leq c_2$, con $c_1, c_2$ da specificare), se sono del tipo diverso da o compreso fra.

Nel seguito di questa trattazione, utilizzeremo una notazione semplificata del tipo $\{\underline{X} \in C_0\}$, ovverosia non distinguendo più tra le regioni $R_0$ e $C_0$

Per ricordare:

- le ipotesi statistiche di un test creano una bipartizione dello spazio parametrico mediante la distinzione tra ipotesi nula ed ipotesi alternativa
- la regola di decisione crea una bipartizione dello spazio campionario, distinguendo tra *regione critica* $C_0$ e regione di *accettazione*, definita come complementare alla precedente

Si può dunque vedere il test delle ipotesi come una corrispondenza tra lo spazio campionario $\mathbb{R}^n$ e lo spazio parametrico $\Omega(\theta)$, a cui si attribuisce una valutazione

probabilistica.
Se il campione casuale appartiene alla RC $C_0$ si decide per il rifiuto di $H_0$ , altrimenti si decide di non rifiutare $H_0$. Tale corrispondenza non è certa e ad essa si può attribuire solo un valore probabilistico perchè non è possibile affermare, in generale, che se il campione determina un test che appartiene a $C_0$ non sia comunque possibile che sia vera $H_0$, oppure se $\underline{X} \notin C_0$ non sia comunque possibile che sia falsa $H_0$. Solo una corrispondenza deterministica (matematica) tra lo spazio campionario e quello parametrico in cui alcuni campioni potessero esser generati solo da $H_0$ ed altri solo da $H_1$, allora il problema del test delle ipotesi si risolverebbe in un problema di logica per cui "dal risultato $\underline{X}$ segue necessariamente l'affermazione $H_0$". Nei problemi reali questo accade molto raramente.

In generale si può affermare che **due test sono differenti se**, a *parità di ipotesi* da confrontare, *sono differenti le rispettive RC*: un **buon test** è quello la cui RC presenta proprietà ottimali nel senso che ora dovremo definire

### 5.1.2 Struttura probabilistica del test

Nel campo del test di ipotesi, e in ragione della veridicità di $H_0$ o $H_1$ ci si può trovare in una delle seguenti situazioni ipotetiche:

- $G_1$: si decide di non rifiutare $H_0$ sulla base del campione osservato ed effettivamente $H_0$ è vera. La decisione presa è corretta.
- $G_2$: si decide di rifiutare $H_0$ sulla base del campione osservato ed effettivamente $H_0$ è falsa. La decisione presa è corretta.
- $E_1$: si decide di rifiutare $H_0$ sulla base del campione, ma in realtà $H_0$ è vera. Si prende quindi una decisione errata, ovvero si commette un errore
- $E_2$: si decide di non rifiutare $H_0$ sulla base del campione osservato ma in realtà $H_0$ è falsa. Si prende quindi una decisione errata, ovvero si commette un errore

Ovviamente il ricercatore non sa se ha preso una decisione giusta o errata poichè non sa se $H_0$ è effettivamente vera o falsa: dopo aver deciso di rifiutare o meno può solo accadere che si sia presa la decisione giusta o sbagliata.

Gli eventi $E_1$ ed $E_2$ si definiscono rispettivamente **errore del 1° e del 2° tipo**; le probabilità di tali eventi si indicano nel seguente modo:

- $\alpha = P(\text{Rifiuto } H_0 | H_0 \text{ è vera}) = P(\underline{X} \in C_0 | H_0 : \theta \in \omega_0)$
- $\beta = P(\text{Non rifiuto } H_0 | H_0 \text{ è falsa}) = P(\underline{X} \notin C_0 | H_1 : \theta \notin \omega_0)$
- $1 - \alpha = P(\text{Non rifiuto } H_0 | H_0 \text{ è vera}) = P(\underline{X} \notin C_0 | H_0 : \theta \in \omega_0)$
- $\gamma = 1 - \beta = P(\text{Rifiuto } H_0 | H_0 \text{ è falsa}) = P(\underline{X} \in C_0 | H_1 : \theta \notin \omega_0)$

Tutte queste probabilità sono funzioni della RC $C_0$ (quindi si potrebbe scrivere $\alpha(C_0), \beta(C_0), \gamma(C_0)$ e per motivi che sarano chiari in seguito prendono il nome di:

- $\alpha$ probabilità dell'errore di primo tipo, detto anche livello di significatività del test

- $\beta$ probabilità dell'errore di 2° tipo
- $\gamma = 1 - \beta$ potenza del test, ovvero probabilità di rifiutare correttamente $H_0$

E ragionevole richiedere che la RC sia tale che le due probabilità $\alpha, \beta$ siano entrambe sufficientemente piccole ma, per una dimensione campionaria prefissata, non è possibile farle tendere a zero contemporaneamente.
La relazione che sussiste tra $\beta$ e $\alpha$ è generalmente esprimibile mediante una funzione decrescente, quindi se si vuole far diminuire una probabilità di errore, a parità di campione bisogna accettare un aumento dell'altra probabilità di errore.
E' necessario trovare una RC che produca un compromesso ragionevole tra $\alpha$ e $\beta$, *tenendo conto della natura differente dei due possibili errori che si possono commettere*. Nella tradizione classica del test delle ipotesi si ritiene più grave commettere l'errore del 1° tipo rispetto a quello del 2° tipo.

