Principles of Statistical Inference

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR

Santiago Pérez-Hoyos i Alex Sánchez-Pla

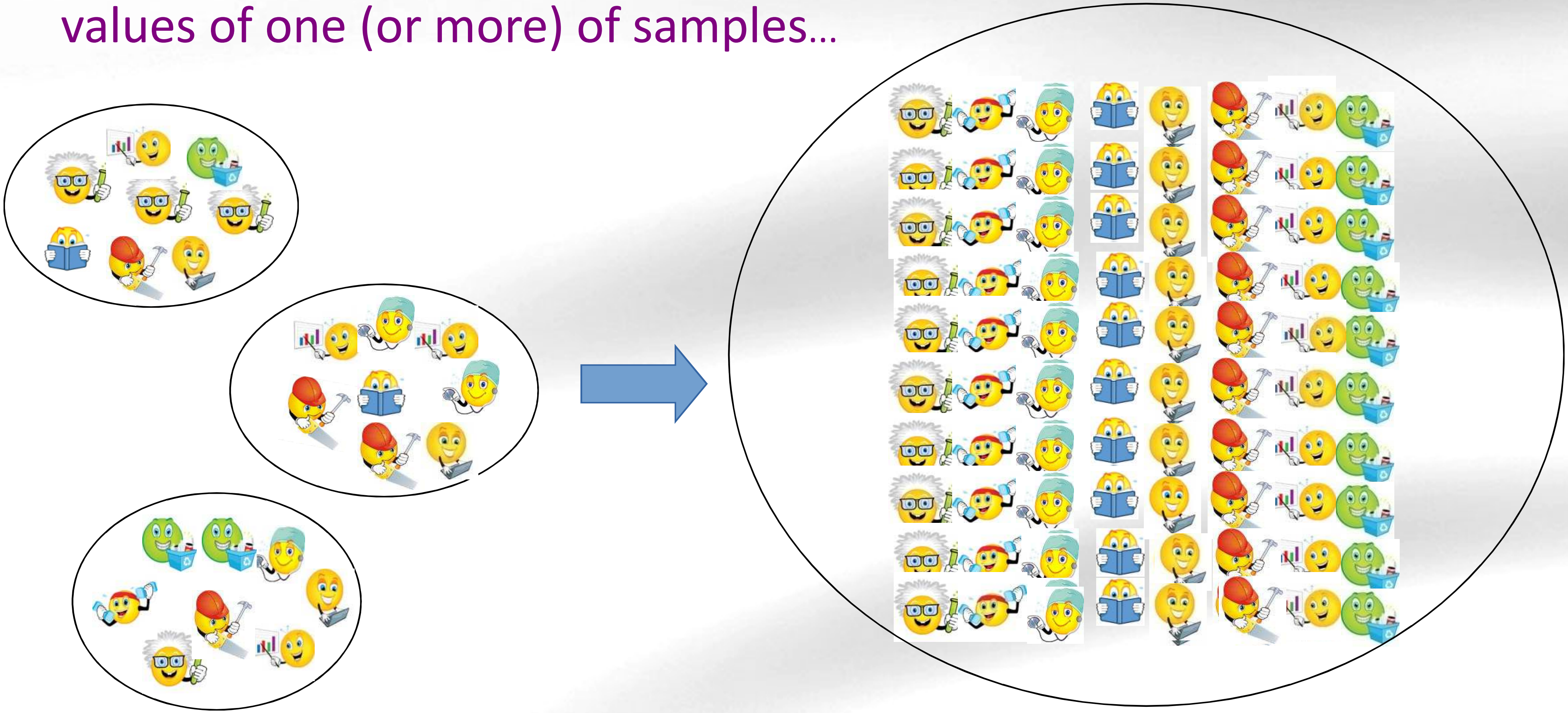
santi.perezhoyos@vhir.org

alex.sanchez@vhir.org

Vall d'Hebron Institut de Recerca

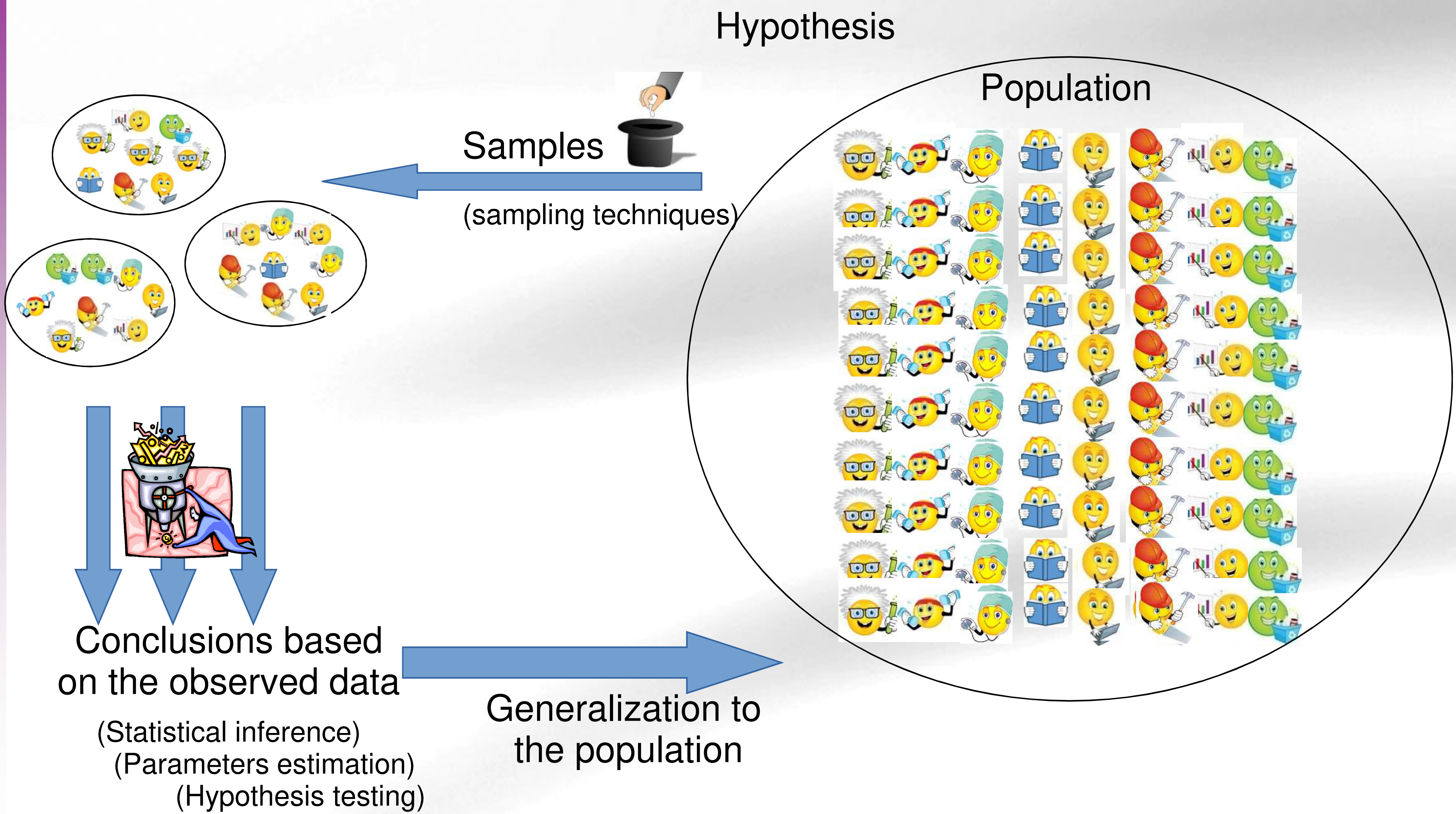
The objective of statistical inference

Taking the observed (measured) values of one (or more) of samples…



... Determine ("***infer***") the properties of the entire population.

The objective of statistical inference



Estimation

- The aim of estimation is ***to infer properties (parameters) of the distribution of population data from sample data***
- **Some key concepts**
  - **Point estimate:** Give a numerical value to the parameter of interest.
  - **Estimator:** Mathematical function to obtain the estimate
  - **Interval Estimation:** Give two values between which is the value of the population parameter with a preset confidence level (or probability)
  - **Random error:** Difference between estimation and real value if the sample is random

Point estimation (1)

- Data from qualitative variables
  - Parameter: Probability to observe a certain category
  - Estimate: Sample proportion: % of that category in the sample
  - Example: *In the Osteoporosis dataset, what is the probability of observing a woman without ostheoporosis*

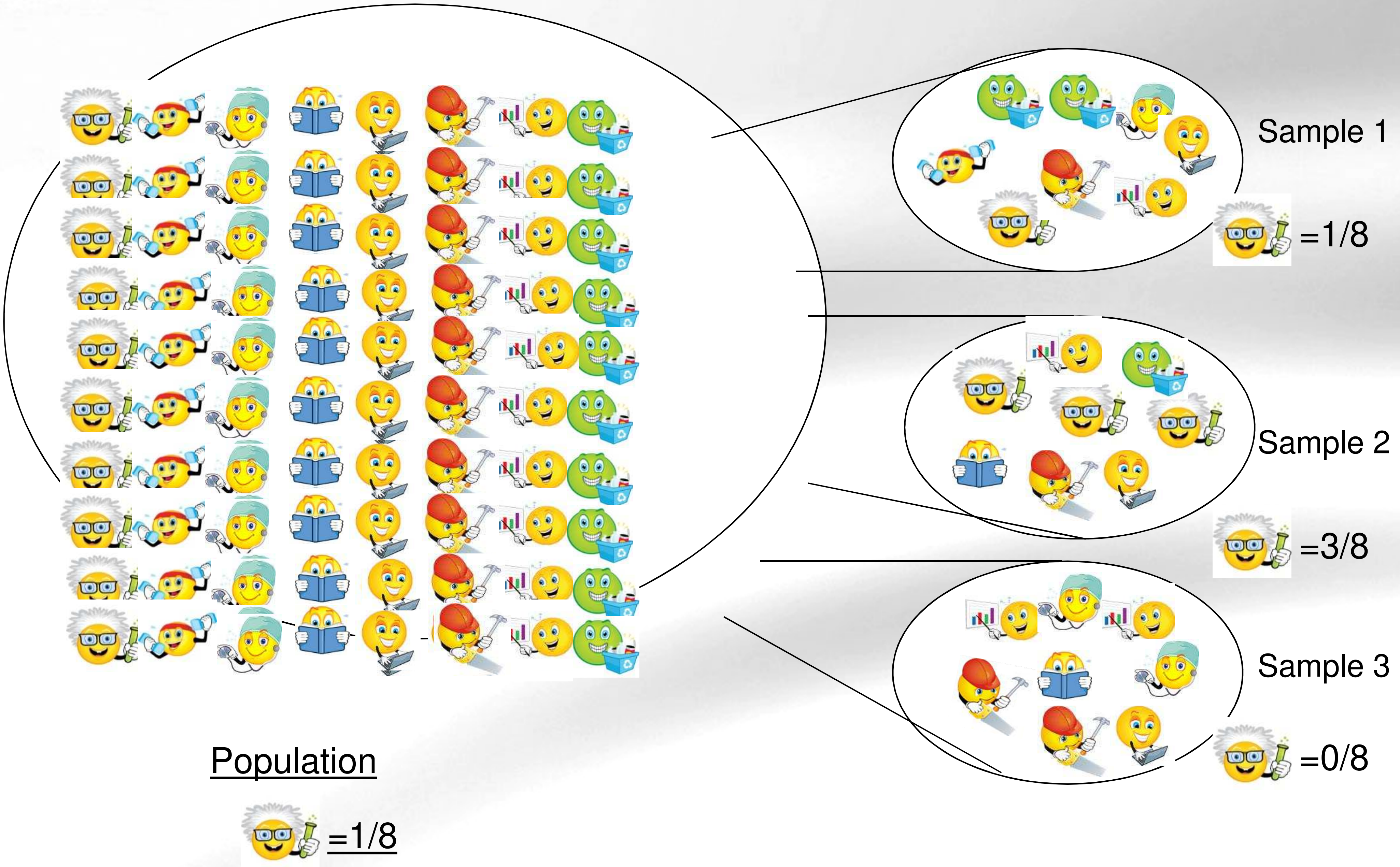
Point estimation (II)

