Principles of Statistical Inference

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR

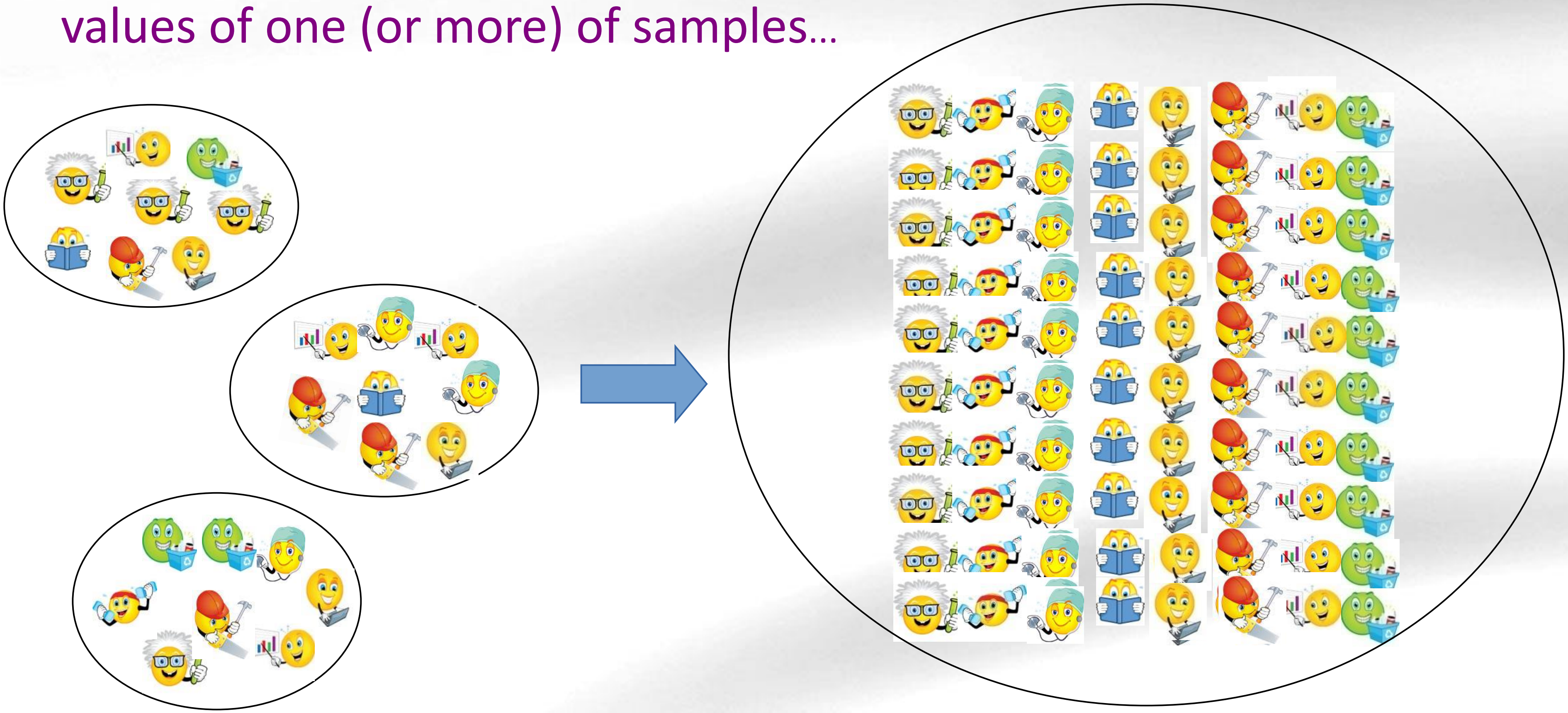
Santiago Pérez-Hoyos i Alex Sánchez-Pla

santi.perezhoyos@vhir.org

alex.sanchez@vhir.org

The objective of statistical inference

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

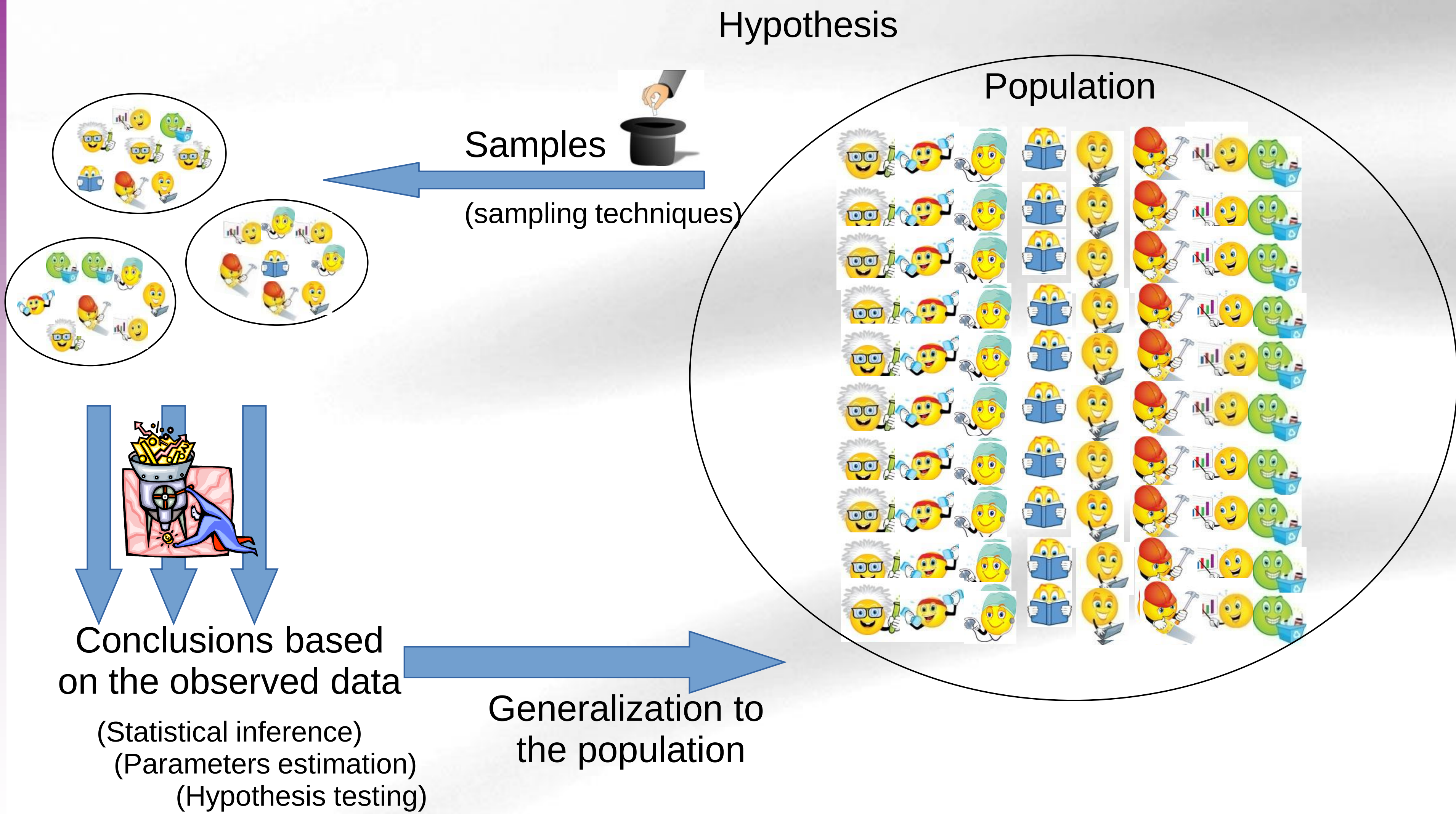


… Determine (“***infer***”) the properties of the entire population.

Vall d'Hebron
Institut de Recerca

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

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

Vall d'Hebron
Institut de Recerca

Point estimation (1)

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

- 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*

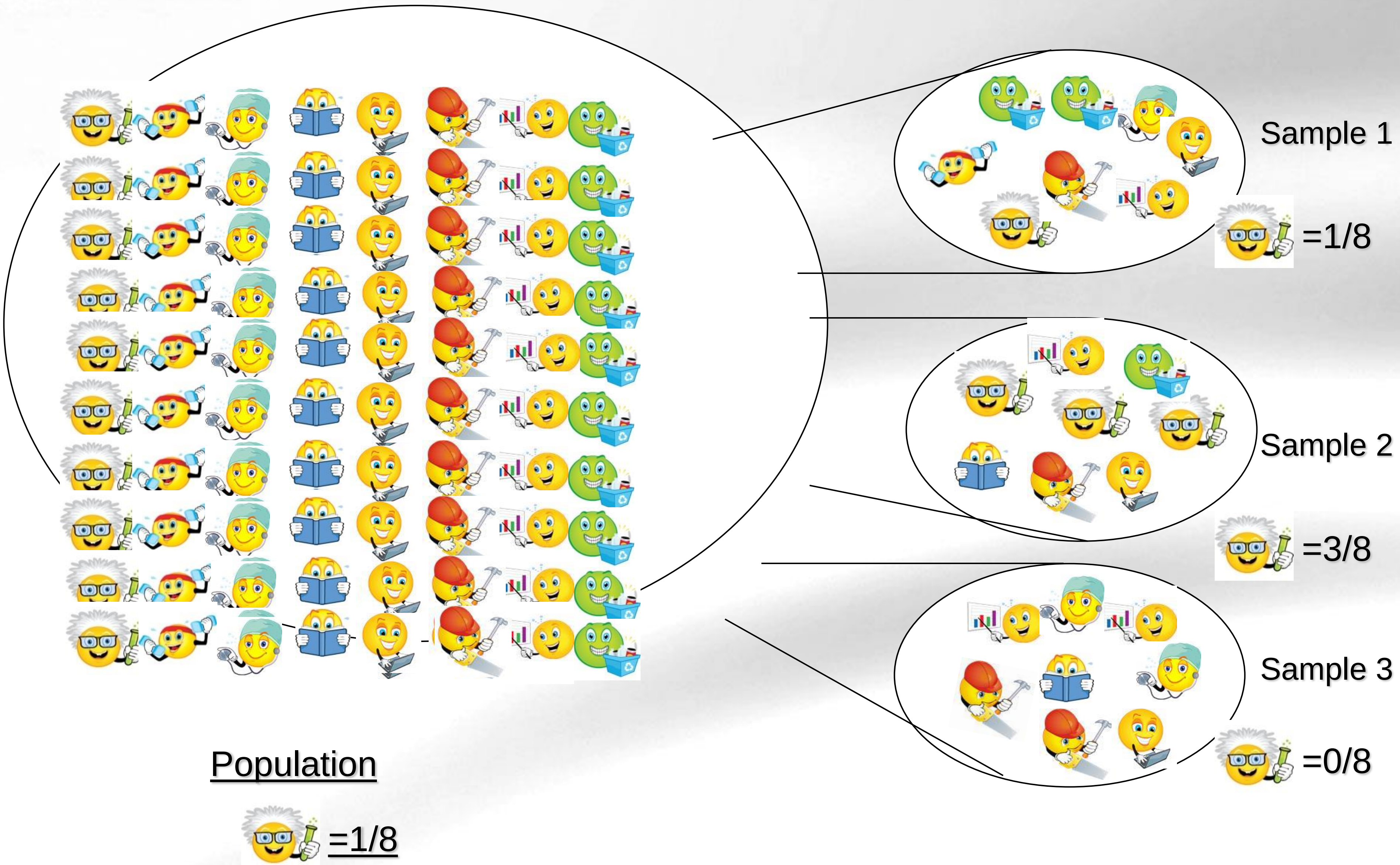
Point estimation (II)

