LESSONS 3 & 4

Introduction to the Theory of Estimation and Statistical inference.

Lessons 3 & 4–Plan

1. Probability and Inference
2. As a summary...
3. Introduction to Sampling Theory
4. Theory of Estimation–Introduction
5. Point estimation: estimator and estimate
6. Properties of point estimators
7. Methods of Estimation
    - The method of Moments
    - The Maximum Likelihood Method
    - Properties of Maximum Likelihood Estimators
    - The maximum likelihood in R
8. Most Common Estimators and Sampling Distributions –Summary

Probability and Inference

It can be said that Probability and Inference have different objectives: while in Probability one starts from a given scheme or model to calculate the probability that certain results or events can be observed; Inference is based on data or observations and one seeks to know or to infer something about the model.

Inference is the “way of going from particular to general.”
Statistical Inference aims at defining procedures which, applied to a sample drawn from the population, allow us to estimate unknown parameters of the population or to estimate something about the population model.

Indeed a particular sample is just one of many samples (in an infinity number if the population is infinite) that can be obtained for one sampling process.

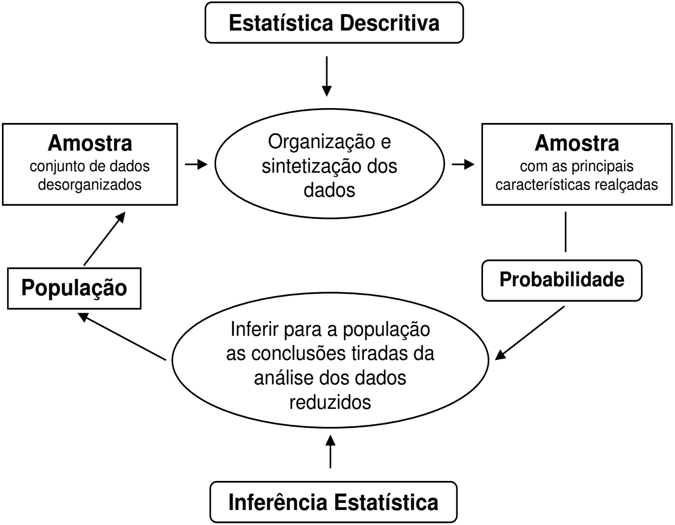
Probability and Inference

The extension of the particular to the general is called inductive inference. It is the way to acquire new knowledge.
The degree of uncertainty that is linked to the inductive inferences can be rigorously measured in terms of probability, if the experiment was conducted according to certain principles (probabilistic or random).

The procedures that allow to extract samples from the population are from the domain of Sampling Theory.

While Sampling refers, as a general rule, to procedures of data collection representative of finite populations, Experimental Design refers to the “production” of data, when controlling the variation of some variables, leaving one or more free

As a summary...



Estatística Descritiva

Amostra
conjunto de dados desorganizados

Organização e sintetização dos dados

Amostra
com as principais características realçadas

População

Probabilidade

Inferir para a população as conclusões tiradas da análise dos dados reduzidos

Inferência Estatística

Introduction to Sampling Theory

Let X be a variable of interest in a population under study.

In the sampling theory procedures are used to extract a collection of population elements to obtain a sample.
We use that sample to infer properties of the variable being studied.
So it is concern of the sampling theory:

- to collect representative samples of the population;
- to determinate the sample size (for which it is crucial to estimate the characteristic variability) to obtain estimates of parameters of interest with a given precision

Before we mention some preliminary concepts in sampling, let's have look at an exercise:

Exercise 1

Let’s suppose that we have a rectangular field of 5 km × 4 km and we intend to choose at random areas of 100 m × 100 m.
We can define a pair of coordinates (i, j), with $i = 1, 2, \cdots, 50$ and $j = 1, 2, \cdots, 40$.
Then the random choice of, for example, 10 plots could be done as follows:

```
> ndim<-10; nlinhas<-seq(1:50);ncol<-seq(1:40)
> i<-sample(nlinhas,ndim);j<-sample(ncol,ndim)
> plots<-cbind(i,j);plots
```

```
      i  j
 [1,] 27 34
 [2,] 14  3
 [3,] 36  4
 ...
```

Sampling

Definition 1

The random variables X_1, X_2, X_n constitute a random sample of dimension n, which is usually represented by $\underline{X} = (X_1, X_2, \ldots, X_n)$, taken from a population X, if they are mutually independent and have the same distribution as X.