- Data from quantitative variables
  - Population parameters: μ, σ, etc.
  - Population parameters:
    - Estimate the mean, μ, with the sample mean, $\overline{X}$
    - Estimate , σ with the sample standard deviation, $\hat{s}$

Exercise

- *In the osteoporosis dataset (osteo100) estimate the mean bone density (BUA)*
  - *for all the population indistinctly*
  - *depending on the CLASSIFIC variable*

Biological variability. Sampling



Sampling distribution

- Population is 5 Children with age

  $x_1=6,\ x_2=8,\ x_3=10,\ x_4=12,\ x_5=14$

  - Mean $\mu=10$
  - Variance $\sigma^2=8$
- Extract all possible samples with replacement and compute the mean in each sample

In this problem we can compute the population parameters because we know all the population values!!!

25 Samples n=2

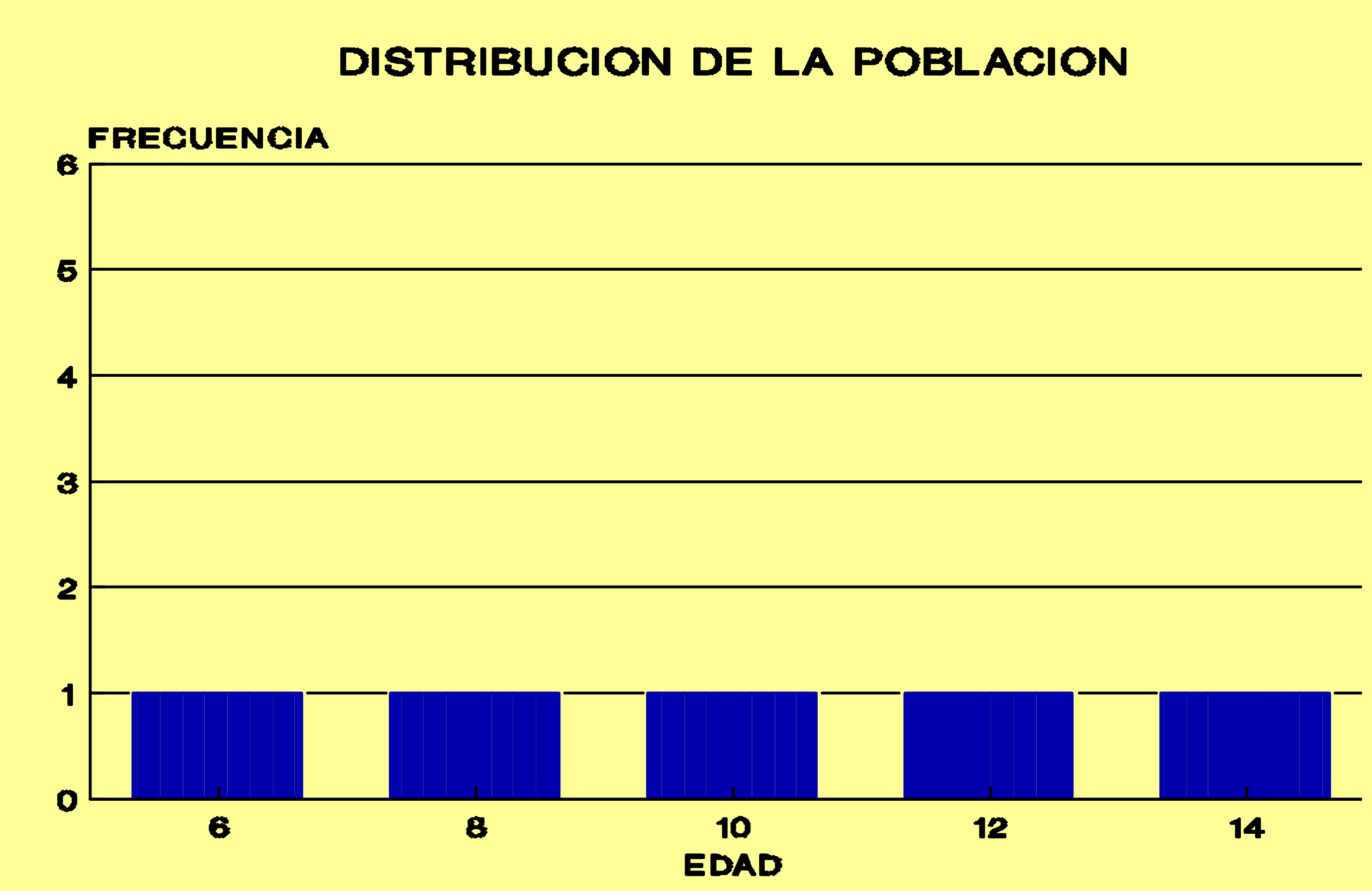
		Second Data				
Fist Data		**6**	**8**	**10**	**12**	**14**
	6	6,6 (6)	6,8 (7)	6,10 (8)	6,12 (9)	6,14 (10)
	8	8,6 (7)	8,8 (8)	8,10 (9)	8,12 (10)	8,14 (11)
	10	10,6 (8)	10,8 (9)	10,10 (10)	10,12 (11)	10,14 (12)
	12	12,6 (9)	12,8 (10)	12,10 (11)	12,12 (12)	12,14 (13)
	14	14,6 (10)	14,8 (11)	14,10 (12)	14,12 (13)	14,14 (14)

Frequency table

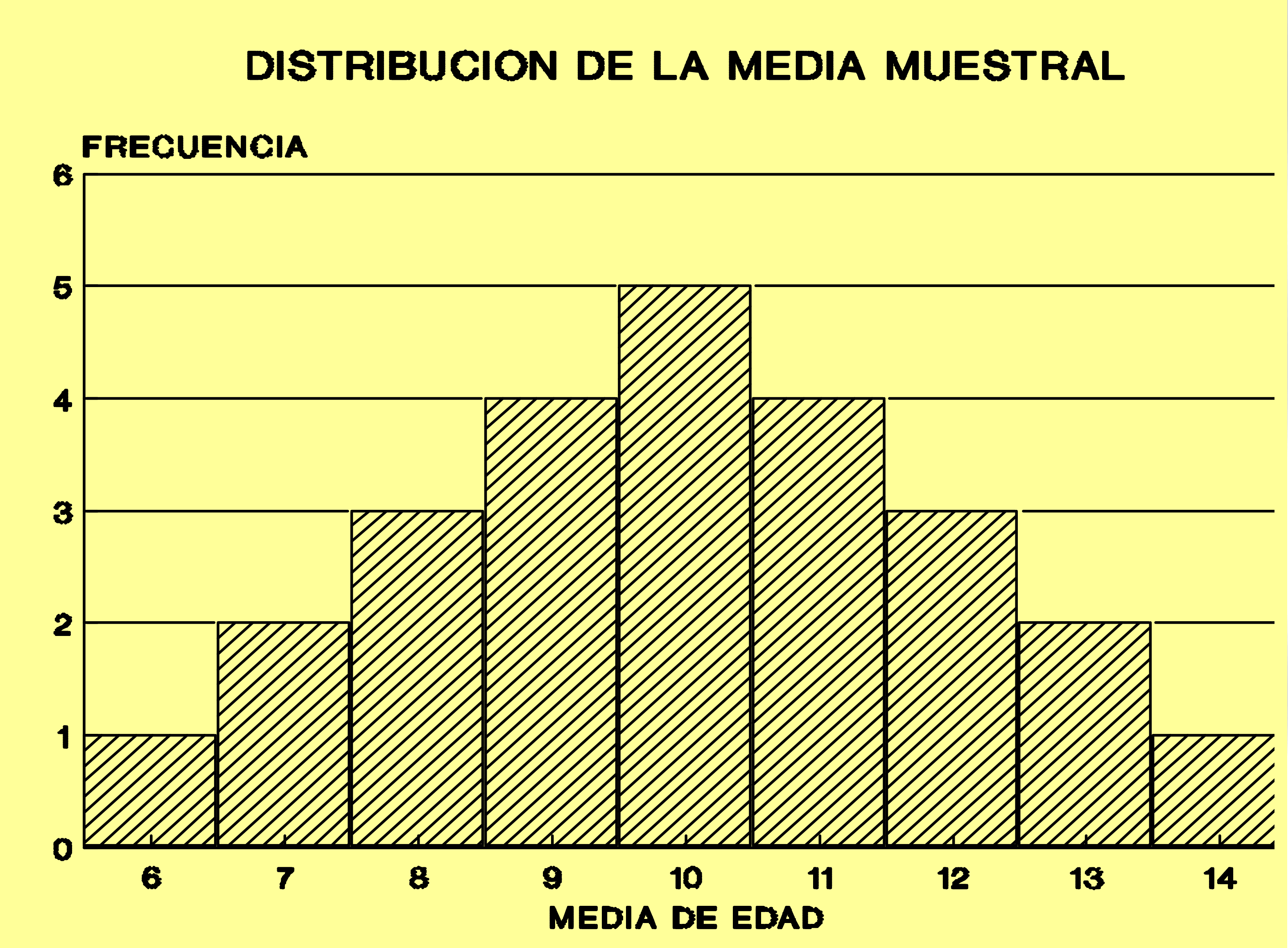
media	frecuencia	frec relativa
6	1	1/25
7	2	2/25
8	3	3/25
9	4	4/25
10	5	5/25
11	4	4/25
12	3	3/25
13	2	2/25
14	1	1/25