- 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

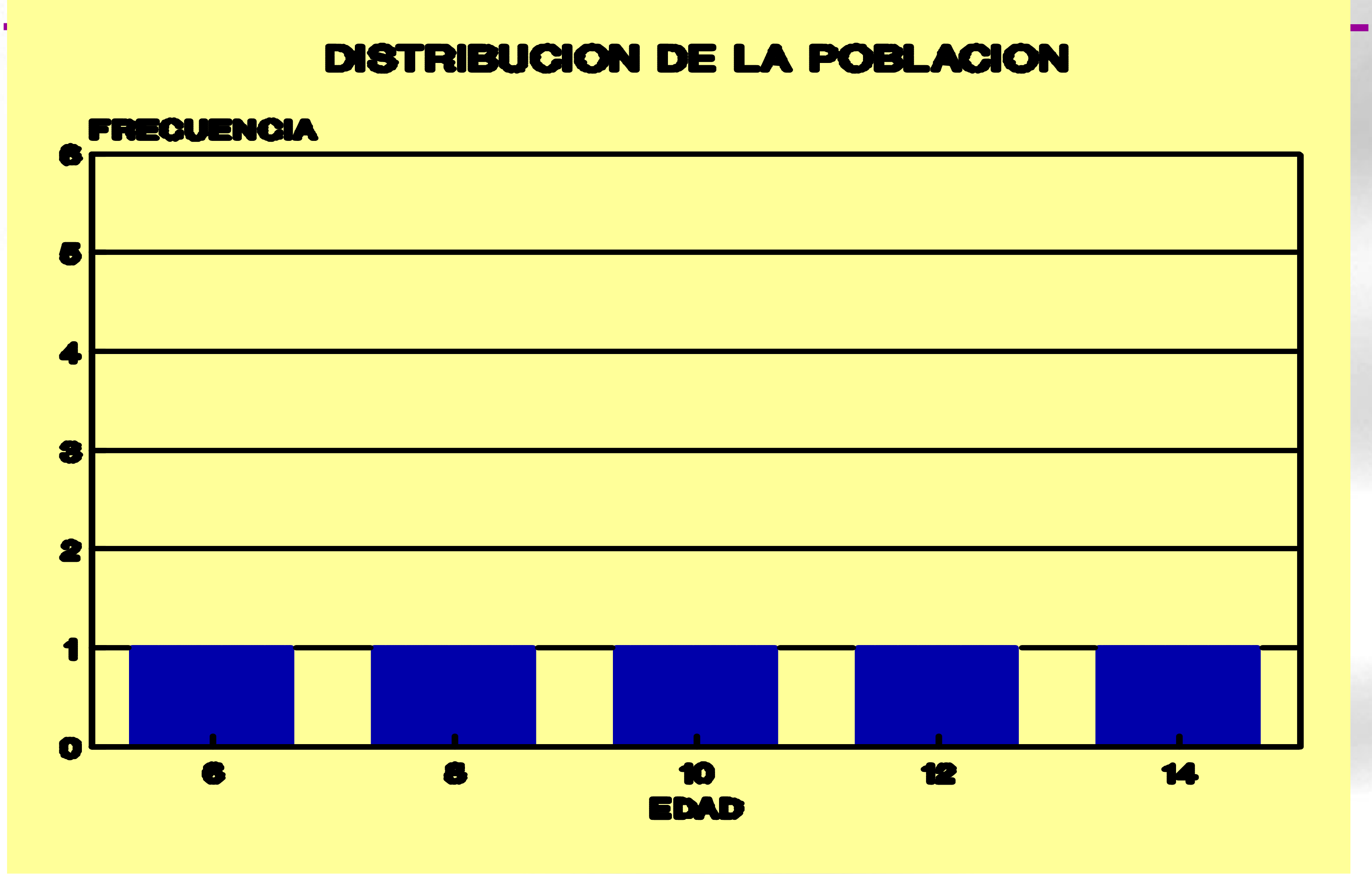
In this problem we can compute the population parameters because we know all the population values!!!

25 Samples n=2

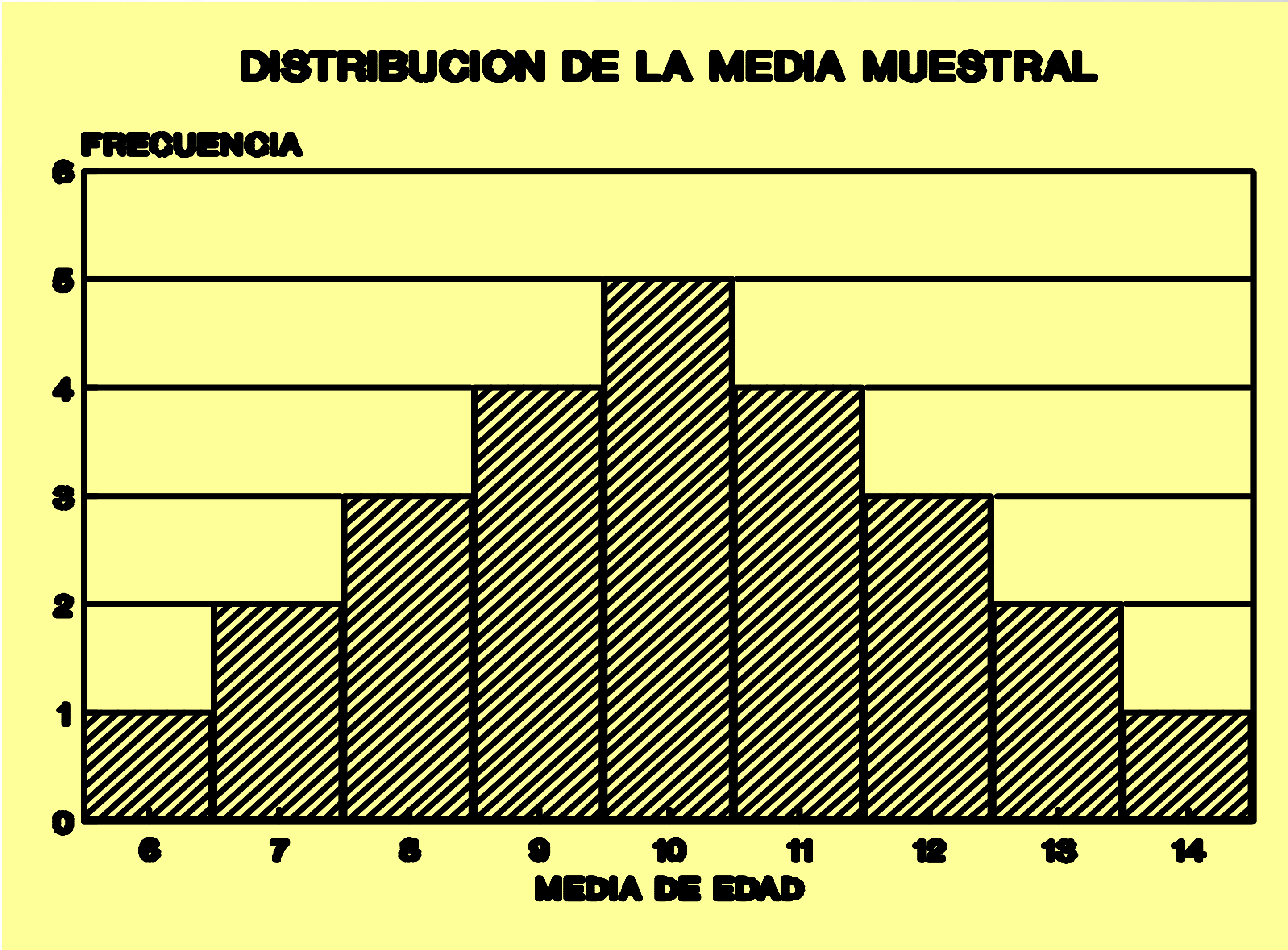
		Second Data				
		6	8	10	12	14
Fist Data	6	6,6 (6)	6,8 (7)	6,10 (8)	6,12 (9)	6,14 (10)
	8	8,6 (7)	8,8 (8)	8,10 (9)	8,12 (10)	8,14 (11)
	10	10,6 (8)	10,8 (9)	10,10 (10)	10,12 (11)	10,14 (12)
	12	12,6 (9)	12,8 (10)	12,10 (11)	12,12 (12)	12,14 (13)
	14	14,6 (10)	14,8 (11)	14,10 (12)	14,12 (13)	14,14 (14)

