Sample Size calculations

Curs d'Estadística Bàsica per a la Recerca Biomèdica

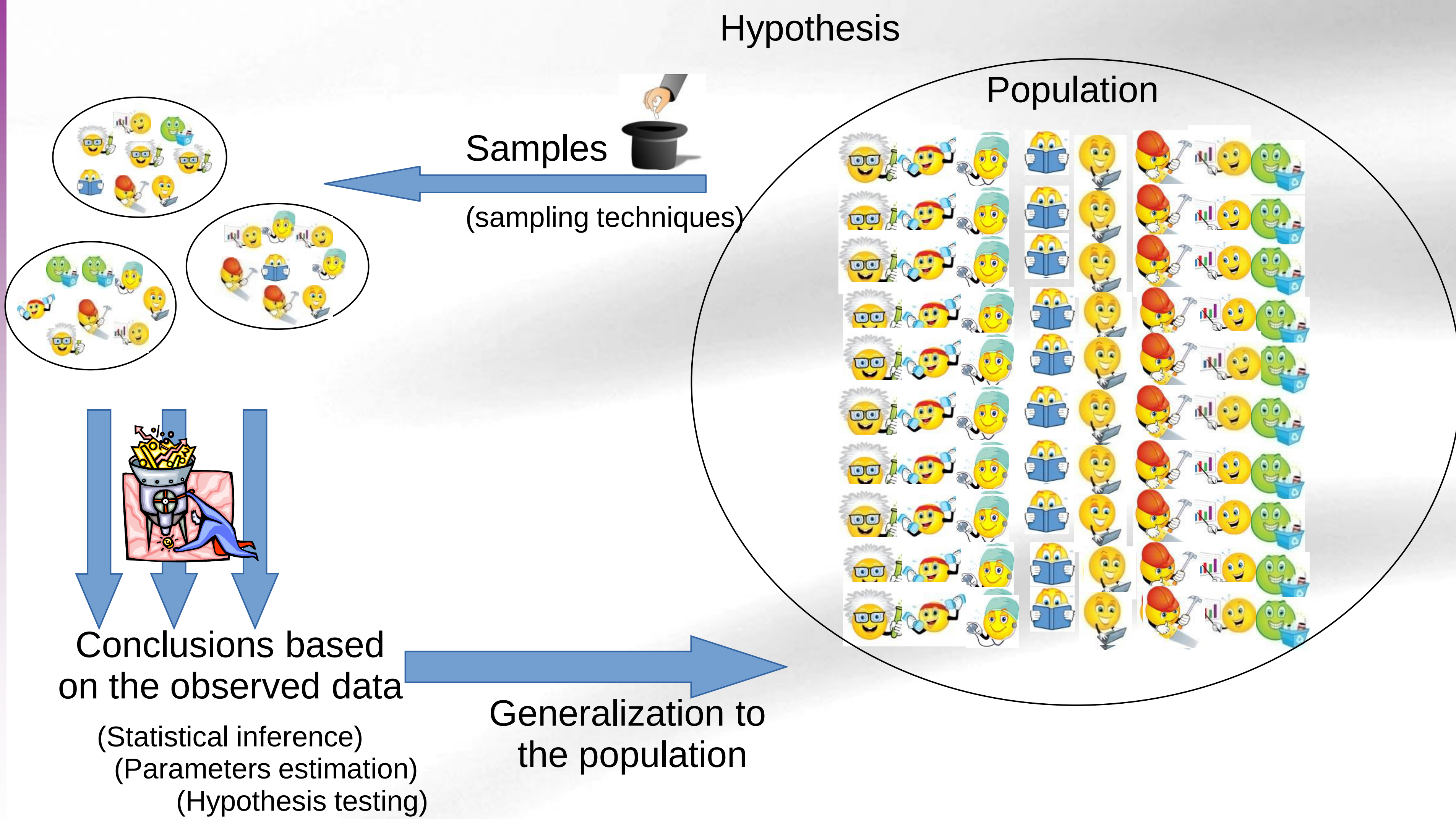
UEB – VHIR

Santiago Pérez-Hoyos i Alex Sánchez-Pla

santi.perezhoyos@vhir.org

alex.sanchez@vhir.org

The objective of statistical inference



Sample Size in Statistical Studies

- Statistical inference is used to *generalize*,
  - It helps obtain conclusions from samples,
  - and apply them to populations,
  - with a certain degree of (known) precision.
- This can be made only if
  - Some assumptions hold (e.g. Normality)
  - The sample size is ***big enough*** as to warrant the desired precision.