Vall d'Hebron Institut de Recerca

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA



Histograma


DISTRIBUCION DE LA MEDIA MUESTRAL
FRECUENCIA
0
1
2
3
4
5
6
6
7
8
9
10
11
12
13
14
MEDIA DE EDAD

Summary

- Mean of 25 sample means

$$\mu_{med}=(6+7+...+14)/25=10$$

- Variance of 25 sample means

$$\sigma^2_{med}=\{(6-10)^2+(7-10)^2+..+(14-10)^2\}/25=4$$

- The mean of sample means is population mean

$$\sigma_{med}^2= \sigma^2/2=8/2=4$$

- Variance of 25 sample means equals population variance divided by sample size

Standard error

- Standard deviation of the distribution of sample means
- Usually it is defined as population standard deviation divided by squared root of sample size

$$\text{standard error} = \frac{\sigma}{\sqrt{n}}$$

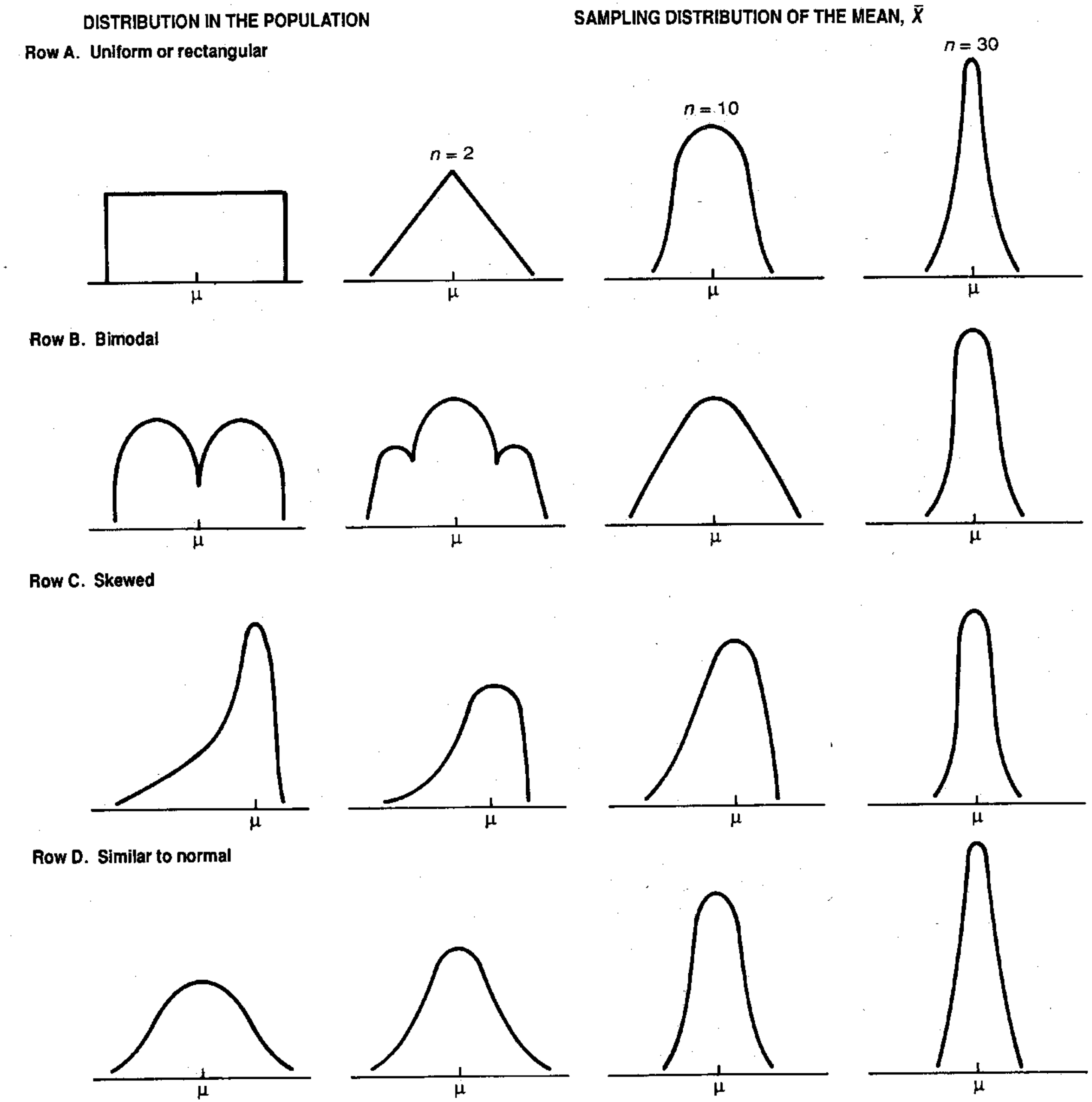


Figure 6-3. Illustration of ramifications of central limit theorem.

Unbiased estimators

- An estimator is unbiased if the mean of the sample estimates is the parameter we are looking for.
  - Sample mean and proportion are unbiased estimators of population mean and probability (percentage)
  - Sample variance is a biased estimator of population variance, but not if we divided by n-1
    - That is why computers compute sample variance dividing by (n-1) instead of dividing by n.

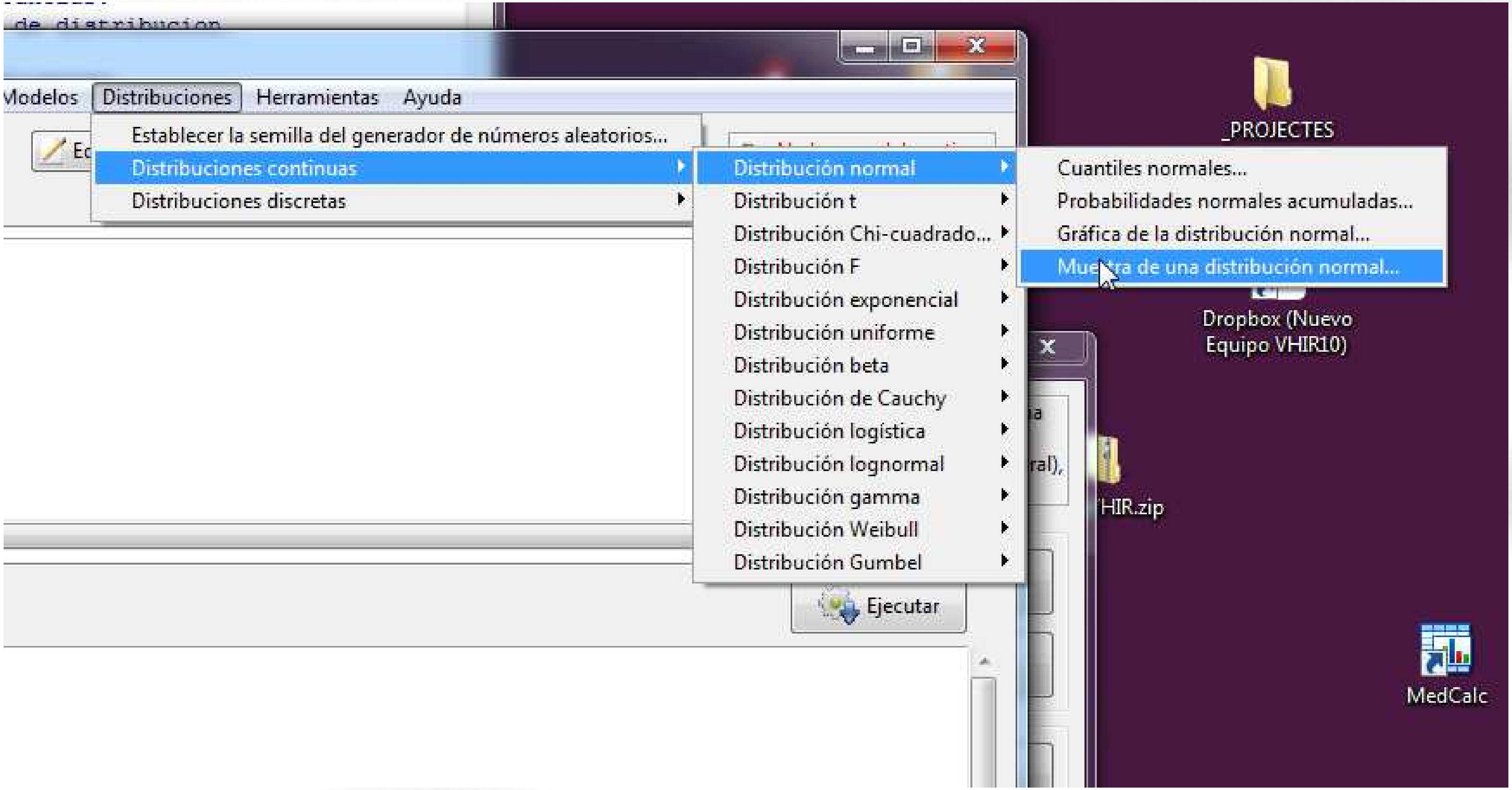
Confidence interval of Mean

- Population blood pressure in hipertensives is normally distributed with mean μ and standard deviation 12
- We extract a sample of n=186 and we observe a *sample* mean m=118,8)
- We can compute a *confidence interval* for the mean:

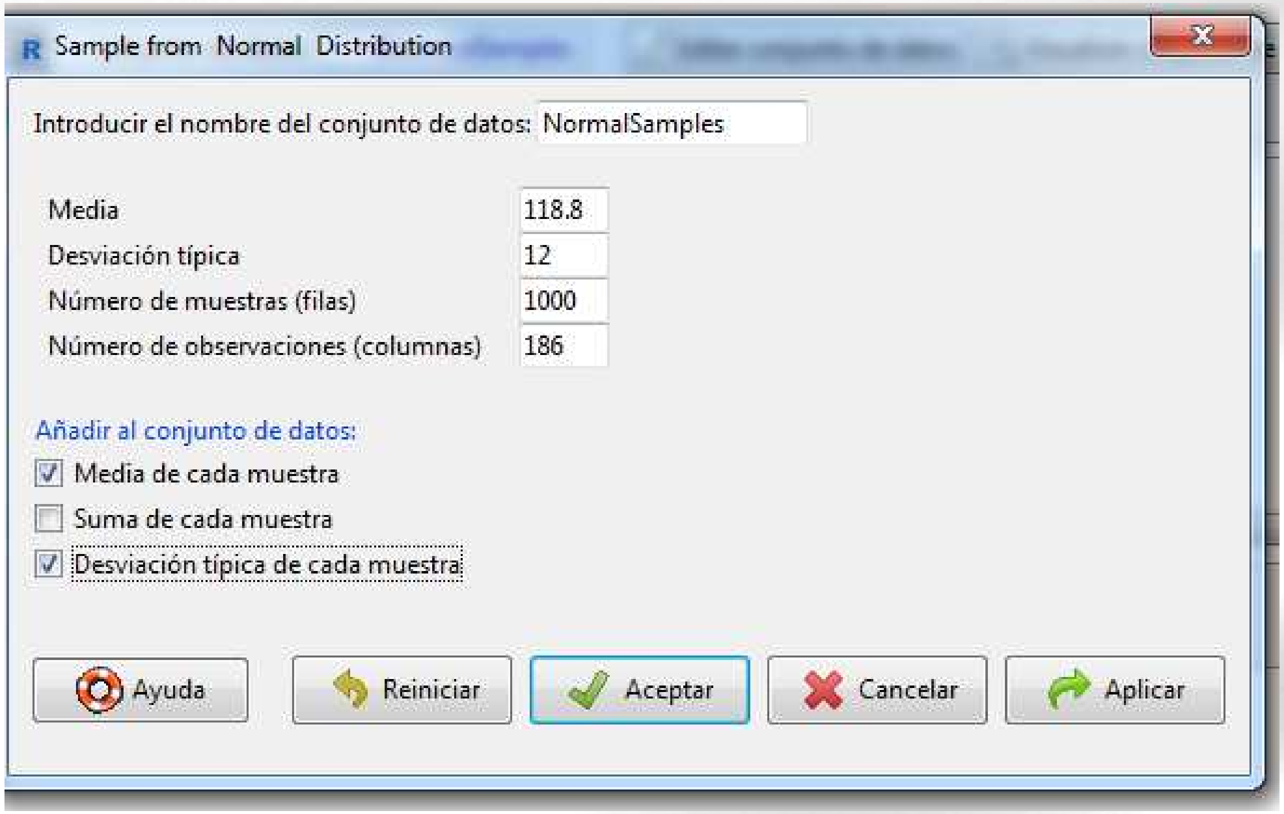
$$\bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} = 118 \pm 1,96 \times 12/\sqrt{186}$$