Let $(\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots \boldsymbol{x_n})$ be a sample of n observations of the characteristic, obtained after a sampling process.

Each of those values is an observed value of the $\boldsymbol{n}$ variables $(X_1, X_2, \cdots X_n)$ which are "replicas" of the variable $\boldsymbol{X}$. So it is a concrete sample.

Theory of Estimation–Introduction

The Statistical Inference is intended to address two major problems:

- to calculate approximate values (estimates) and to get confidence intervals for unknown population parameters;
- to formulate hypotheses and check for agreement between what is assumed and the facts - Tests of Hypothesis

The first problem is in the domain of Theory of Estimation. If:

- if we want to infer for the unknown parameter value θ using the knowledge, even approximately, of the population distribution, we are in a (parametric estimation);
- the unknown distribution function of the variable under study, F or parameters without the assumption of knowledge of a model for population (non-parametric estimation).

Point estimation: estimator and estimate

Given a random sample $(X1, X2, ..., Xn)$, a **Statistic** is a function of the random sample that does not involve unknown parameters.

Definition

Given an unknown parameter θ, an **estimator** of is a Statistic that for each observed sample gives a value that estimates θ, which we call an **estimate** of $\underline{\theta}$.

$$\begin{array}{cl} \Theta^*(X_1, X_2, ..., X_n) & \text{is an estimator} \\ \theta^*(x_1, x_2, ..., x_n) & \text{is an estimate} \end{array}$$

Make no confusion

- estimator - random variable

- estimate - approximate value of the parameter, obtained by the value of the estimator using an observed sample

Point estimation: estimator and estimate

Examples of estimators:

i) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i/n$ e ii) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Given, for example, the sample $(1, 2, 0, 3, 1, 5)$

The estimates associated with those estimators are:

i) $\overline{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = 2$ e ii) $s^2 = \frac{1}{5}\sum_{i=1}^{6}(x_i - \overline{x})^2 = 3.2$

Let us consider the case of having a given population with a distribution $F(x|\theta)$, with an unknown parameter θ.

Since we can define several estimators of a parameter, we have the problem of choosing, if possible "the best". Then one has to consider certain properties that an estimator should verify.

Properties of point estimators

Let's look at some properties:

1. Consistency

An estimator Θ is said to be convergent for the parameter θ if $\boldsymbol{\Theta^* \xrightarrow{P} \theta}$.

It is proved that this is a sufficient convergence condition for an estimator Θ^* para θ que

$$Var(\Theta^*) \to 0 \qquad \mathrm{e} \qquad E(\Theta^*) \to \theta$$

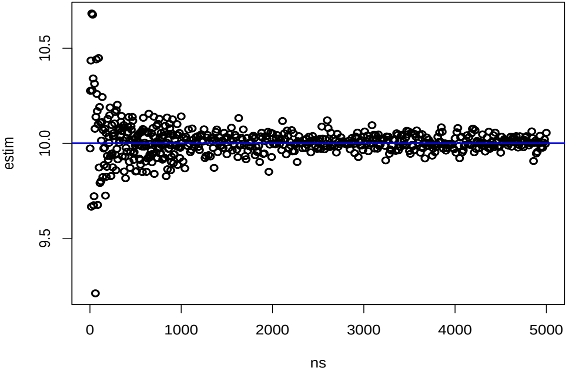
Let us see the following example illustrating the convergence of an estimator.

Example

Consider a population $\mathcal{N}(10, 2)$ and see, using R, the behavior of sample mean for several sample sizes, $n = 2, 5, 10, 15, 20, ..., 1000, 1010, 1020, ..., 5000$

```
> ns <- c(2, seq(5, 1000, by = 5), seq(1010, 5000, by = 10))
> estim <- numeric(length(ns))  # we could also do estim<-c()
> for (i in 1:length(ns)) {
+ amostra <- rnorm(ns[i], 10, 2)
+ estim[i] <- mean(amostra)
+ }
> plot(ns, estim,lwd=2)
> abline(h = 10,lwd=2,col=4)
```

Illustrating Convergence



Properties of point estimators

2. Unbias

The estimator Θ^* is said to be unbiased if $\boldsymbol{E(\Theta^*) = \theta}$.