Vall d'Hebron
Institut de Recerca

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Preliminaries

- Before discussing sample size calculations there are several things to keep in mind
  - Type of calculations depend on study goal.
    - Estimation
    - Testing
  - Preliminary concepts to be used
    - Standard error of an estimator
    - Confidence interval
    - *Type I and type II errors. Power of a test*

Standard error of the mean

- A measure of how variable is the sample mean when computed in distinct samples.
  - Standard deviation of the distribution of sample means
- Usually it is defined as population standard deviation divided by squared root of sample size
  - It is estimated substituting population by sample deviation

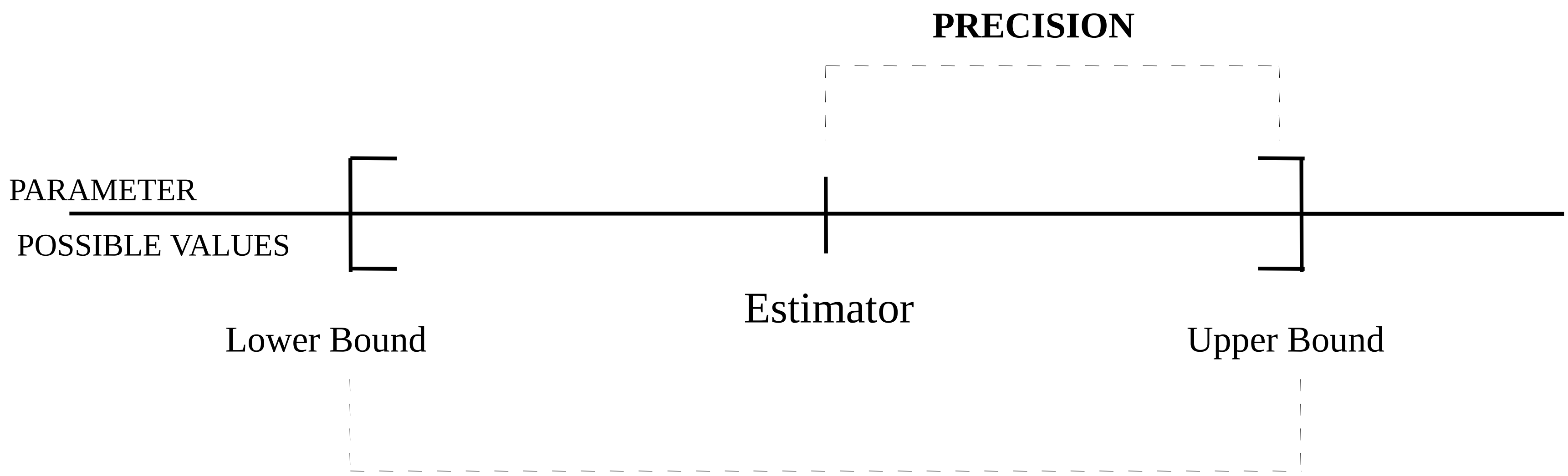
$$\text{standard error} = \frac{\sigma}{\sqrt{n}} \cong \frac{s}{\sqrt{n}}$$

Standard error of a proportion

- The standard error of a proportion (SEP) is computed similarly to the SEM.
  - Instead of the standard deviation it uses the population proportion in the formula.
  - Because *p* is usually unknown it is subsituted by its estimator.
  - It is common to set: *q=1-p*

$$SEP(Std.Err.Prop) = \sqrt{\frac{p \cdot (1-p)}{\mathrm{n}}} \cong \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{\mathrm{n}}}$$

Confidence interval



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

Formulas for confidence intervals

- Data normally distributed
  - Population variance known (unrealistic assumption)

$$\bar{X}_n - z_{\varepsilon/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{\varepsilon/2}\frac{\sigma}{\sqrt{n}}$$

  - Population variance unkown, estimated by sample variance

$$\bar{X}_n - t_{\varepsilon/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X}_n + t_{\varepsilon/2}\frac{s}{\sqrt{n}}$$

- Data: Counts of presence or absence of an event
  - Sample must be "big enough"

$$\hat{p} \pm z_{\varepsilon/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}; \quad n \ge 30, n\hat{p} \ge 5, n\hat{q} \ge 5$$

- $z_{\varepsilon/2}$ are quantiles of standard Normal N(0,1) distribution

$1-\varepsilon$	0,90	0,95	0,99
$z_{\varepsilon/2}$	1,64	1,96	2,58

Example: Confidence interval for the mean

- Goal: Estimate ureic nitrogen concentracion in serum (SUN) in rats that have been eating a certain diet.
- A sample of size 10 has been taken.
- Confidence interval is computed from formula (2) above

```
> x10
1.648943 20.960346 22.915030 27.348437 14.613271 10.705787 -
5.131364, 22.863318 41.924915 27.298092

> t.test(x10)
        One Sample t-test
data:  x10
t = 4.3016, df = 9, p-value = 0.001986
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  8.778153 28.251202
sample estimates:
mean of x
 18.51468
```

Confidence interval for a proportion

Problem

- A molecular diagnosis lab is doing tests to detect hereditary venous pathology (PVH).
- In a series of **150** affected patients **18** show in their genètic profile the AGx allele for the gene related with the disease.
- With a confidence of 99% which is the estimation for the percentatge of AGx individuals between people affected by PVH?