  – This provides an interval such that we are *highly confident* that the true population may be between the upper and lower value of the interval.
    - In practice this means that if we repeated the process of sampling and building the interval we would expect that 95% of the times it would contain the true population value

Let’s simulate 1000 samples of size 186 with mean 118.8 and standard deviation 12



Let's simulate 1000 samples of size 186 with mean 118.8 and standard deviation 12

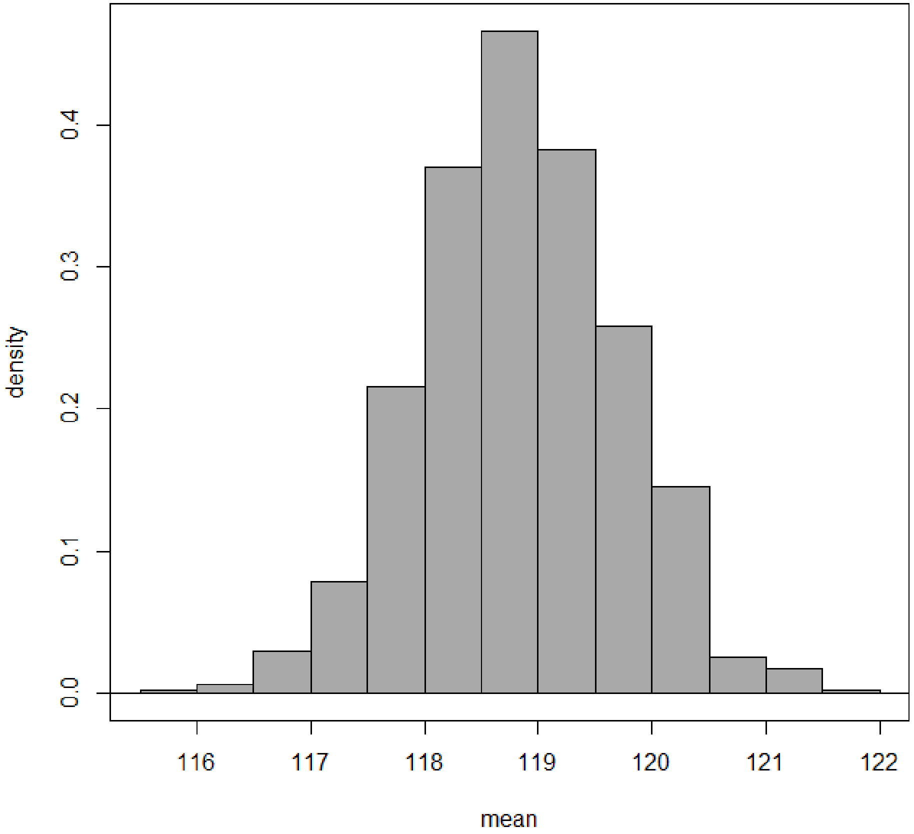


Calculate the mean and standard deviation of the mean of the 1000 samples
Calculate the mean and standard deviation of the sd of the 1000 samples

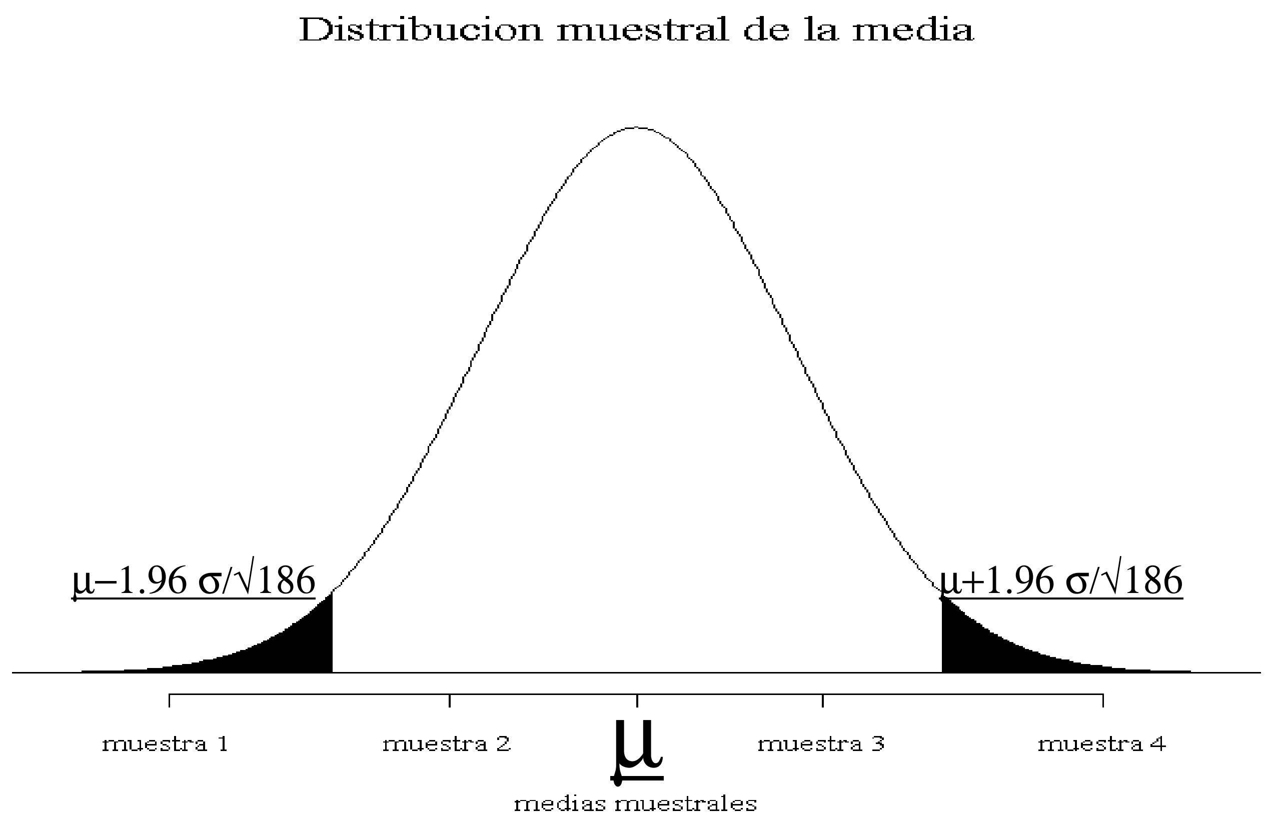
	mean	sd
mean	118.84143	0.8746236
sd	11.97547	0.6613225