It is actually a property that many estimators do not have and hence having more interest in defining a measure of the difference between the estimator and the parameter that it intends to estimate.

It is the **bias** that we will represent by $\boldsymbol{b_\theta(\Theta^*)}$ and is defined as

$$\boldsymbol{b_\theta(\Theta^*) = E(\Theta^*) - \theta}$$

Properties of point estimators

3. Mean Squared Error

The Mean Squared Error of the estimator $\mathbf{\Theta^*}$ a $\boldsymbol{EQM\left(\Theta^*\right) = E\left[\left(\Theta^* - \theta\right)^2\right]}$.

This property is one of the most used criteria to compare estimators.

It is very easy to show that

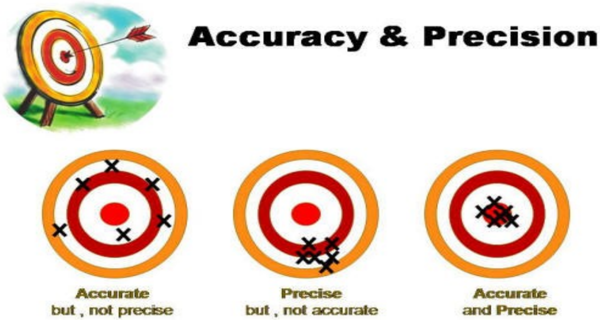
$$EQM\left(\Theta^*\right) = Var\left[\Theta^*\right] + \left[b_\theta(\Theta^*)\right]^2$$

If Θ^* is an unbiased estimator of a parameter then the mean squared error $\equiv$ variance, i.e.

$$E\left[\left(\Theta^* - \theta\right)^2\right] = Var\left(\Theta^*\right)$$

Precision and accuracy illustration

The term accuracy refers to the proximity of a measurement or estimate to the true value.
The term precision or variance (precision) refers to the "degree of agreement in a succession of measurements".



Properties of point estimators

Desirable situation: unbiased estimator with minimum variance. One says that the estimator is the most efficient.

An estimator of a parameter θ is the most efficient if it has the least probability of moving away from θ, i.e., with minimum dispersion.

If we can not get unbiased estimators we are interested in looking for estimators that have a minimal EQM, i.e.,

$$\forall\tilde{\Theta} \qquad \mathrm{E}\left[(\Theta^* - \theta)^2\right] \leqslant \mathrm{E}\left[\left(\tilde{\Theta} - \theta\right)^2\right] \qquad \forall\theta$$

There is an established criterion in the so-called inequality of Fréchet-Cramer-Rao which, under certain conditions, provides a lower limit for the EQM of an estimator (for more information see Murteira e Antunes (2012) e Casella e Berger (2002))

Methods of Estimation

So far we have talked about estimators and the properties they must have. Interested in having procedures estimators with good properties.

Let us then speak of main parametric estimation methods.

Among the **parametric estimation methods** we will only refer to two :

The method of Moments e
The Maximum Likelihood Method

The method of Moments

Introduced by Karl Pearson in the twentieth century, it was the first method of estimation that appeared and that has a very simple philosophy.

This method consists of:

– to consider as estimators of the unknown parameters the solutions of the equations that are obtained by matching the theoretical moments to the empirical moments.

It is a general method, with the only condition that the distribution of the underlying population has a sufficient number (theoretical) moments.

The method of Moments

Let $\theta_1, ..., \theta_k$ be unknown parameters of the r. v. X.
The method of moments consists of equalizing theoretical moments and empirical moments, i.e.,

$$\begin{array}{lll} E[X] = m'_1 & \text{c/} & m'_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \\ E[X^2] = m'_2 & \text{c/} & m'_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \vdots & & \\ E[X^k] = m'_k & \text{c/} & m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \end{array}$$

$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ are the empirical moments, calculated based on the sample $(x_1, ..., x_n)$.

Those equalities give us estimates which are the concretization of estimators with the corresponding expressions.

