Basic statistics

Hugo Harari-Kermadec

EPOG - Econometrics

September, the 26th, 2019

1 Introduction

- Start R
- Clic in the consol (down left) and type 1+23
- Set `x<-3` and calculate x^2.

- Start R
- Clic in the consol (down left) and type 1+23
- Set `x<-3` and calculate x^2.
- Set `income <- c(873,1050,1401,4102,2350)` and calculate the mean, var, sd, max, min, range, sum, length.

- Start R
- Clic in the consol (down left) and type 1+23
- Set `x<-3` and calculate x^2.
- Set `income <- c(873,1050,1401,4102,2350)` and calculate the mean, var, sd, max, min, range, sum, length.
- and with `age<-c(21,22,NA,51)`.

- Start R
- Clic in the consol (down left) and type 1+23
- Set `x<-3` and calculate x^2.
- Set `income <- c(873,1050,1401,4102,2350)` and calculate the mean, var, sd, max, min, range, sum, length.
- and with `age<-c(21,22,NA,51).Try mean(age,na.rm=T)`

- Start R
- Clic in the consol (down left) and type 1+23
- Set `x<-3` and calculate x^2.
- Set `income <- c(873,1050,1401,4102,2350)` and calculate the mean, var, sd, max, min, range, sum, length.
- and with `age<-c(21,22,NA,51).Try mean(age,na.rm=T)`
- Set `nationality<-c("French","Spanish","Brasilian")`

- Start R
- Clic in the consol (down left) and type 1+23
- Set `x<-3` and calculate x^2.
- Set `income <- c(873,1050,1401,4102,2350)` and calculate the mean, var, sd, max, min, range, sum, length.
- and with `age<-c(21,22,NA,51).Try mean(age,na.rm=T)`
- Set `nationality<-c("French","Spanish","Brasilian")`

Copy and paste everything !

create a new script by clicking on the green + just below the *File* menu. Save it as `basics.R`
Comment your script with ♯. You can organize the script with titles: ♯♯
You can run the script directly with `Ctrl+Enter` or `Cmd+Enter`. This will execute the whole script or the selected part.

Definition (Population)

The population is the given of all the individuals that could be in the sample. It's not "what we'd like to talk about".

Example (French actives' income: population)

Let's consider the $N = 30$ millions French actives. For each individual j, let's a_j be her income. The population is then $[|1; N|]$ whose incomes are $\{a_1, \ldots, a_N\}$.

- Download the data from `http://www.cepn-paris13.fr/epog/?p=1111` and save it in a specific folder
- Import in R with the button *Import Dataset* (up right), "from text (base)"
- copy and paste the produed code in basics.R

```
population <- read.csv("C:/Users/harari/Desktop/Dropbox/ens
```

Definition (i.i.d. sample)

$J_1, \ldots, J_n$ are n individual drown independently and with the same sampling scheme. This implies sampling with replacement.

Example (French actives' income: sample)

Sampling is performed on individuals: we sample j's, $J_1, J_2, \ldots$
The income is then deterministic for each individual: $A_1 = a_{J_1}$

Definition (Parameter)

A parameter a is a real (a numerical value) of interest.

Example (Expected income)

Let's consider a, the mean of French actives' incomes:

$$a = \frac{1}{n}\sum_{j=1}^{n} a_j, \text{ where } a_j \text{ is the income of individual j}$$

and a_M and a_F, the means of male and female incomes:

$$a_M = \frac{1}{n_M}\sum_{j \text{ male}}^{n} a_j, \qquad a_F = \frac{1}{n_F}\sum_{j \text{ female}}^{n} a_j$$

Definition (Estimator)

An estimator $\hat{a}_n$ of a parameter a is a function of the data, aiming to be close to the real value of the parameter.

Example (Estimator of the expected income: sample mean)

Suppose we observe a sample of incomes $a_1, \ldots, a_100$

$$\hat{a}_{100} = \frac{1}{100} \sum_{j=1}^{100} a_j,$$

with 45 males and 55 females

$$\hat{a}_{M,45} = \frac{1}{45} \sum_{j \text{ male}}^{n} a_j, \qquad \hat{a}_{F,55} = \frac{1}{55} \sum_{j \text{ female}}^{n} a_j$$

How good is my estimator?