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


DISTRIBUCION DE LA POBLACION
FRECUENCIA
6
5
4
3
2
1
0
6
8
10
12
14
EDAD

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

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

- Variance of 25 sample means

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

- 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

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

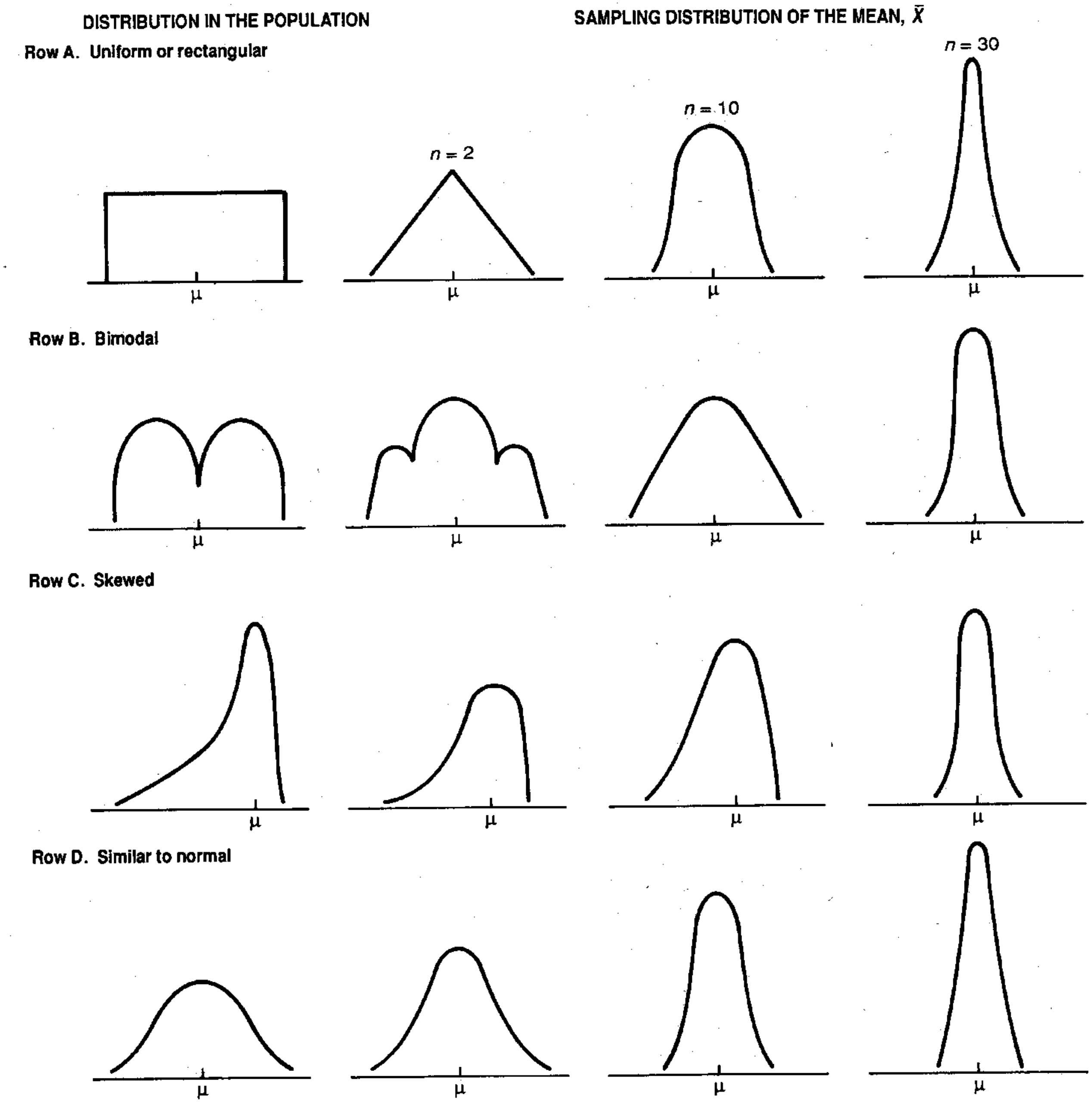


Figure 6–3. Illustration of ramifications of central limit theorem.

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.

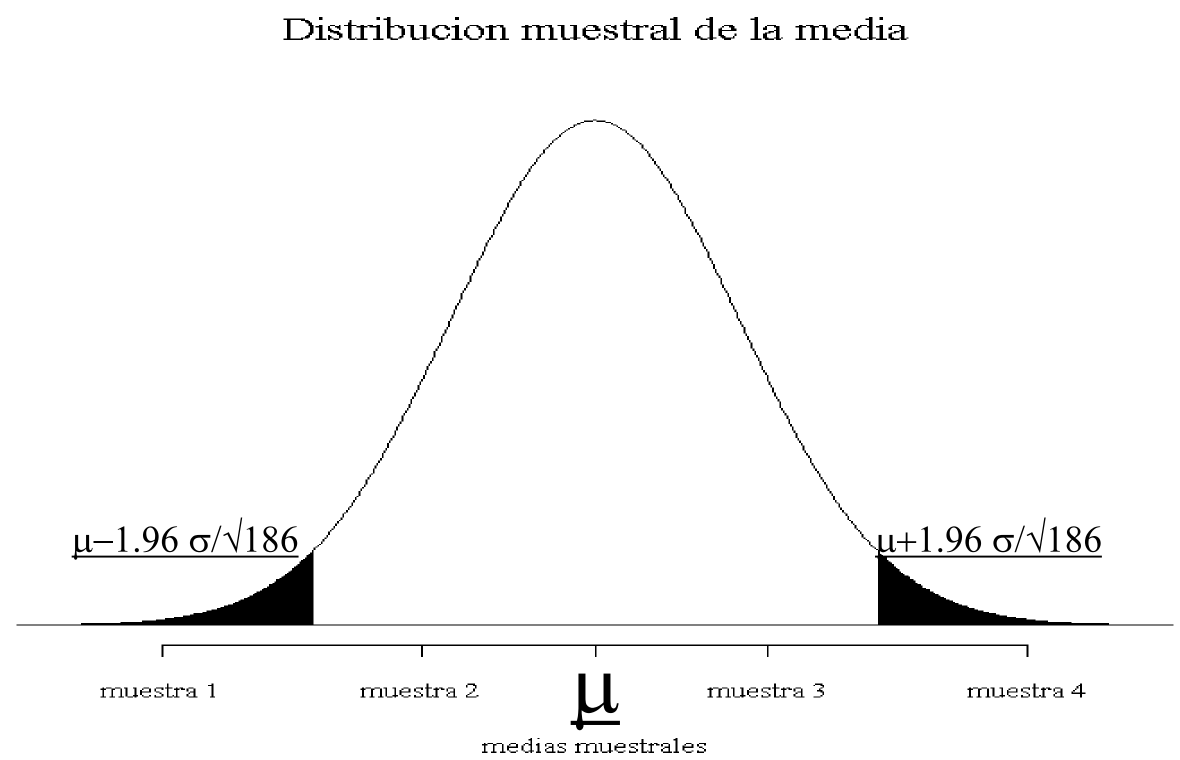
Confidence interval of Mean

- 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:

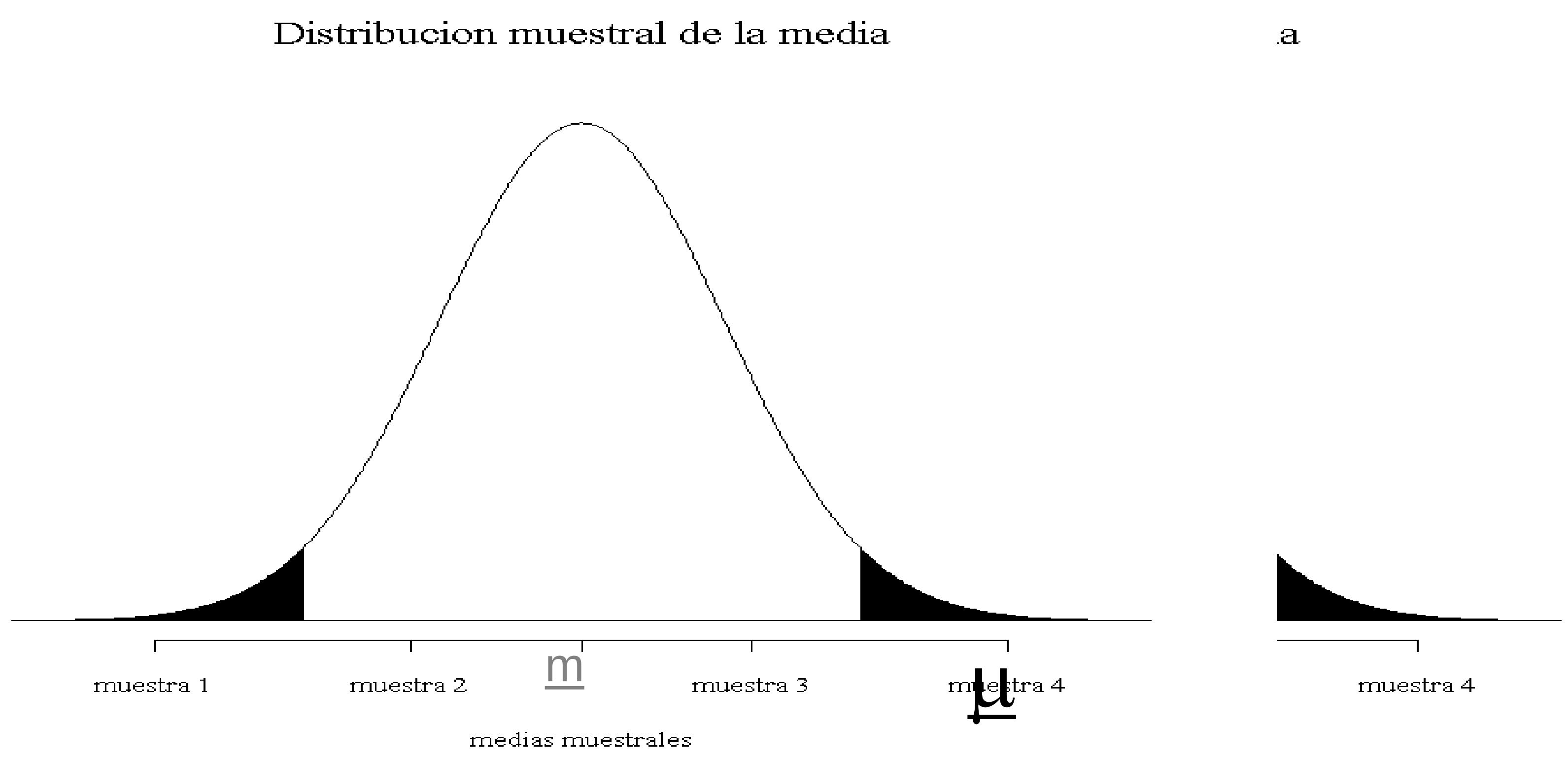
$$\overline{x} \pm z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} = 118 \pm 1,96 \times 12/\sqrt{186}$$