The method of Moments– examples

Example

Let us consider $X \frown \mathcal{N}(\mu, \sigma)$. Obtain the moment estimators of μ and σ^2?

We have :

$$\begin{array}{lcl} E[X] = \mu & \text{e} & M_1' = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X} \\ E[X^2] = \sigma^2 + \mu^2 & \text{e} & M_2' = \frac{1}{n}\sum_{i=1}^n X_i^2 \end{array}$$

so

$$\begin{array}{l} \mu^* = \overline{X} \\ (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^n X_i^2 - \mu^{*2} \Rightarrow (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \end{array}$$

Exercise

Let be $X \frown \mathcal{P}(\lambda)$. Obtain the monent estimator of λ.

The Method of Moments

The calculations are not complicated, but the empirical moments of high order appear with high exponents, when there are many parameters, leading to unstable estimates.
Therefore, as a rule of thumb, you should avoid using method of moments for more than four parameters.

Remark: The estimators obtained by the method of moments are less efficient than the maximum likelihood estimators, which we are now talking about.

The Likelihood

Let X be a r.v. whose distribution depends on an parameter θ, unknown, and let $(X_1, ..., X_n)$ be a random sample. Let $(x_1, ..., x_n)$ be the observed sample.

Definition

The **likelihood of the sample**, represented by $\boldsymbol{\mathcal{L}(\theta|x_1, x_2, ..., x_n)}$ is

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \qquad \text{caso contínuo}$$

$$P(X_1 = x_1, ..., X_n = x_n|\theta) = \prod_{i=1}^{n} P(X_i = x_i|\theta) \quad \text{caso discreto}$$

The Maximum Likelihood Method (MLM)

The maximum likelihood method, proposed by Fisher in 1922 and developed in 1925 consists of choosing as an estimate of θ the value that **maximizes the likelihood** $\boldsymbol{\mathcal{L}(\theta|x_1, ..., x_n)}$, given a sample $(x_1, ..., x_n)$.

Let us see an illustration withR

A very simple example

In a small lake that we have in the garden there are some fish. I would like to know how many fish the lake will have. How can I "estimate"?
A common procedure is the so-called capture-recapture method, of which they have certainly spoken when they gave the hypergeometric probability model. How is the method described?
- We caught M = 7 fish, we marked each one and released it afterwards.
"You'll wait a few days to go back to your habits, and we'll fish a few more."
- Let X a.v. be counted the number of captured-marked, i.e. are recaptured.
- Suppose that the second time we got $n = 4$, of which 3 are marked, then $x = 3$.

The Maximum Likelihood Method (MLM)

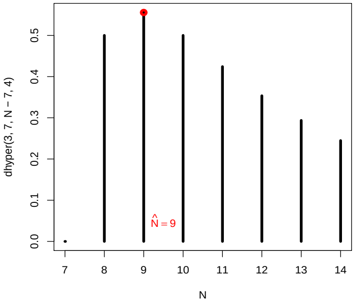
Now we want to know the number N of fish that exist in the lake, $N = M + (N - M)$;
X represents the number of fish recaptured.

$$X \frown \mathcal{H}(N, n = 4, M = 7) \qquad \text{então} \qquad P[X = x] = \frac{\binom{7}{x}\binom{N-7}{4-x}}{\binom{N}{4}}$$

If we found 3 marked fish, what is the most likely number of fish there?
We know that $N \geq 7$, but it could be $N = 8, 9, 10, 11, 12, 13, 14...$?
Sir Ronald Fisher asked: *"What is the value of N which has the highest likelihood?"*
That is, for the possible values for N what makes the above probability maximum?
Let's solve in R, calculating the probabilities and plotting

Exemplo muito simples

```
N<-seq(7,14);N
plot(N,dhyper(3,7,N-7,4),type="h",lwd=4)
points(9,dhyper(3,7,2,4),type="p",col=2,lwd=4)
text(9.5,0.05,expression(hat(N)==9),col=2)
```



The Maximum Likelihood Method (MLM)

In many situations, the likelihood functions satisfy conditions that allow the value for which the likelihood is maximum is obtained by using derivatives.