$$\frac{12}{\sqrt{186}} = 0,879$$

Standard ERROR

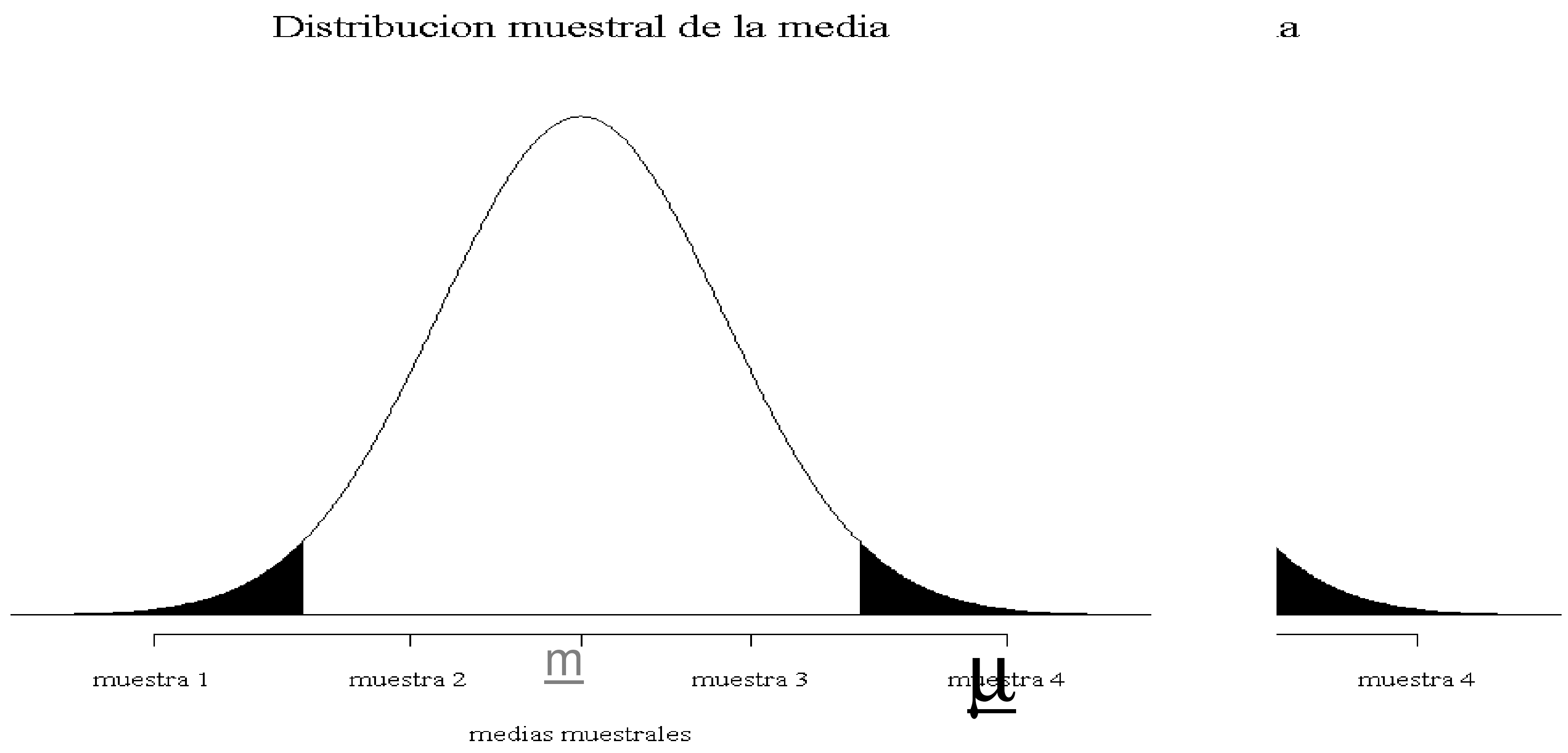


True mean =118,8



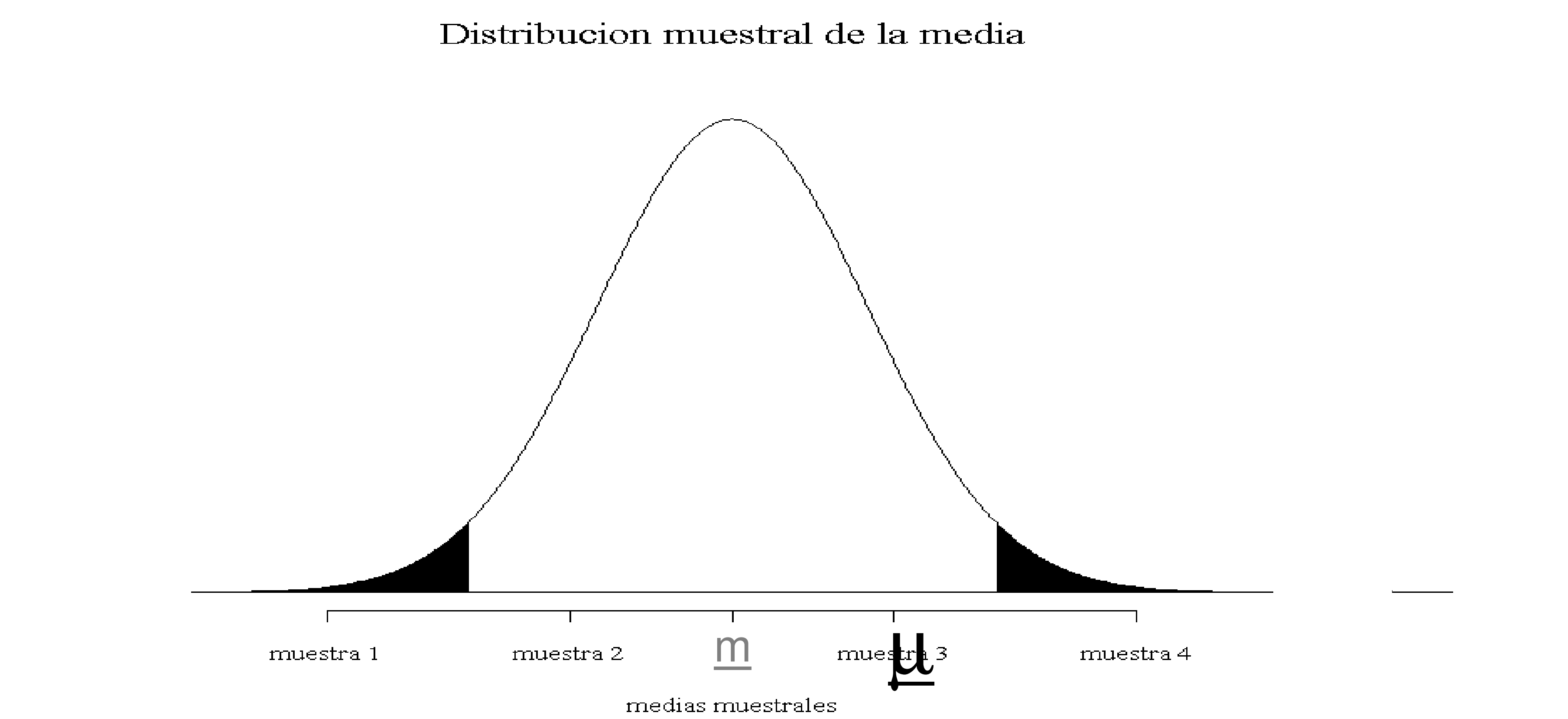
Distribucion muestral de la media
μ−1.96 σ/√186
μ+1.96 σ/√186
muestra 1
muestra 2
μ
muestra 3
muestra 4
medias muestrales

$m \pm 1.96\ \sigma/\sqrt{186}$



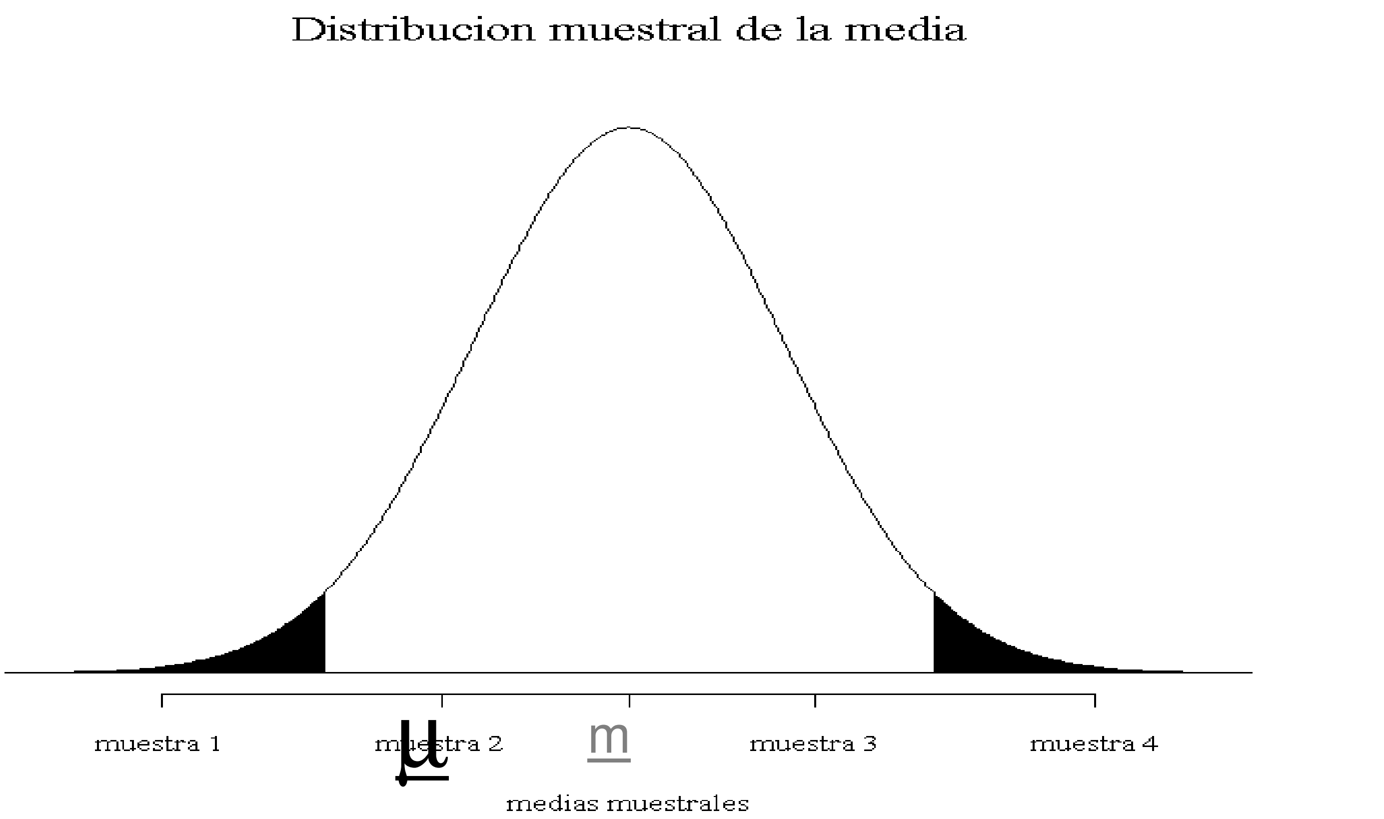
Population mean is outside de confidence interval

$m \pm 1.96\ \sigma/\sqrt{186}$



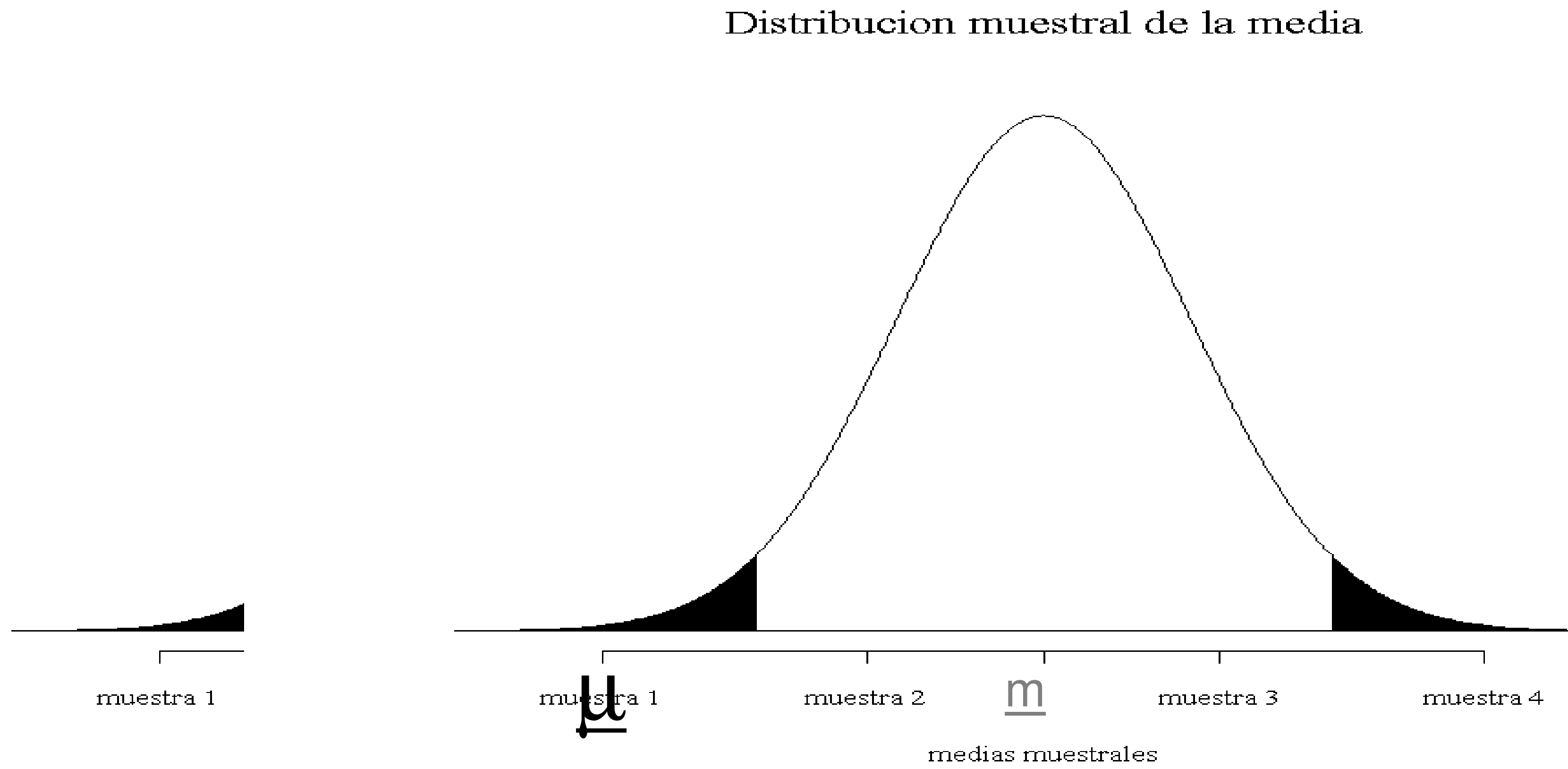
Population mean is inside confidence interval