  – 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

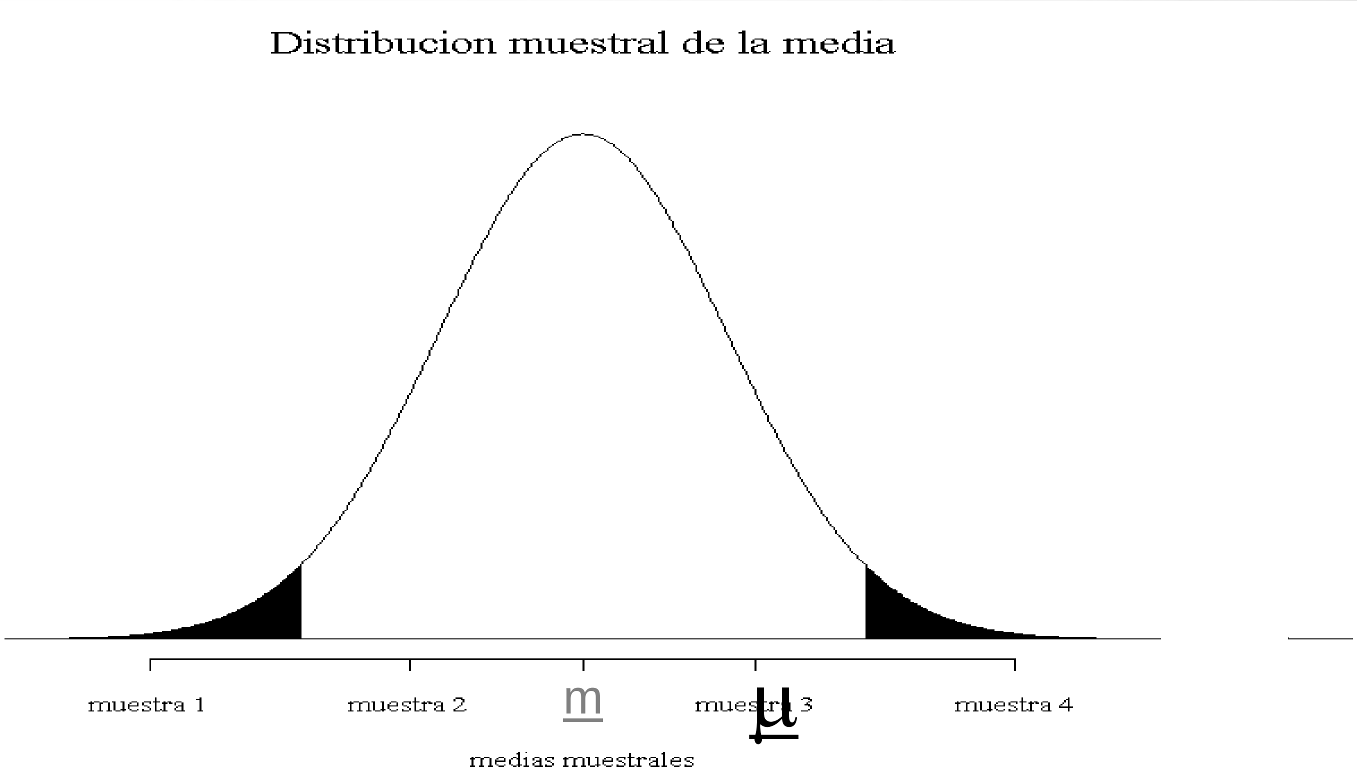

Distribucion muestral de la media
μ−1.96 σ/√186
μ+1.96 σ/√186
muestra 1
muestra 2
μ
muestra 3
muestra 4
medias muestrales

m±1.96 σ/√186



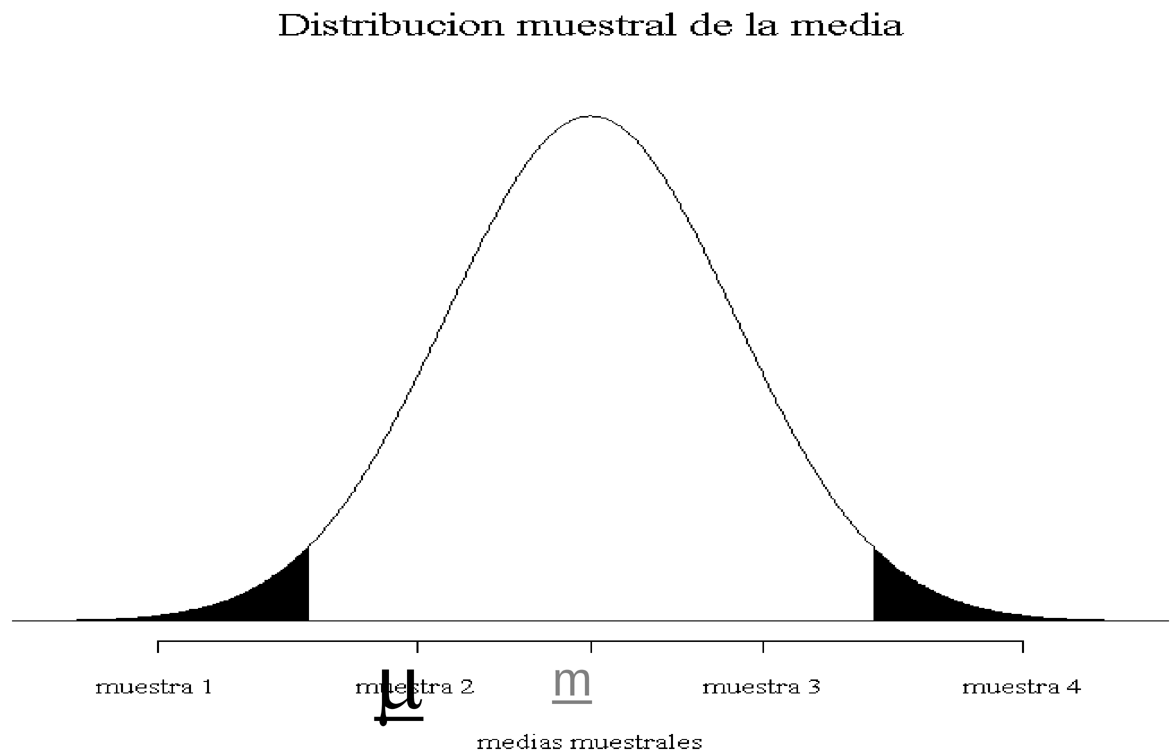
Population mean is outside de confidence interval

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



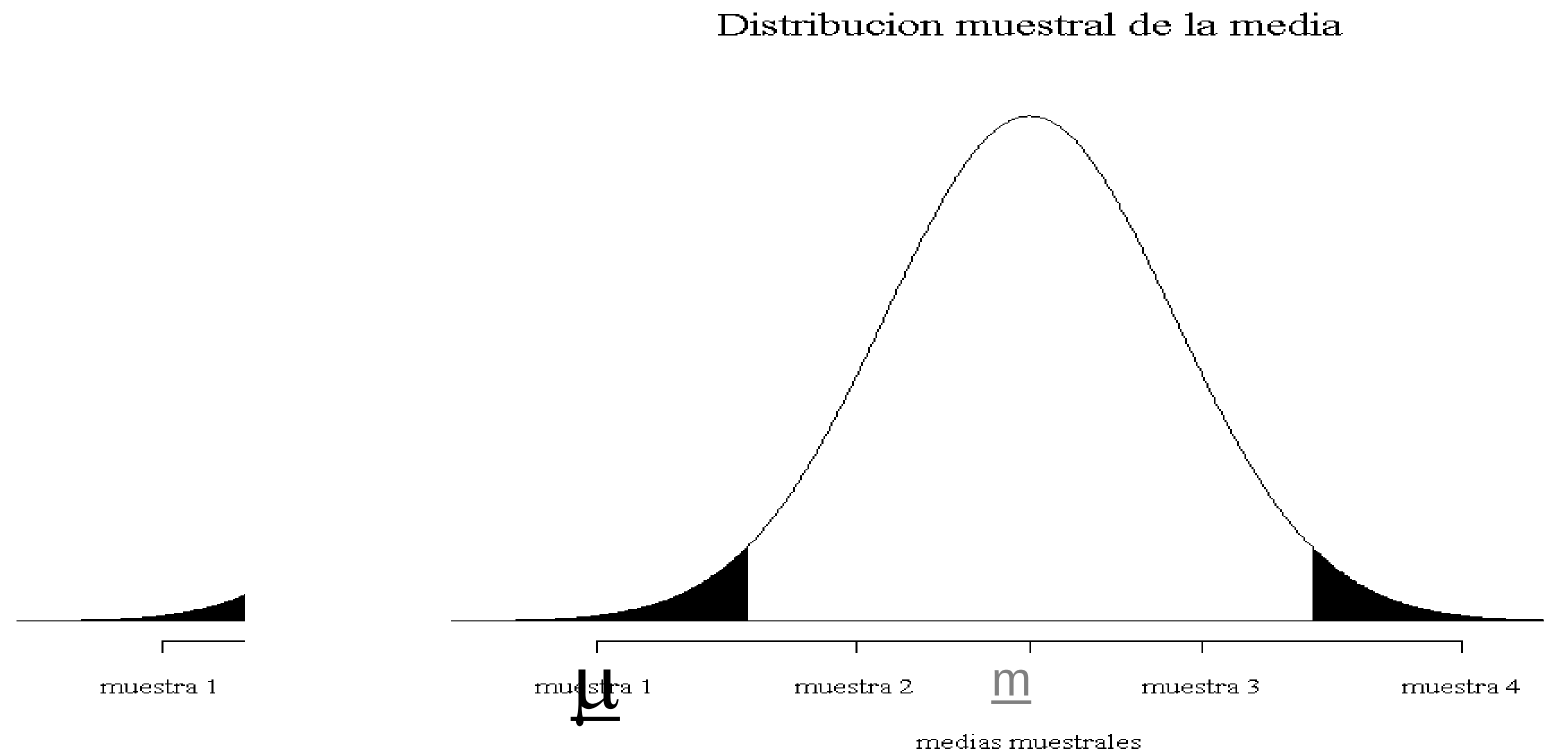
Population mean is inside confidence interval

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



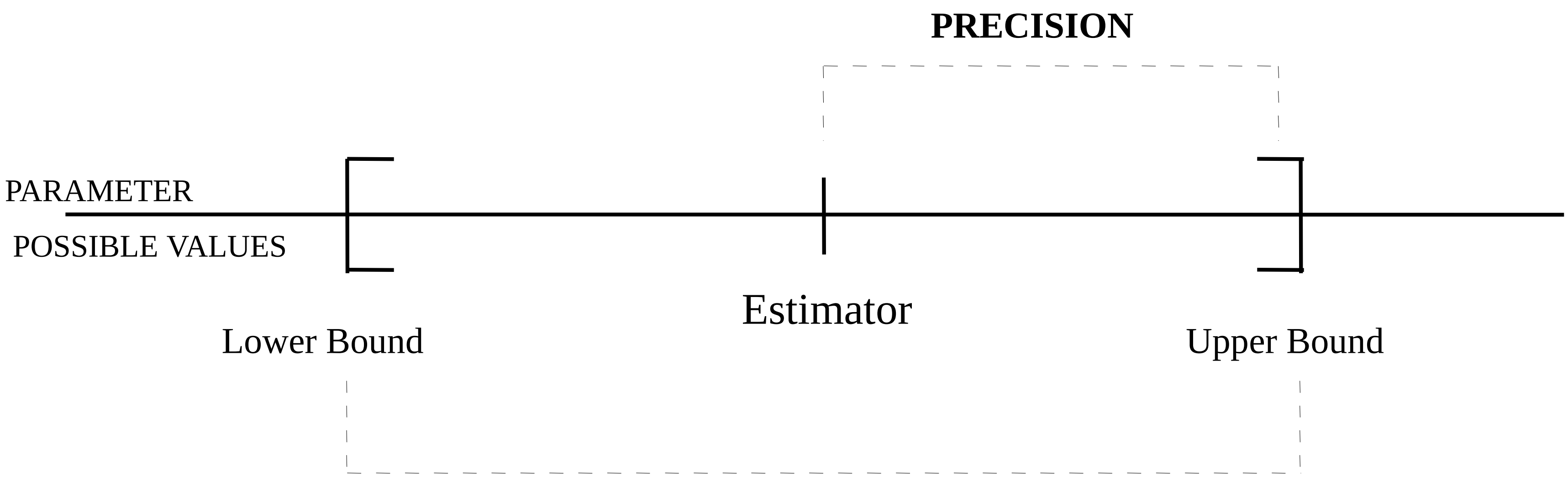
Population mean is inside confidence interval