However, and since the logarithmic function is monotonous, as a general rule it is more convenient to work with the **log-likelihood function**,

$$\log \mathcal{L}(\theta|\underline{x})$$

The Maximum Likelihood Method (MLM)

So, in case there are derivatives (and for a single parameter θ), the value of the maximizer is obtained by solving:

$$\frac{d\log\mathcal{L}}{d\theta}=0 \qquad \mathrm{e} \qquad \frac{d^2\log\mathcal{L}}{d\theta^2}<0$$

Remark that $$\frac{d\log\mathcal{L}}{d\theta}=\sum_{i=1}^{n}\frac{d\log f(x_i|\theta)}{d\theta}$$

The solution, $\hat{\theta}(x_1, ..., x_n)$ is **the maximum likelihood estimate,** which is a realization of the r.v..$\hat{\Theta}=\hat{\Theta}(X_1, ..., X_n)$.

Let's do a very simple exercise – To obtain the maximum likelihood estimator of the β parameter of the exponential model.

Properties of Maximum Likelihood Estimators

Although it may not be unbiased, the maximum likelihood estimator of a generic parameter θ is the unique efficient estimator (if there is an efficient estimator) and

— is asymptotically normal with mean value θ and asymptotic variance given by

$$\frac{1}{E\left[\left(\frac{\partial \log \mathcal{L}}{\partial \theta}\right)^2\right]} = \frac{1}{nE\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right]} = \frac{1}{-E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right]}$$

The maximum likelihood in R

In R there are functions to obtain the maximum likelihood estimates, using iterative procedures.

We have, for example, the functions

`mle()` do *package* stats4

`fitdistr()` do *package* MASS

`optimize ()` – suitable when there is only one parameter
`optim ()` – suitable when there are two or more parameters

Example

Let's consider the R– PlantGrowth data file.
Let's use the sample of observed values of the "weight" variable. Let's assume that the observations are the realization of a r.s. $(X_1, X_2, ..., X_n)$, i.i.d. $\mathcal{N}(\mu, \sigma)$

First we need to write the log-likelihood function

Important note the optimization function, by default calculates minimums, so we must write the negative log-likelihood

```
>minuslogL <- function(mu, sigma){
  -sum(dnorm(x, mean = mu, sd = sigma, log = TRUE)) }
```

To obtain the maximum likelihood estimate, the optimization algorithm requires initial values for the parameters to be estimated.

Resolution

I considered values close to the observed mean and variance, in this case 5 and 0.5, respectively, to illustrate.

```
>data(PlantGrowth)
>x <- PlantGrowth$weight
>media<-mean(x); dp<-sd(x);media;dp
>library(stats4)
library(stats4)
>MLest1 <- mle(minuslogL, start = list(mu = media, sigma = dp))
 # outra tentativa
>MLest2 <- mle(minuslogL, start = list(mu = 5.1, sigma = 0.5))
>summary(MLest1);summary(MLest2)
Maximum likelihood estimation
Call:
mle(minuslogl = minuslogL, start = list(mu = media, sigma = dp))
Coefficients:
       Estimate Std. Error
mu     5.0730000 0.12586801
sigma 0.6894075 0.08900152
-2 log L: 62.82084
Maximum likelihood estimation
```

Most Common Estimators and Sampling Distributions –Summary

Parameters that we are going to refer to, their estimators and estimates

Parameter to estimate	Estimator	Estimate
μ	$\overline{X}=\frac{\sum_{i=1}^{n} X_i}{n}$	$\overline{x}=\frac{\sum_{i=1}^{n} x_i}{n}$
σ^2	$S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$	$s^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$
p	$\widehat{P}=\frac{X^{(a)}}{n}$	$\hat{p}=\frac{x^{(b)}}{n}$
$\mu_1 - \mu_2$	$\overline{X_1} - \overline{X_2}$	$\overline{x_1} - \overline{x_2}$
$\sigma_1^2 \,/\, \sigma_2^2$	$S_1^2 \,/\, S_2^2$	$s_1^2 \,/\, s_2^2$
$p_1 - p_2$	$\widehat{P_1} - \widehat{P_2}$	$\hat{p}_1 - \hat{p}_2$

[a] X - r.v. that counts ... and ˜ [b] x - observed number of successes in the sample of dimension n.