Definition (Confidence interval - CI)

$[l_n; u_n]$ is a 95% confidence interval for a if and only if

$$\mathbb{P}(a \in [l_n \ ; \ u_n]) = 0,95.$$

The parameter is in the interval with probability 0,95.

How good is my estimator?

Definition (Confidence interval - CI)

$[l_n; u_n]$ is a 95% confidence interval for a if and only if

$$\mathbb{P}(a \in [l_n \ ; \ u_n]) = 0,95.$$

The parameter is in the interval with probability 0,95.

Most CI are only asymptotic:

$$\mathbb{P}(a \in [l_n \ ; \ u_n]) \xrightarrow{n\to\infty} 0,95.$$

Example (Election survey)

Macron' popularity is estimated to 23%, on the basis of a sample of size 1.000, means that with probability 0.95 they are in $[20\% \ ; \ 26\%]$

CI: formula

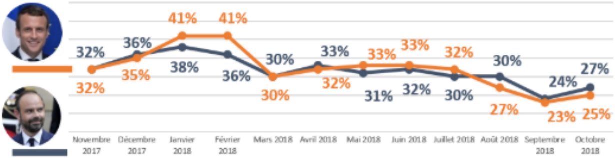
CI around the mean writes:

$$[\hat{a}_n - q_{0.95}\frac{\sigma}{\sqrt{n}} \;;\; \hat{a}_n + q_{0.95}\frac{\sigma}{\sqrt{n}}]$$

where $q_{0,95} \approx 2$ is the 95 quantile of $\mathcal{N}(0,1)$,
and σ is the standard deviation in the sample.

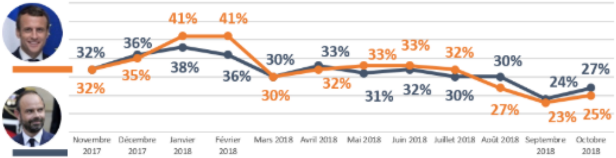
Example: Macron's popularity (YouGov)

Macron's popularity, at 27% in August, has fallen in september to 23% and risen to 25% in october (1 006 individuals)



Example: Macron's popularity (YouGov)

Macron's popularity, at 27% in August, has fallen in september to 23% and risen to 25% in october (1 006 individuals)



95% confidence intervals are:

$$\left[24 - \frac{1/4}{\sqrt{1007}} q_{0,95} \ ; \ 27 + \frac{1/4}{\sqrt{1007}} q_{0,95}\right] = [24 \ ; \ 30]$$

$$\left[19 - \frac{1/4}{\sqrt{1007}} q_{0,95} \ ; \ 23 + \frac{1/4}{\sqrt{1007}} q_{0,95}\right] = [20 \ ; \ 26]$$

A small change, up to 3 pts, is not statistically significant.

Definition (Test)

If a given value m_0 for the parameter is inside the 95% confidence interval, the "null" assumption that H_0 : "$m = m_0$" holds at confidence 95%. Else, H_0 is rejected.
The probability to reject by mistake H_0 is noted α, here 5%.

Definition (Test)

If a given value m_0 for the parameter is inside the 95% confidence interval, the "null" assumption that H_0 : "$m = m_0$" holds at confidence 95%. Else, H_0 is rejected.
The probability to reject by mistake H_0 is noted α, here 5%.

The assumption of Macron's popularity stability between sept. and oct. ($m = 0.23$) holds, at 95% confidence.

Test main caracteristics

	Accepts H_0	Rejects H_0
H_0 true	True negative $\mathbb{P}_{H_0} = 1 - \alpha$ Test level	False positive $\mathbb{P}_{H_0} = \alpha$ Type I error
H_1 true (H_0 false)	False negative $\mathbb{P}_{H_1} = \beta$ Type II error	True positive $\mathbb{P}_{H_1} = 1 - \beta$ Test's power

More precisely, the assumption of stability is weaker from aug. to sept. (27% $\to$ 23%) than from sept. to oct. (23% $\to$ 25%)

More precisely, the assumption of stability is weaker from aug. to sept. (27% $\rightarrow$ 23%) than from sept. to oct. (23% $\rightarrow$ 25%) The p-value quantifies that strength:

Definition (p-value)

The p-value is the larger error α such as assumption H_0 holds. The smaller p-value, the weaker the assumption.

In june, $p = 0.1$ and in september, $p = 0.05$.