Solution

- Relative frequency in the sample: $\hat{p} = \frac{18}{150} = 0.12$
- Conditions that make the approximation reliable are verified

$$n \geq 30, n\hat{p} \geq 5, n\hat{q} \geq 5$$

- From this one may compute:

$$\hat{p} \pm z_{\varepsilon/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.12 \pm 2.56\sqrt{\frac{0.12 \times 0.88}{150}}$$

- With a 99% confidence proportion is between 0.052 and 0.188

Computing confidence interval with R

```
> prop.test(x=18, n=150, conf.level = 0.99, correct = TRUE)

        1-sample proportions test with continuity correction

data:  18 out of 150, null probability 0.5
X-squared = 85.127, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.0648676 0.2088192
sample estimates:
   p
0.12
```

Exercise

- Simulate 3 random samples from a normal population of mean 15 and Standard deviation 2.
  - Sample sizes must be 9, 25 and 100 respectively
  - Compute a 95% confidence interval for the mean in each sample
- Use the following code

```
x9 <- rnorm (n=9, mean=15, sd=2)
x25 <- rnorm (n=25, mean=15, sd=2)
x100 <- rnorm (n=100, mean=15, sd=2)
t.test(x9)
t.test(x25)
t.test(x100)
```

- What do you observe?

Exercise

- Simulate 3 random samples from a Binomial population of "true probability"=0.12=12%.
  - Sample sizes must be 25, 100 and 500 respectively
  - Compute a 95% confidence interval for the proportion in each sample
- Use the following code

```
y25 <- rbinom (n=25, size=1,prob=0.12)
y100 <- rbinom (n=100, size=1,prob=0.12)
y500 <- rbinom (n=500, size=1,prob=0.12)
prop.test(sum(y25), length(y25), conf.level = 0.95, correct = TRUE)
prop.test(sum(y100), length(y100), conf.level = 0.95, correct = TRUE)
prop.test(sum(y500), length(y500), conf.level = 0.95, correct = TRUE)
```

- What do you observe?

Sample Size

Sample Size Calculation

- Some questions must be answered before we can compute the sample size needed to estimate the mean or percentage.
  - Precision (interval range) of estimations (*" how accurate I want the estimate to be"?*)
  - Level of confidence of estimations (*" how confident will I be on the estimation"?*)

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

Remember: Confidence Interval for the mean

$$\bar{X}_n \pm z_{\varepsilon/2} \frac{\sigma}{\sqrt{n}} = \bar{X}_n \pm precision$$

$$precision = z_{\varepsilon/2} \times \frac{\sigma}{\sqrt{\mathrm{n}}} \Rightarrow$$