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



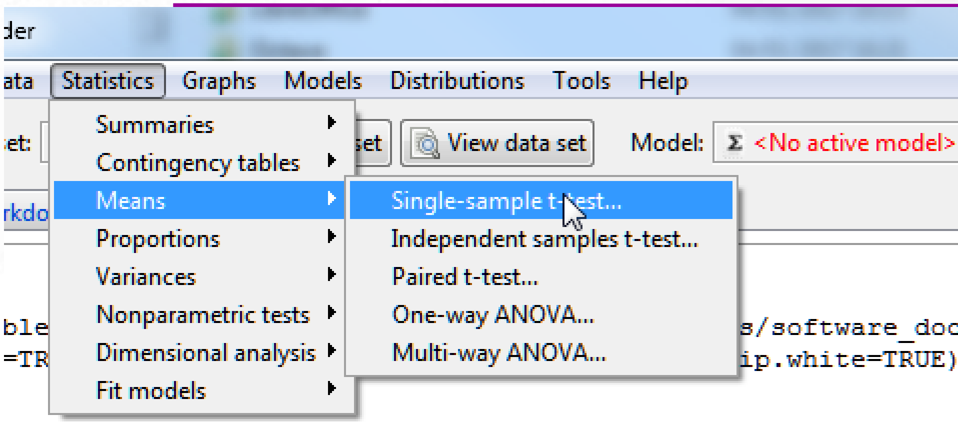
Population mean its outside confidence interval

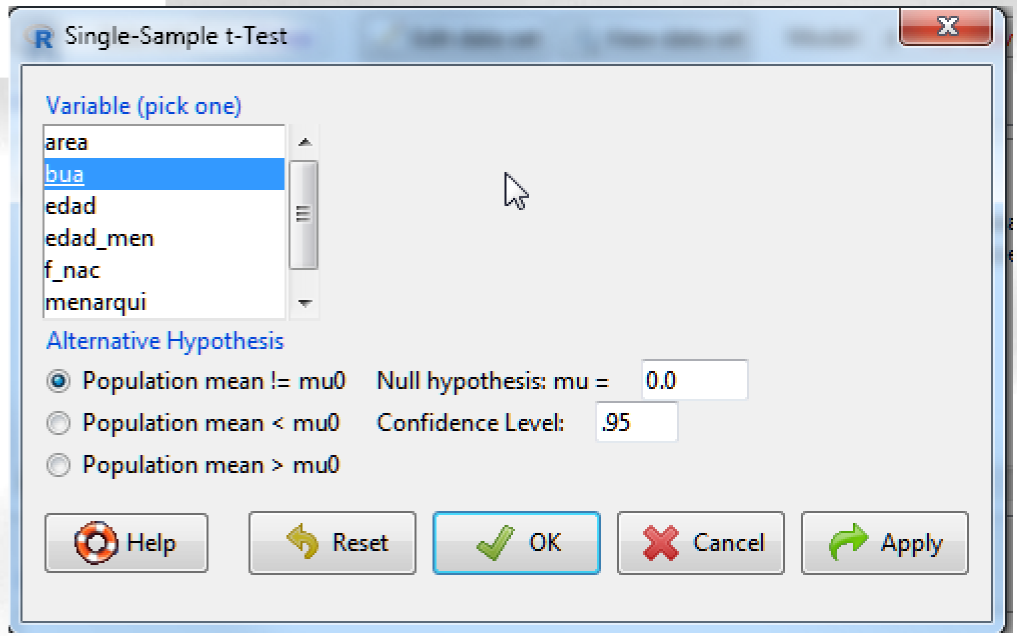
Confidence interval



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:

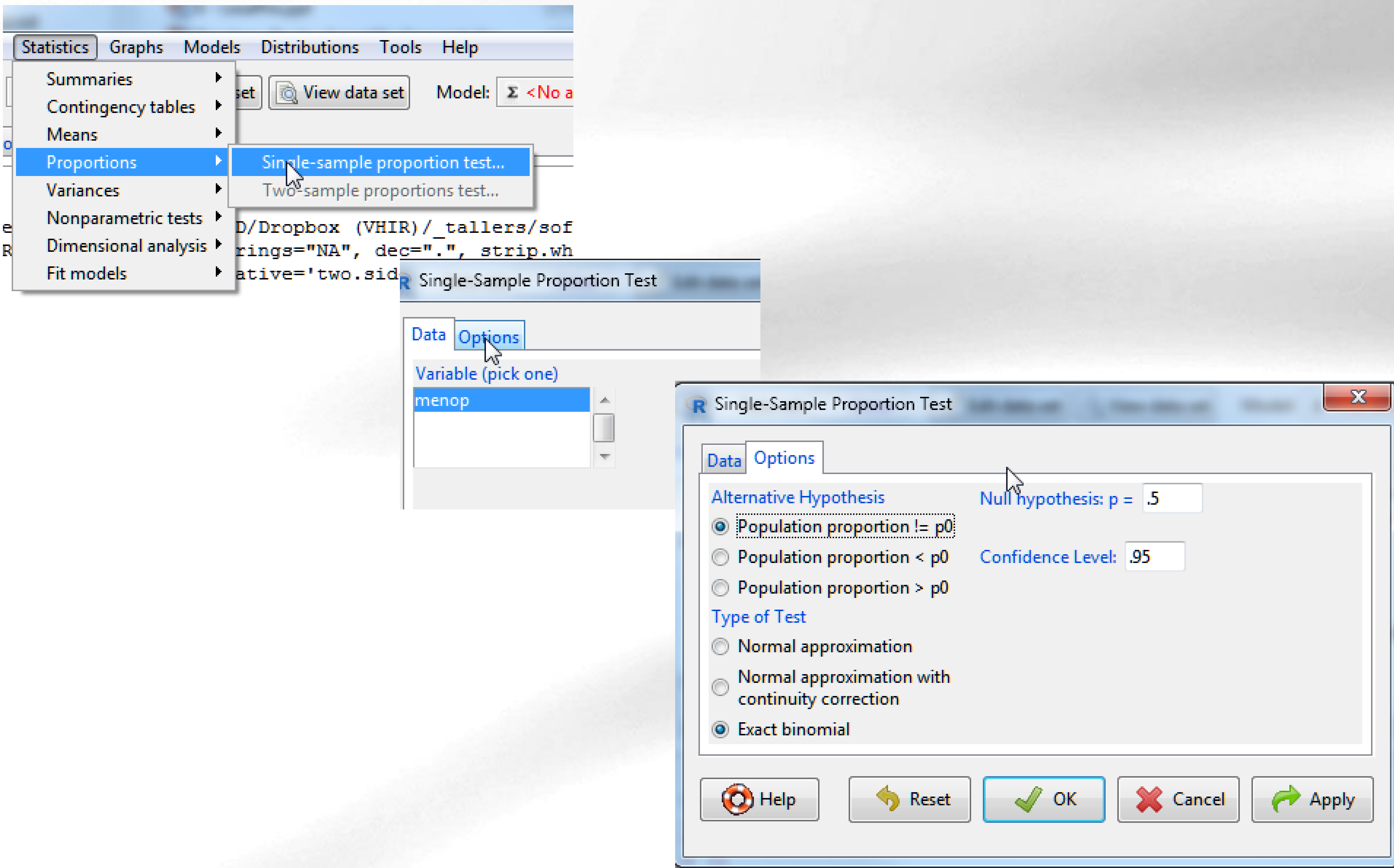
72.2539 74.3401

sample estimates:

mean of x

73.297

Confidence interval in RCmdr



Proportion Test Normal Aproximation

Frequency counts (test is for first level):
menop
 NO SI
303 697
1-sample proportions test without continuity correction

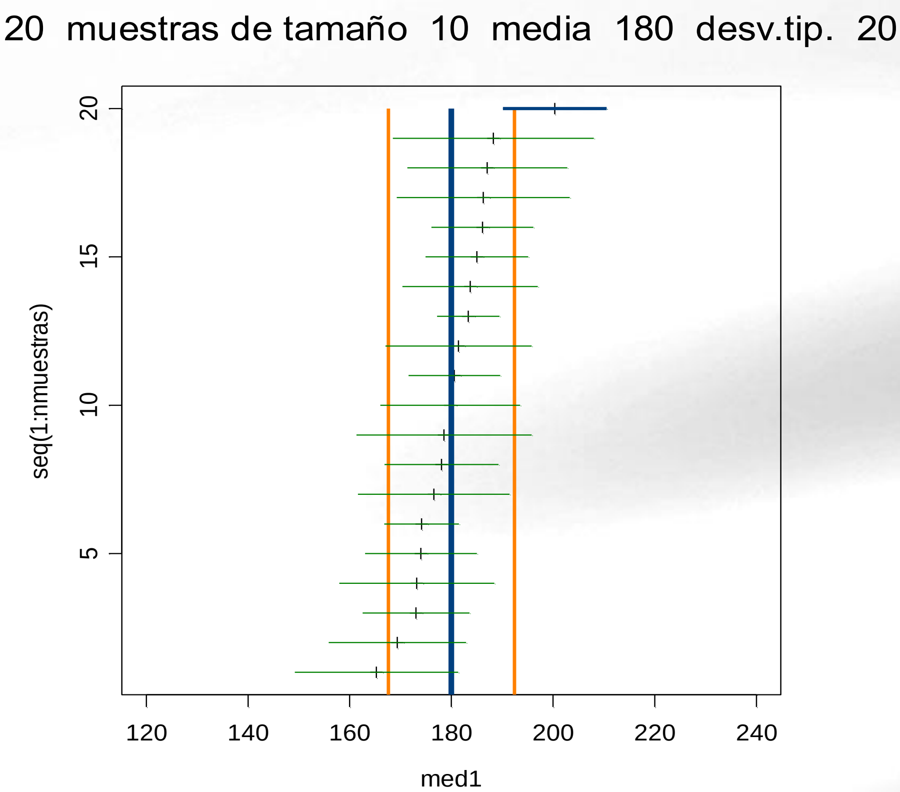
data: rbind(.Table), null probability 0.5
X-squared = 155.24, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2753154 0.3321923
sample estimates:
p
0.303