Ovvero spesso si desidera privilegiare nella costruzione della RC, quelle che pongono in maggior rilievo la proabilità di errore del primo tipo, senza trascurare la probabilità di errore del secondo.
Tale esigenza conduce al concetto di *regione critica ottimale di ampiezza prefissata*, nella quale in primo luogo è stabilita la probabilità di errore $\alpha$.
Definiamo **Regione Critica Ottimale di ampiezza** $\alpha$, indicata con $RCO(\alpha)$ una RC $C_0$ per $H_0$ tale che la probabilità dell'errore del primo tipo sia $\alpha$ ($P(\underline{X} \in C_0|H_0) = \alpha$) e che, per qualsiasi altra RC $C_0^{'}$ di uguale ampiezza $\alpha$, RCO possieda la più piccola probabilità dell'errore del secondo tipo $\beta$, ovvero

$$\beta(C_0) = P(\underline{X} \notin C_0|H_1) < P(\underline{X} \notin C_0^{'}|H_1) = \beta(C_0^{'}) \tag{5.6}$$

o equivalentemente

$$\gamma(C_0) = P(\underline{X} \in C_0|H_1) > P(\underline{X} \in C_0^{'}|H_1) = \gamma(C_0^{'}) \tag{5.7}$$

In altre parole $RCO(\alpha)$ è la regione critica più potente tra quelle di errore $\alpha$ prefissato.

La definizione della $RCO(\alpha)$ qualifica il significato di un buon test delle ipotesi; infatti il test che possiede tale RC è quello che , ad u un livello prefissato di probabilità di errore del primo tipo, consente la minima probabilità di errore del secondo tipo e di conseguenza, la più elevata potenza.
Nel seguito si introduce il metodo ce, sotto determinate condzioni, perviene alla costruzione di $RCO(\alpha)$ per ogni prefissato test delle ipotesi.

### 5.1.3 Lemma di Neyman e Pearson

Il risultato teorico che illustriamo costituisce l'elemento essenziale per costruire RCO nei casi di ipotesi semplici e, grazie ad alcune generalizzazioni, in alcuni casi più generali.

Sia $\underline{X} = (X_1, X_2, \ldots, X_n)$ un campione casuale generato da $X \sim f(x;\theta)$ e si voglia verificare $H_0 : \theta = \theta_0$ contro $H_1 : \theta = \theta_1$. Se $\mathcal{L}(\theta; \underline{X})$ è la funzione di verosimiglianza di $\underline{X}$, allora la $RCO(\alpha)$ per $H_0$ contro $H_1$ è quella regione $C_0$ dello spazio campionario che soddisfa:

$$\frac{\mathcal{L}(\theta_1; \underline{X})}{\mathcal{L}(\theta_0; \underline{X})} \geq c \tag{5.8}$$

$$P(\underline{X} \in C_0|H_0) = \alpha \tag{5.9}$$

Rispettivamente considerabili come primo e secondo vincolo del lemma.

Alcuni commenti:

- Il lemma individua una particlare $RCO(\alpha)$ quando si specifica la costante $c$, la quale può esser univocamente determinata perchè e prefissata l'ampiezza $\alpha$ della RC.
- Essendo un risultato teorico esso esprime la RCO tramite il campione casuale $\underline{X}$
- Operativamente la prima disequazione vincolo di cui sopra determina la forma della RC, la seconda specifica la sua ampiezza
- una conseguenza di grande rilievo del lemma è che la $RCO(\alpha)$ è una funzione statistica sufficiente $T_n$ per $\theta$ (quando essa esiste). Infatti se esiste per $\theta$ uno stimatore sufficiente $T_n$ allora si dimostra che il rapporto tra le funzioni di verosimiglianza (presenti nel primo vincolo del lemma) include l'informazione esplicitata dal campione casuale $(X_1, X_2, X_n)$ solo attraverso la statistica sufficiente. Quindi la RCO individuata dal Lemma di Neyman e Pearson è funzione del campione casuale solo attraverso la statistica sufficiente $T_n$ per $\theta$
- Il lemma individua la RCO per una ipotesi nulla semplice ed una alternativa anch'essa semplice; tuttavia la natura e la specificazione della RCO non cambiano se si incrementa/decrementa nella stessa direzione il valore numerico di $\theta_1$ specificato in $H_1$.