$m \pm 1.96\ \sigma/\sqrt{186}$



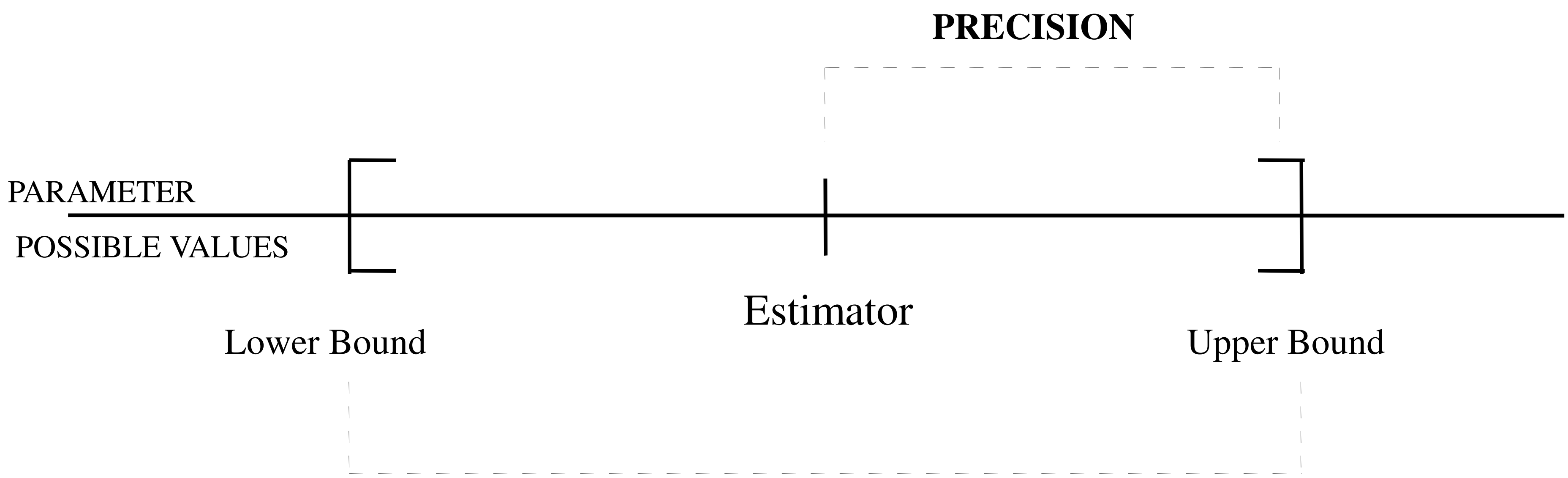
Population mean is inside confidence interval

m±1.96 σ/√186



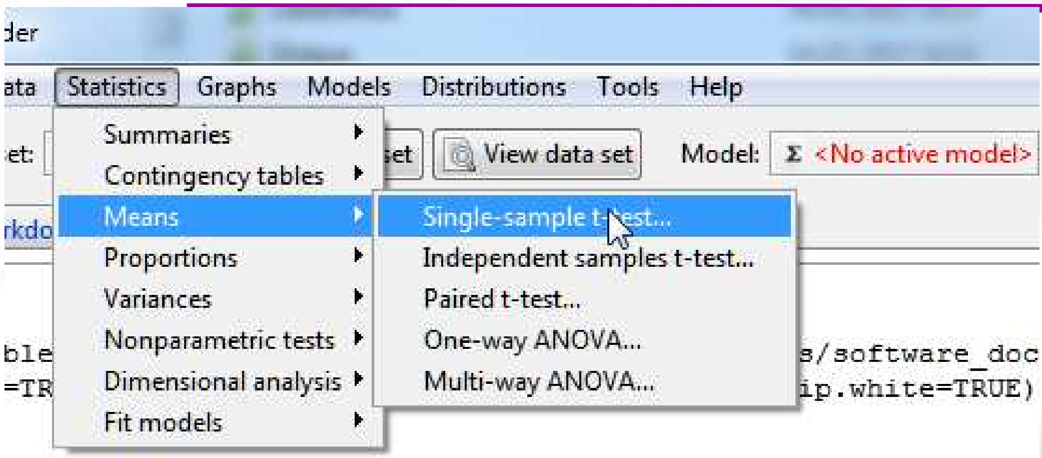
Population mean its outside confidence interval

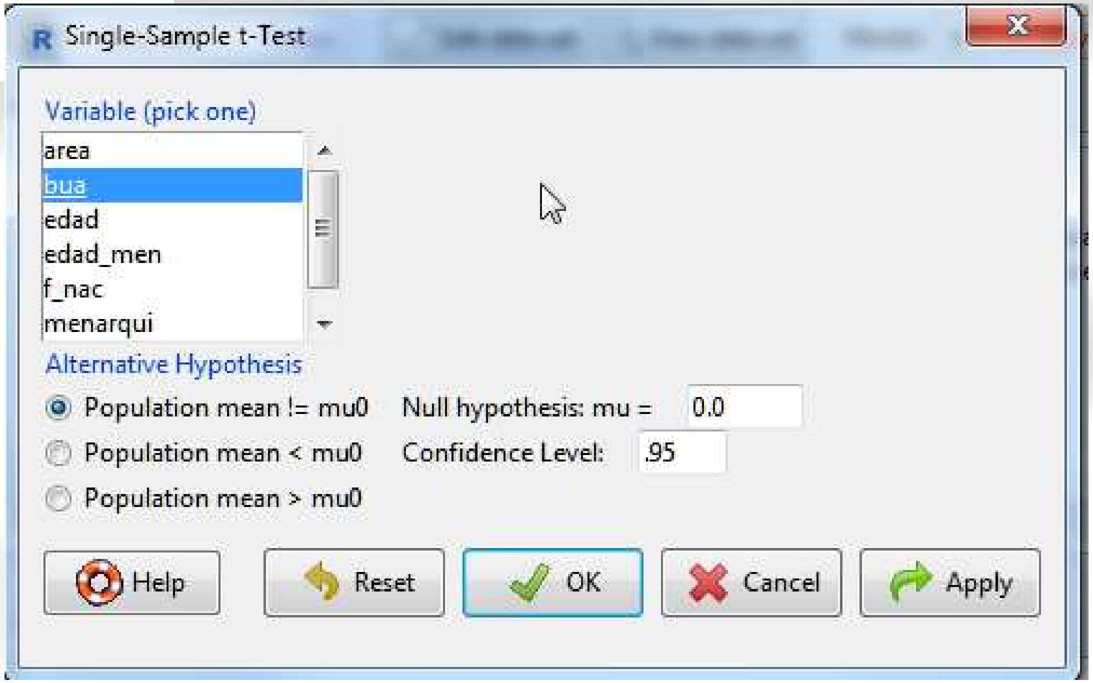
Confidence interval



Values in which we are confident that real population parameter is inside
With a prefixed confidence level (Usually 95%)

Confidence intervals in RCmdr

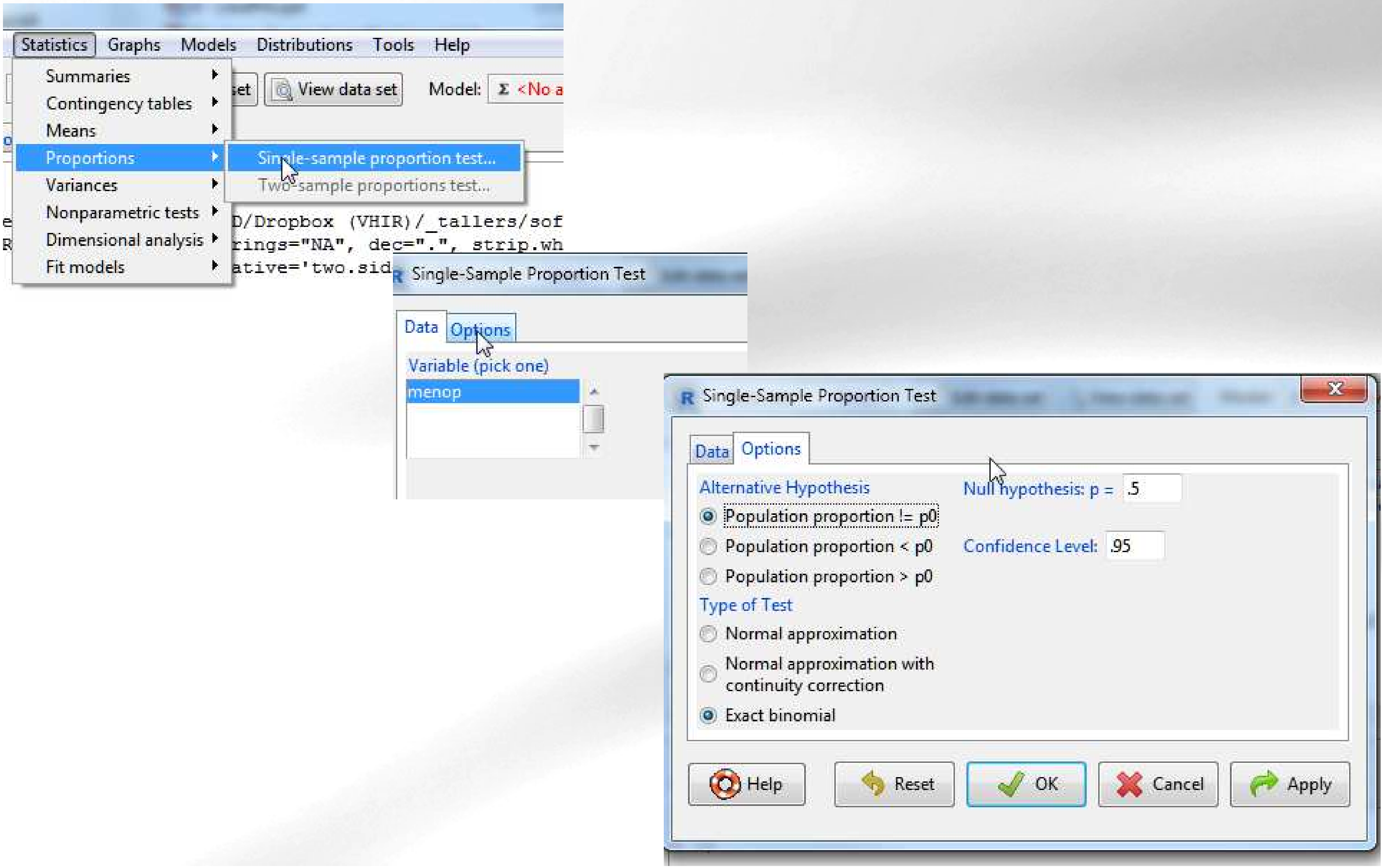




One Sample t-test

data: bua
t = 137.89, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
72.2539 74.3401
sample estimates:
mean of x
73.297