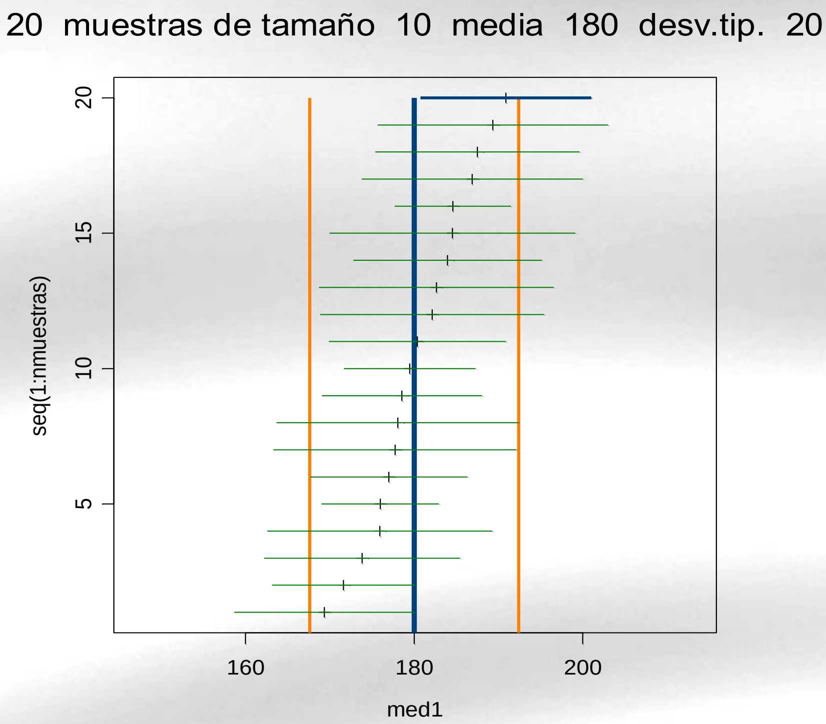
Frequency counts (test is for first level):
menop
NO SI
303 697

Exact binomial test

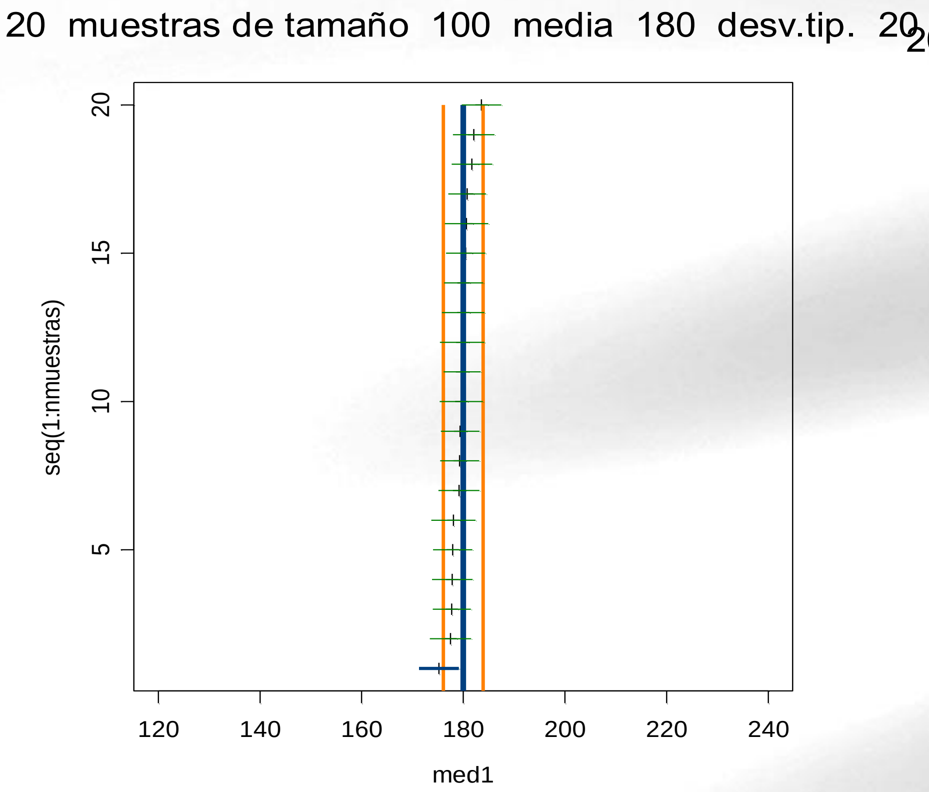
data: rbind(.Table)
number of successes = 303, number of trials = 1000, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.274632 0.332533
sample estimates:
probability of success
0.303

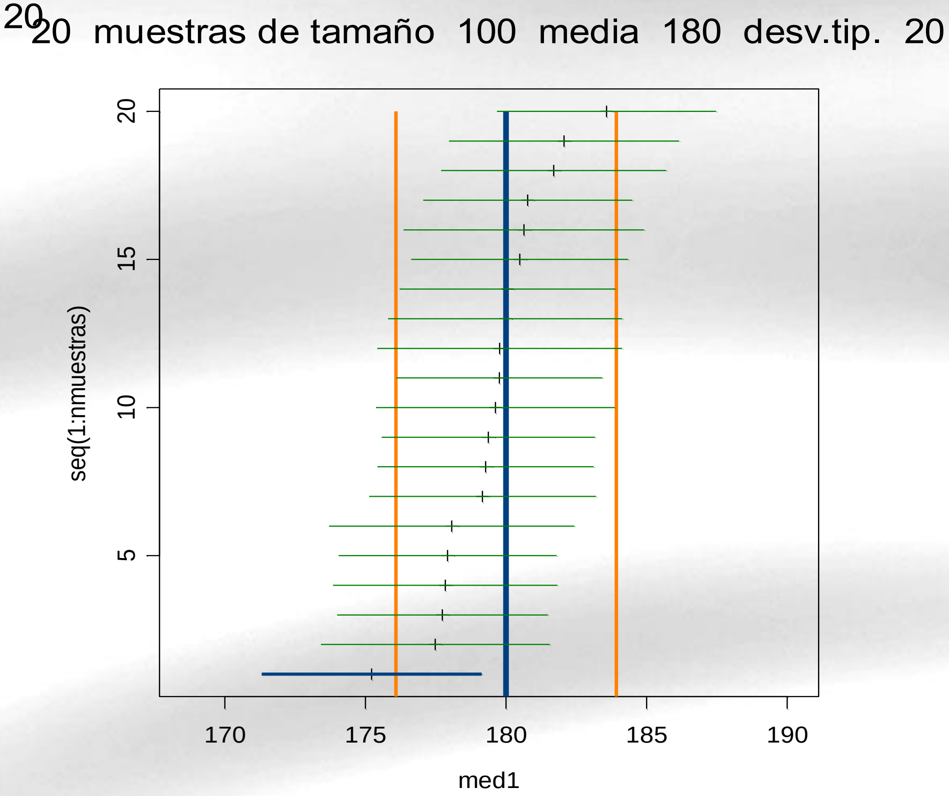
Sample size =10 , Mean=180, sd=20





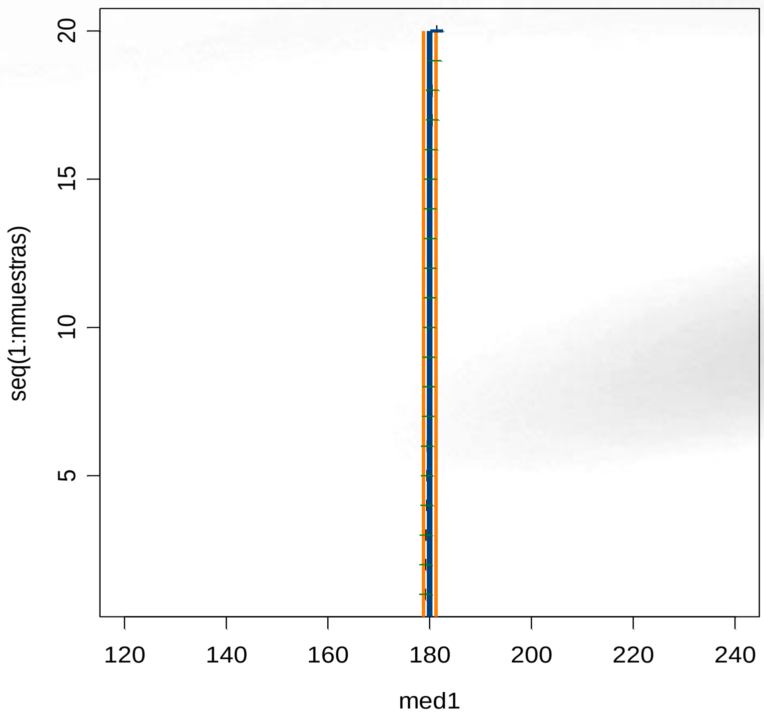
Sample size =100 , Mean=180, sd=20



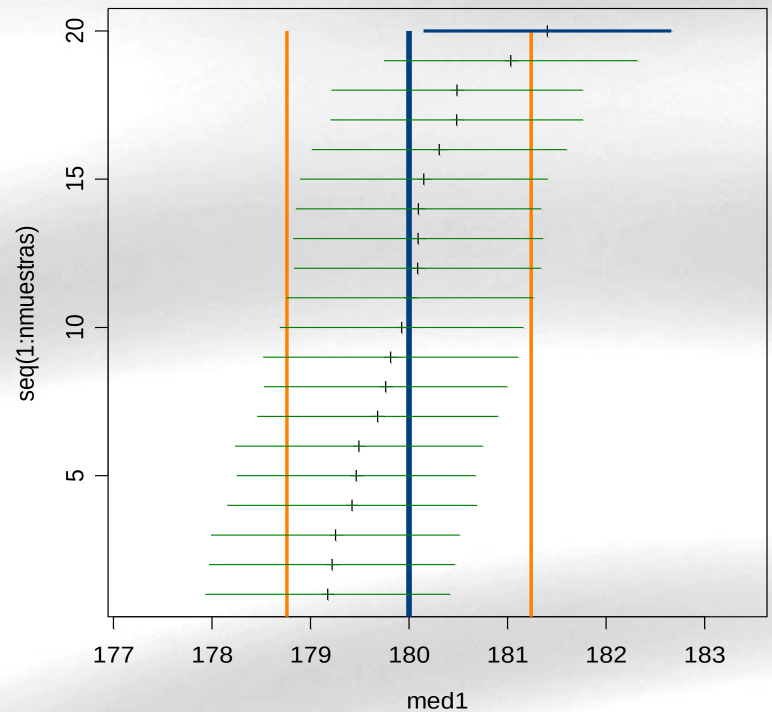


Sample size =100 , Mean=180, sd=20

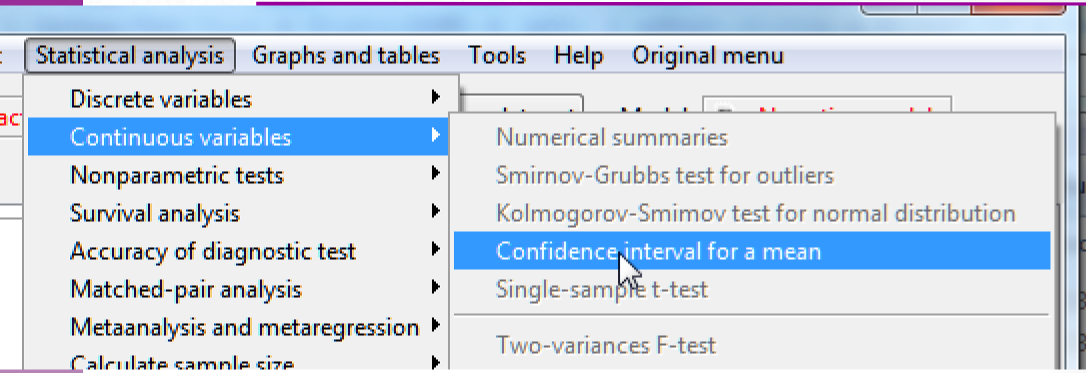
20 muestras de tamaño 1000 media 180 desv.tip. 20

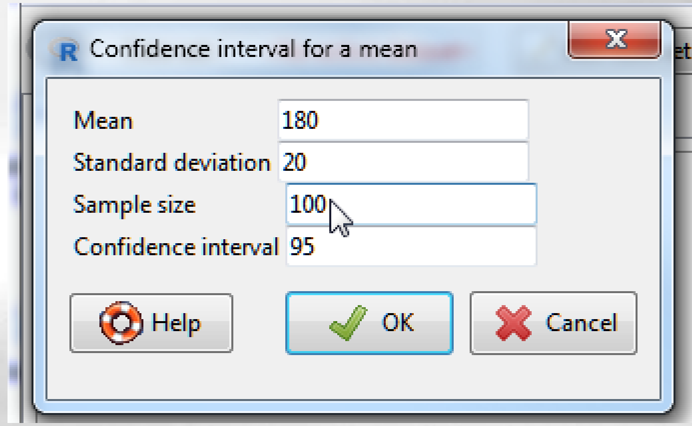


20 muestras de tamaño 1000 media 180 desv.tip. 20

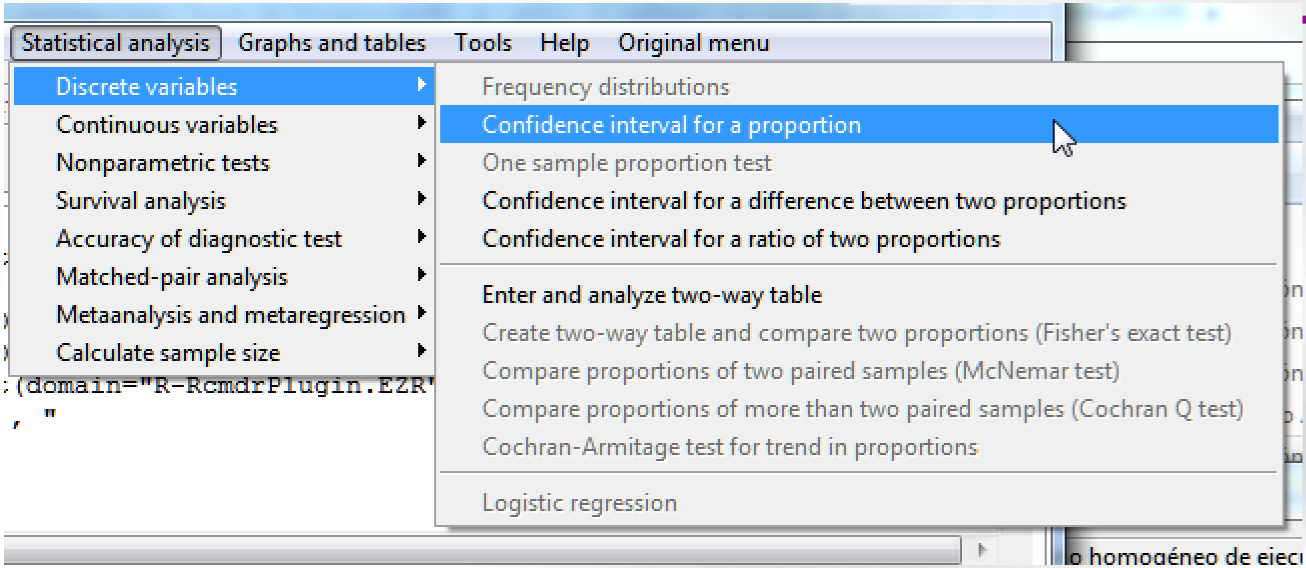


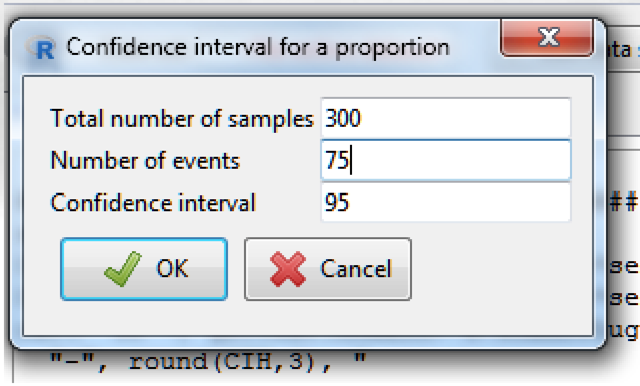
Confidence interval calculator (Plugin EzR)





95 %CI 176.032-183.968





[1] Probability : 0.25

[1] 95% confidence interval : 0.202 - 0.303

Sample Size

Sample Size Calculation

- Some questions must be answered before we can compute "sample size"
  - Precision (interval range) of estimations
  - Level of confidence of estimations
  - *Sometimes (for proportions) it will help to have an idea of the value of the parameter we want to estimate*

Sample Size Calculation

- The question "what is the sample size" must be rephrased as:

  What sample size is needed
  to estimate the mean, so that
  we have a high confidence (say 95%)
  that the estimation error will be less than a given threshold?

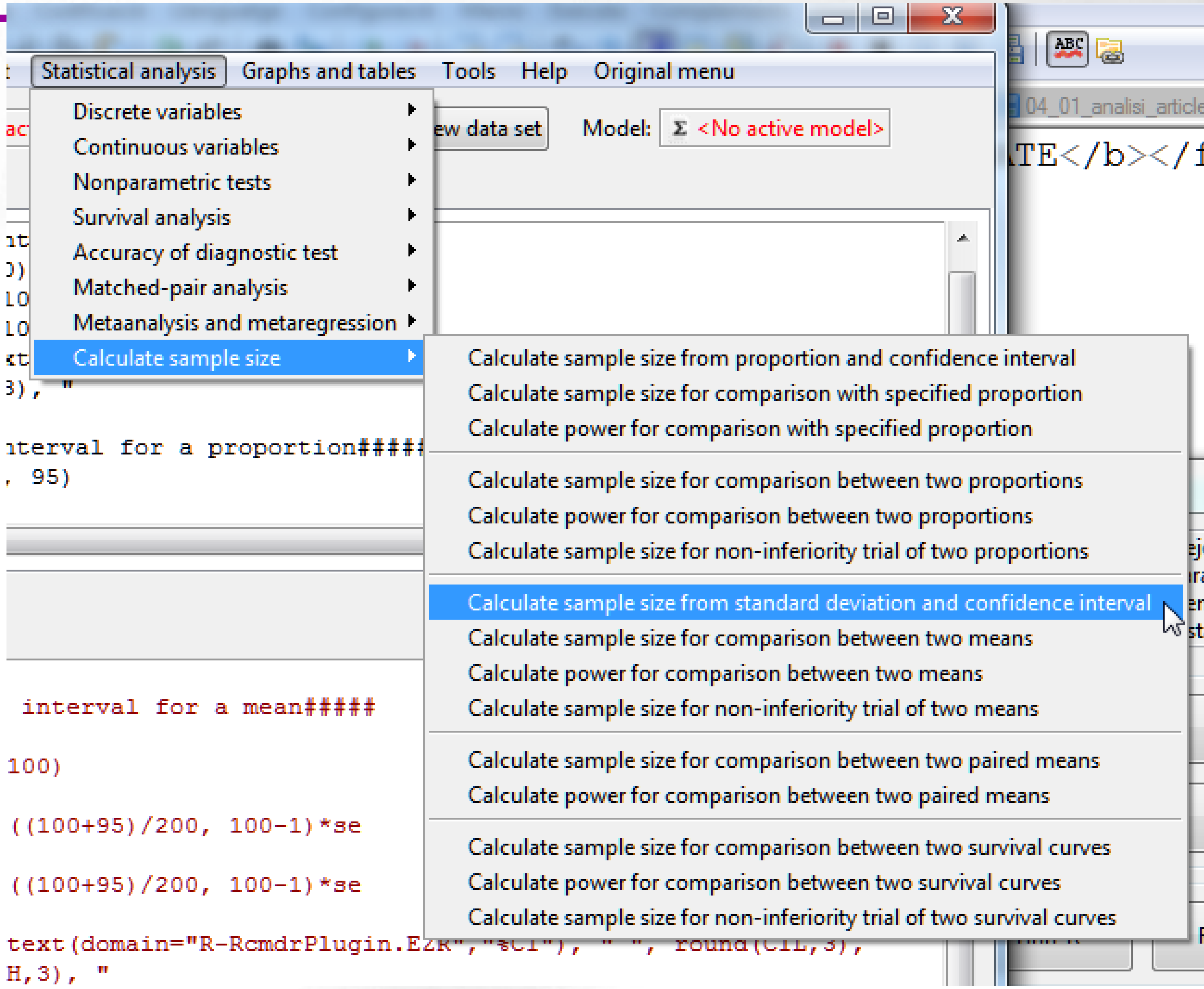
Sample SIZE for mean

$$Precision = z_{1-\alpha/2} * ee = 1.96 * \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{precision^2}$$