Confidence interval in RCmdr



Vall d'Hebron
Institut de Recerca

Proportion Test Normal Aproximation

UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

Frequency counts (test is for first level):
menop
 NO SI
303 697

1-sample proportions test without continuity correction

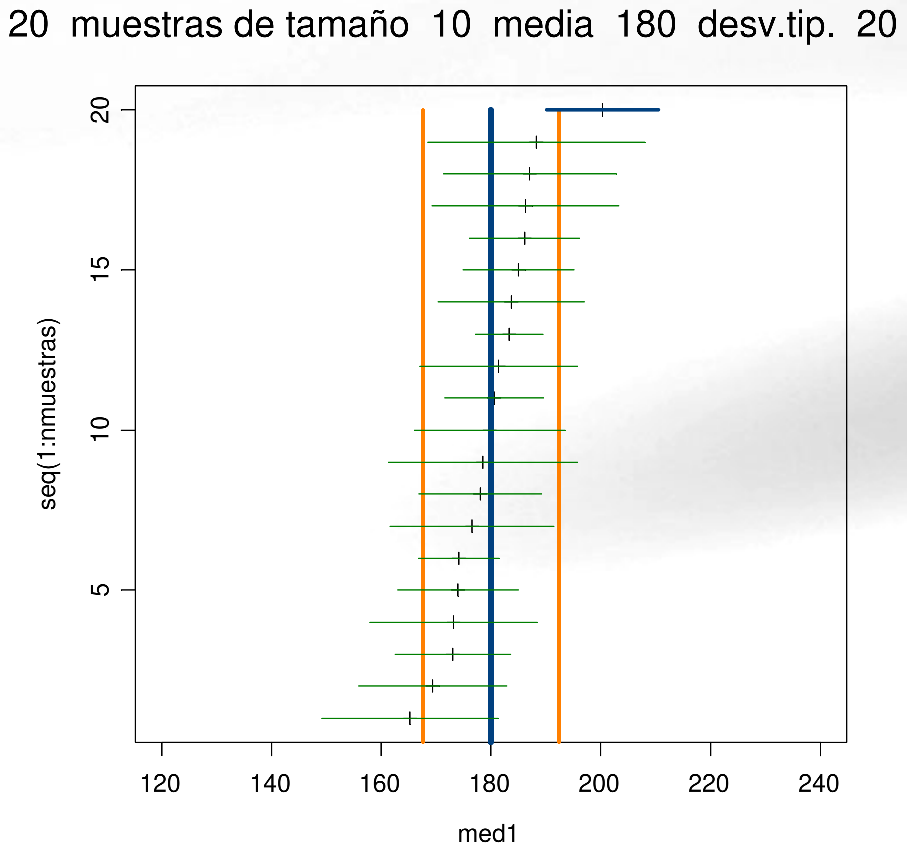
data: rbind(.Table), null probability 0.5
X-squared = 155.24, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2753154 0.3321923
sample estimates:
    p
0.303

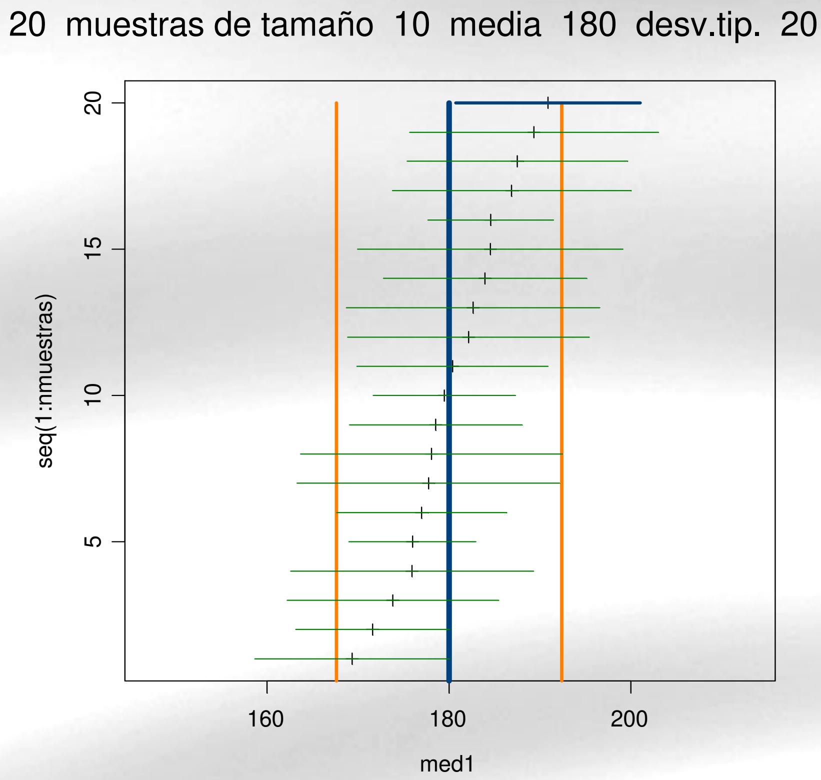
Frequency counts (test is for first level):
menop
 NO SI
303 697

Exact binomial test

data: rbind(.Table)
number of successes = 303, number of trials = 1000, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.274632 0.332533
sample estimates:
probability of success
0.303