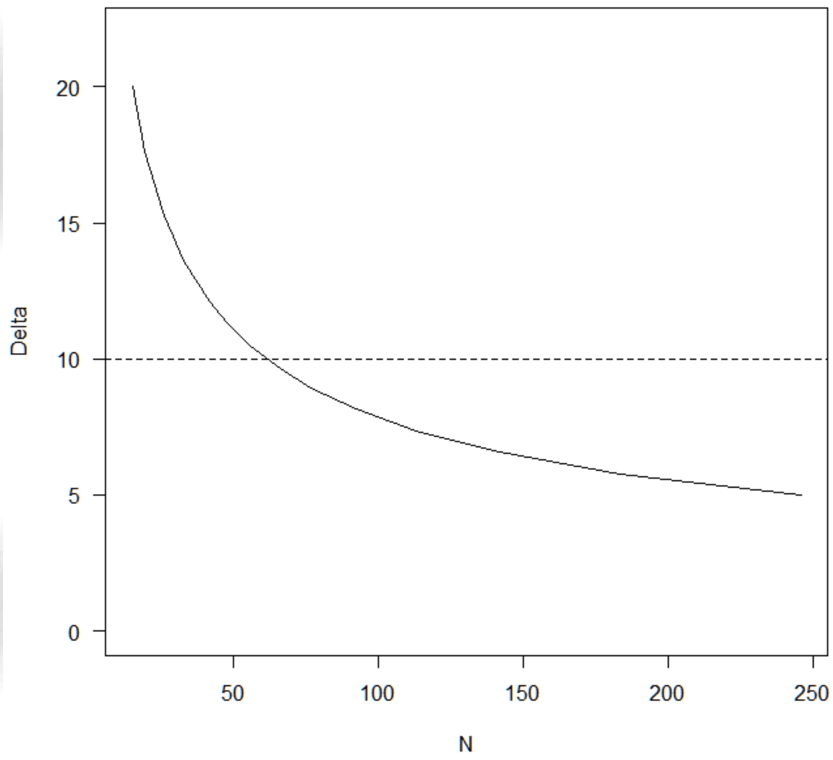
If interval range is 10 (precision = 10/2= 5), confidence levell is 95% and standard deviation is 20, sample size will be

$n = 1.96^2\ 20^2 / 5^2 = 62$


Statistical analysis
Graphs and tables
Tools
Help
Original menu
Model:
<No active model>
Discrete variables
Continuous variables
Nonparametric tests
Survival analysis
Accuracy of diagnostic test
Matched-pair analysis
Metaanalysis and metaregression
Calculate sample size
Calculate sample size from proportion and confidence interval
Calculate sample size for comparison with specified proportion
Calculate power for comparison with specified proportion
Calculate sample size for comparison between two proportions
Calculate power for comparison between two proportions
Calculate sample size for non-inferiority trial of two proportions
Calculate sample size from standard deviation and confidence interval
Calculate sample size for comparison between two means
Calculate power for comparison between two means
Calculate sample size for non-inferiority trial of two means
Calculate sample size for comparison between two paired means
Calculate power for comparison between two paired means
Calculate sample size for comparison between two survival curves
Calculate power for comparison between two survival curves
Calculate sample size for non-inferiority trial of two survival curves

```
> SampleMeanCI(20, 10, 95)
                         Assumptions
Standard deviation                20
Confidence interval               10
Confidence level                0.95

                           Estimated
Required sample size              62
```



Vall d'Hebron Institut de Recerca

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Sample size for proportion

$$Precision = z_{1-\alpha/2} * ee = 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = \frac{1.96^2\, \hat{p}(1-\hat{p})}{precision^2} =$$

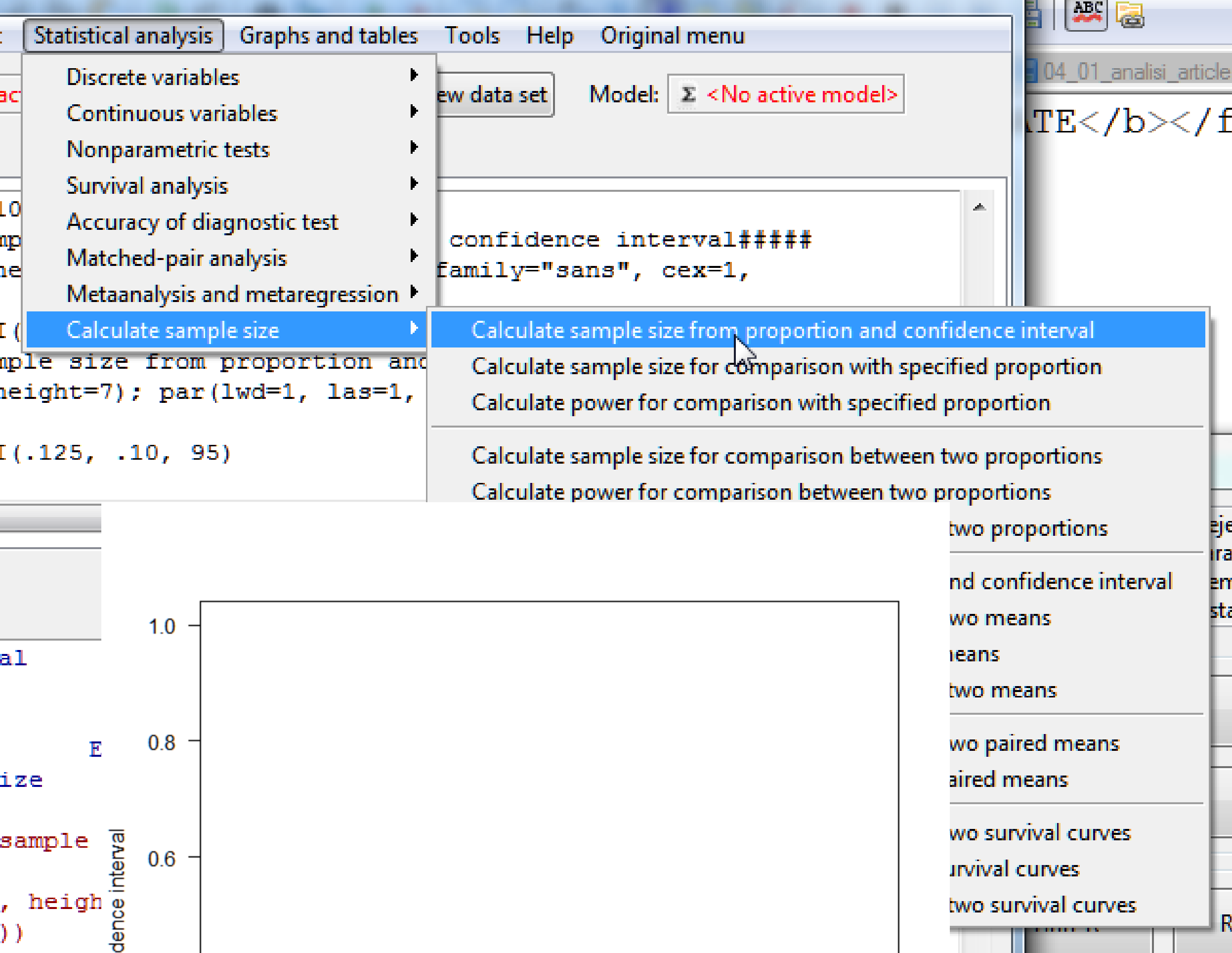
Assume precision is 5% (Interval = p±.05) and confidence level is 95%

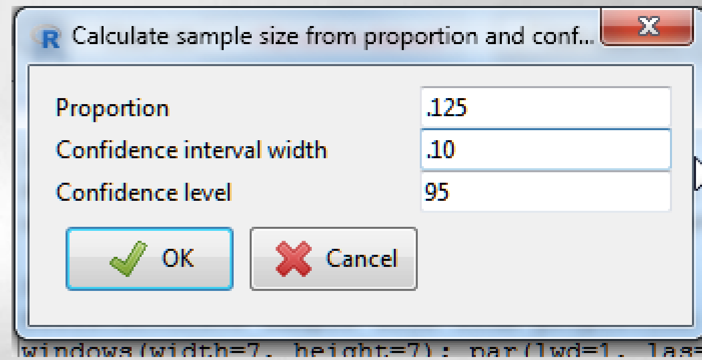
- If it is known that p is around 12.5%

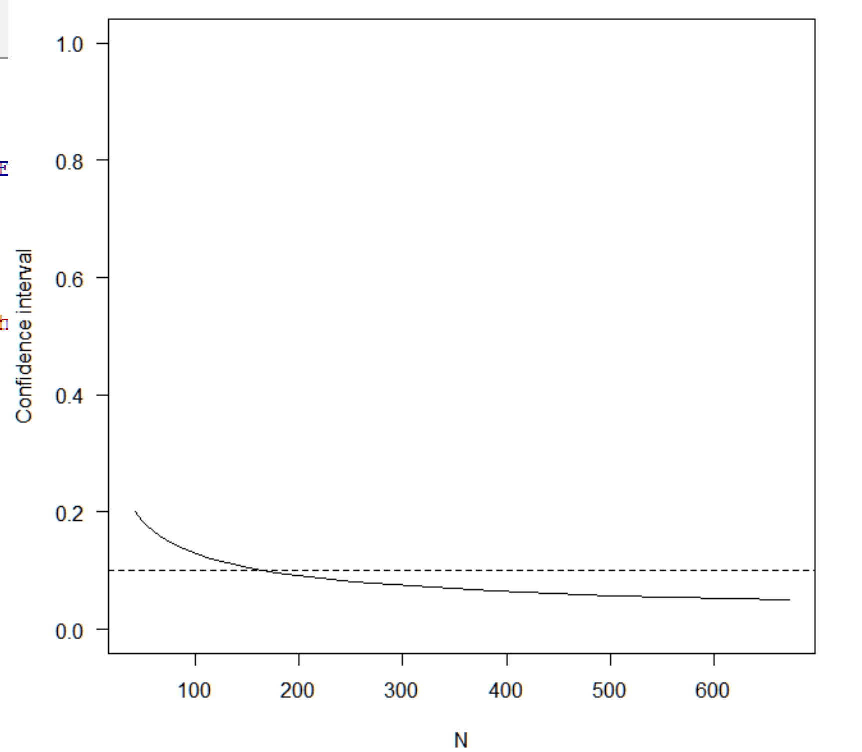
  $n = 1.96^2 \, .125\,(1-.125)/.05^2 = 168$

- If p is unknown maximum sample size will be if p=.50

  $n = 1.96^2 \, .5\,(1-.5)\,/.05^2 = 384$







Assumptions	
P	0.125
Confidence interval	0.1
Confidence level	0.95

	Estimated
Required sample size	169