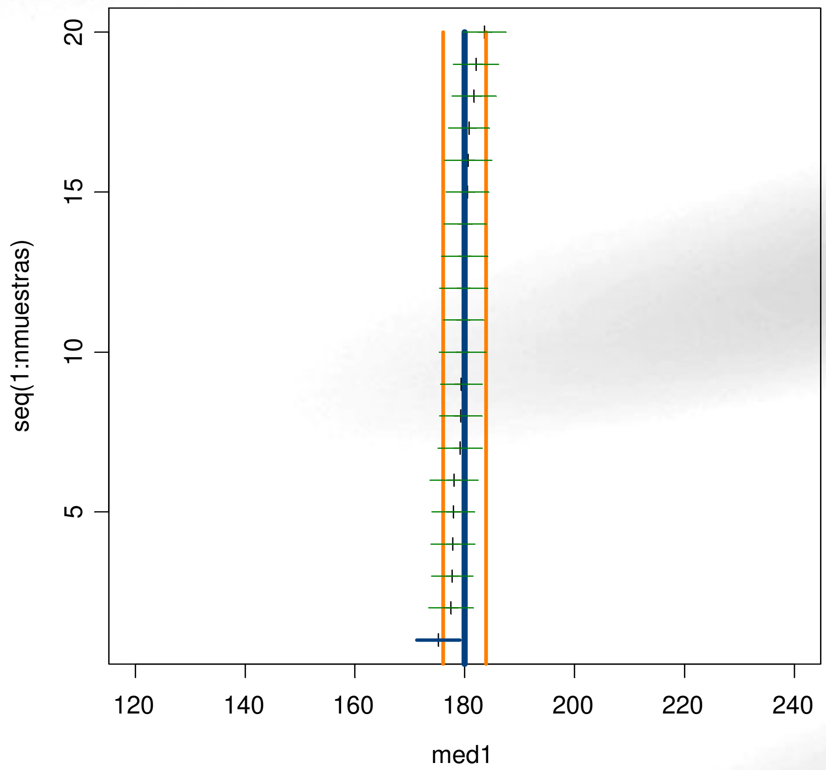
Sample size =10 , Mean=180, sd=20



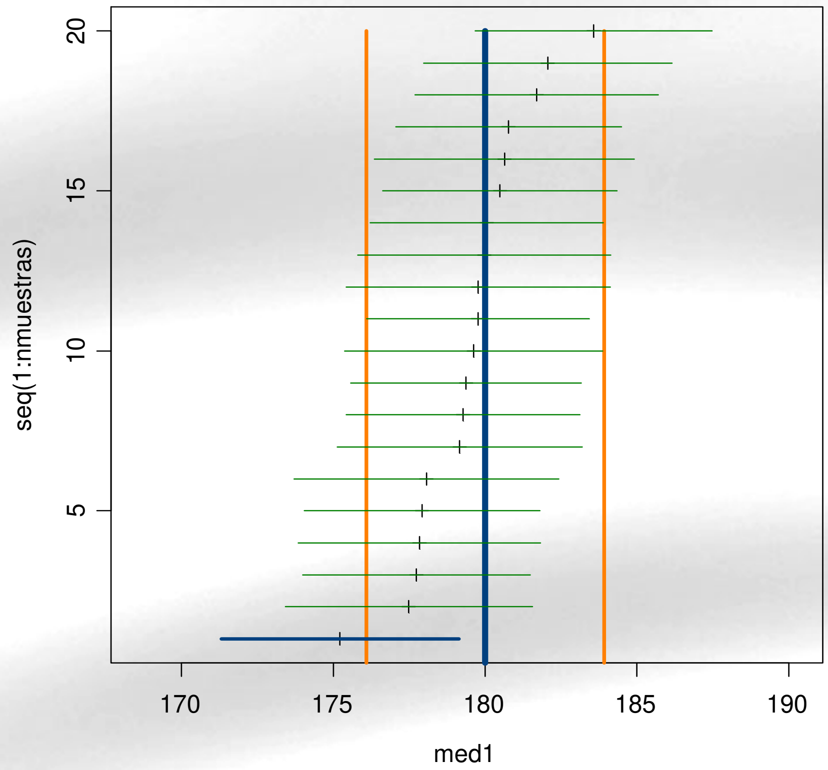


Sample size =100 , Mean=180, sd=20

20 muestras de tamaño 100 media 180 desv.tip. 20



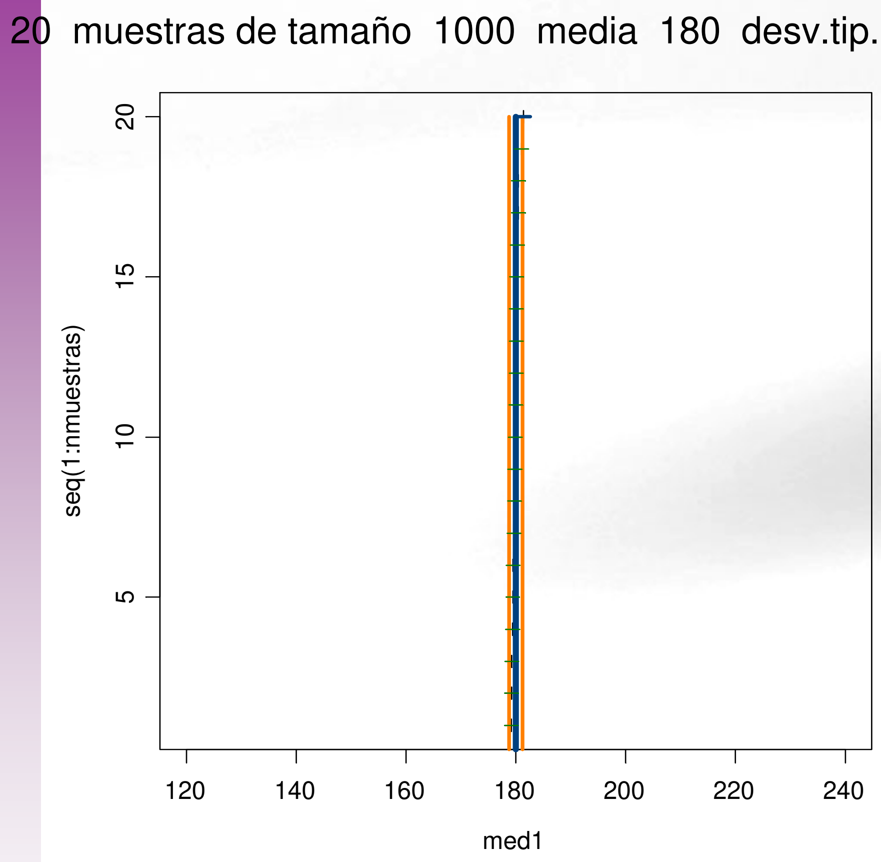
20 muestras de tamaño 100 media 180 desv.tip. 20

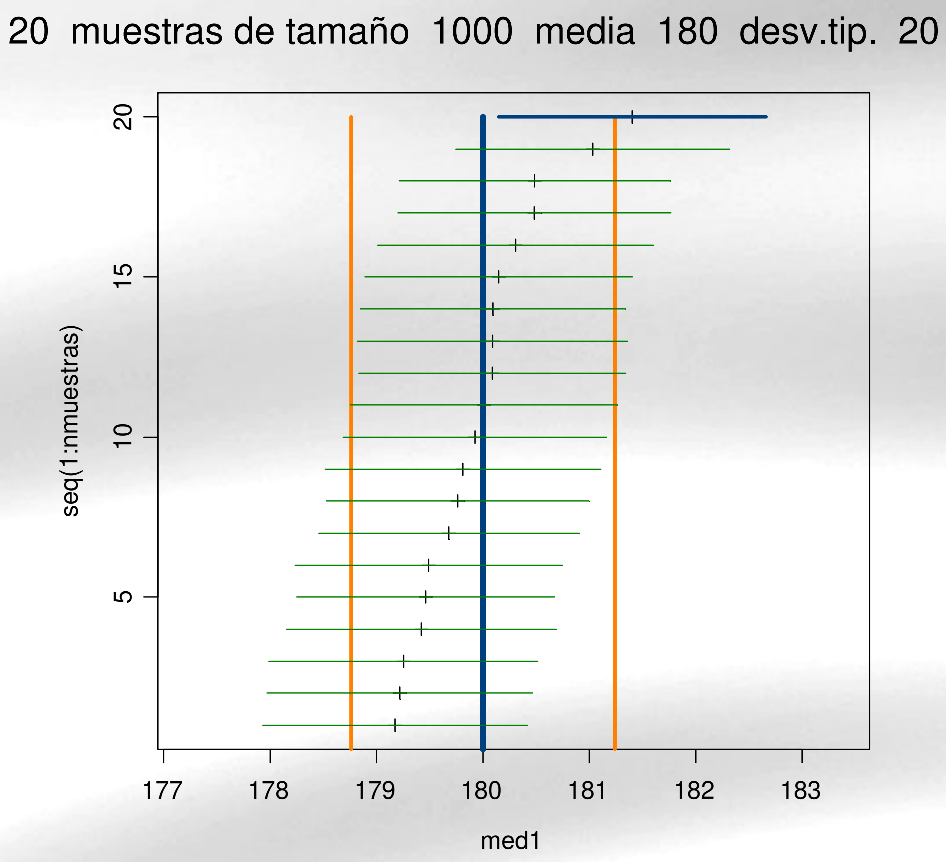


Vall d'Hebron Institut de Recerca

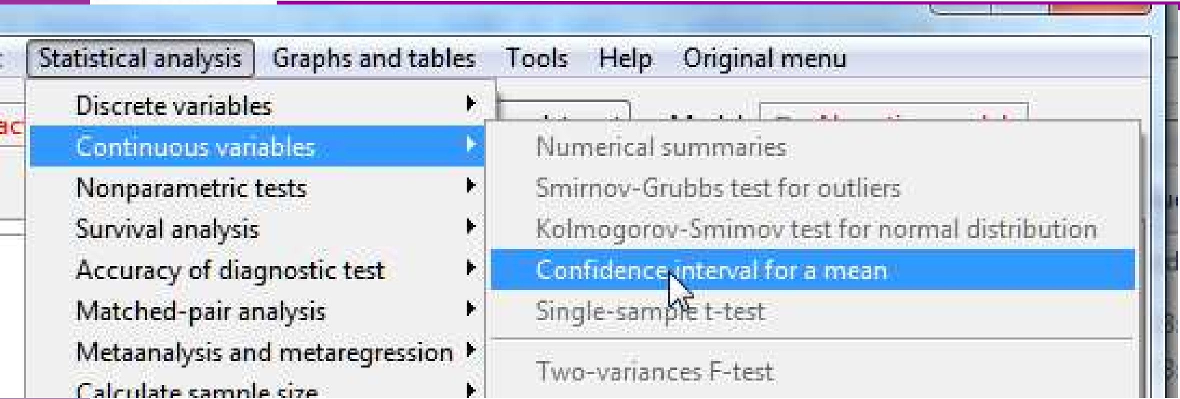
UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

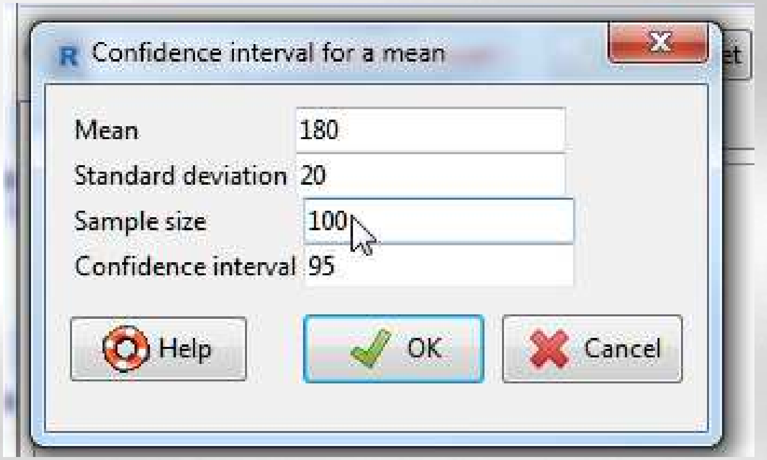
Sample size =100 , Mean=180, sd=20



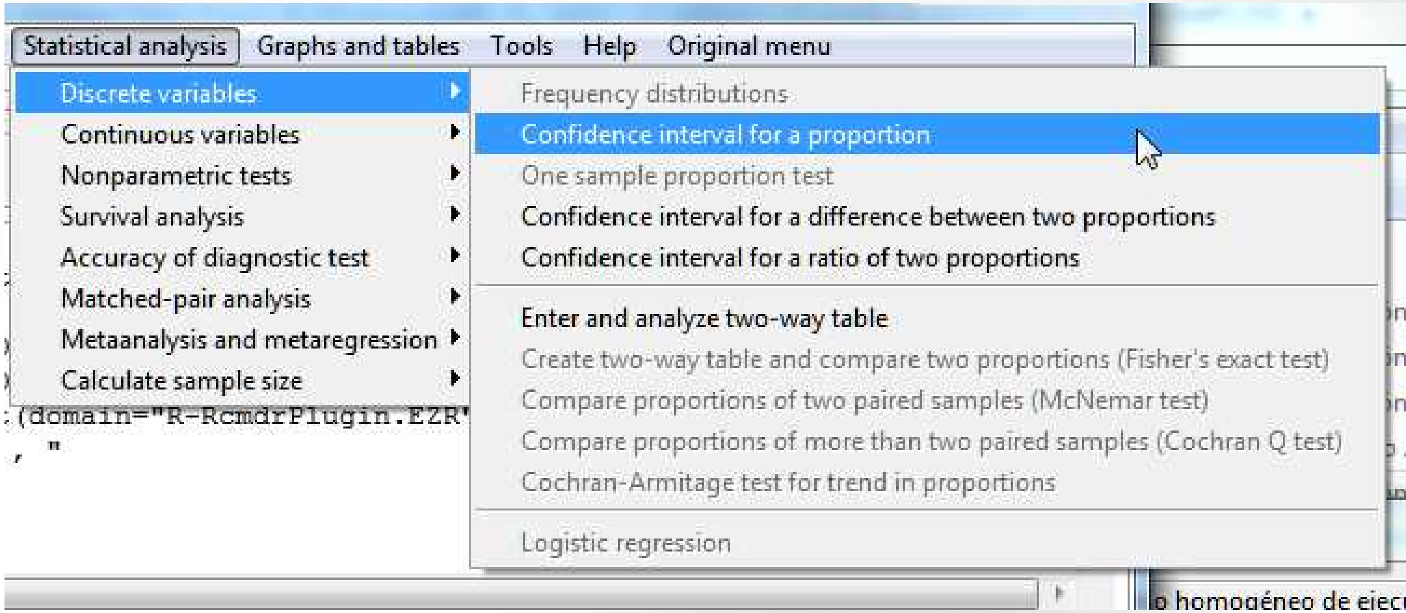


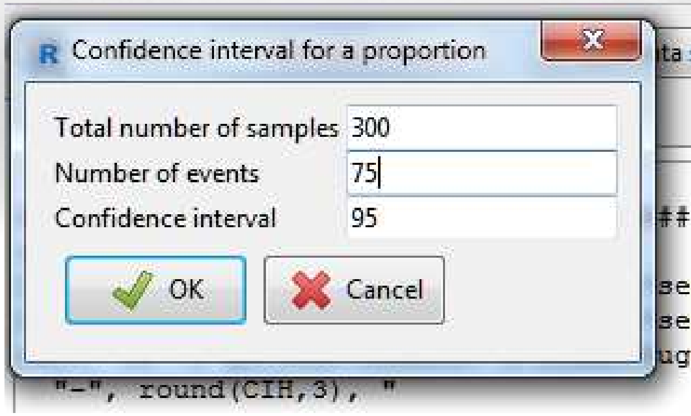
Confidence interval calculator (Plugin EzR)





95 %CI 176.032-183.968





[1] Probability : 0.25
[1] 95% confidence interval : 0.202 - 0.303