$$n = \frac{z^2{}_{\varepsilon/2}\sigma^2}{precision^2}$$

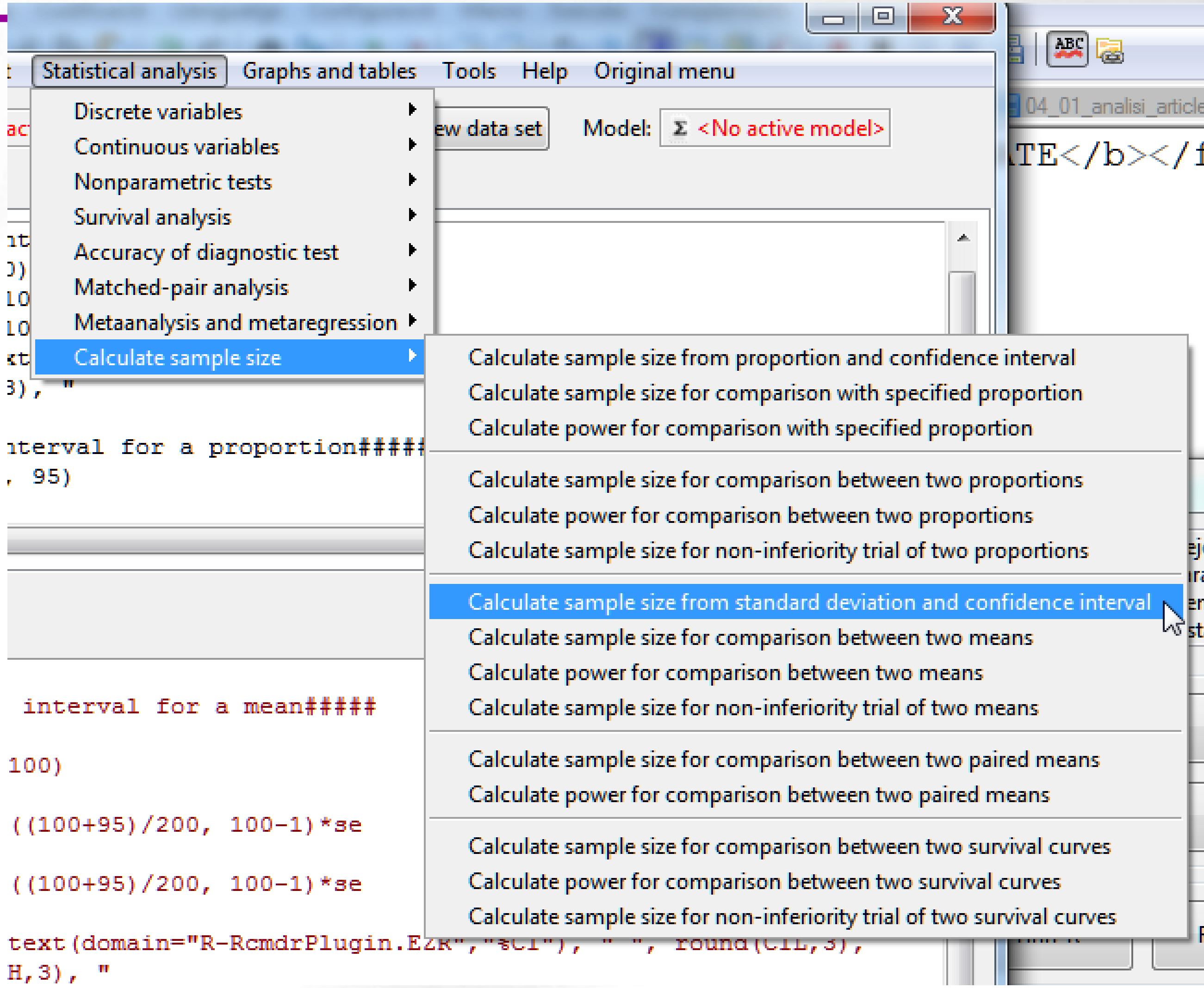
Example:
The sample size needed to estimate the mean with a confidence interval of width 10 (precision =10/2=5), a confidence level of 95%, if we know that the standard deviation is 20, will be:

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

Vall d'Hebron Institut de Recerca

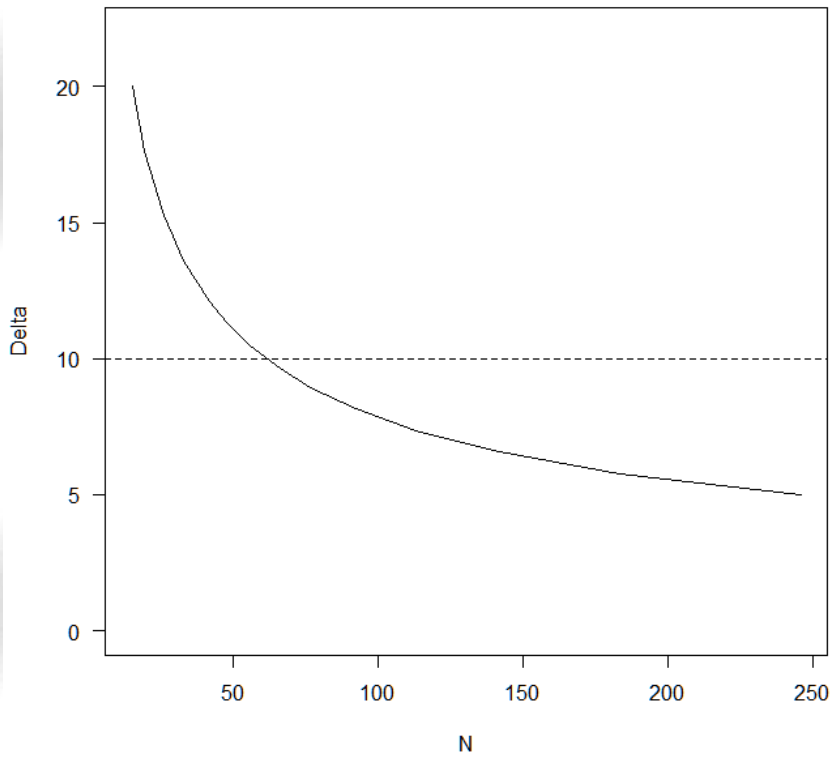
Computing Sample Size with eZR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA



Computing Sample Size with eZR

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

                                    Estimated
Required sample size                       62
```



Sample size for proportion

$$\Pr ecision = z_{\varepsilon/2} \times ee = z_{\varepsilon/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow$$

$$n = \frac{{z_{\varepsilon/2}}^2 \hat{p}(1-\hat{p})}{precision^2}$$

If p is unknown one can take p=q=0.5

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

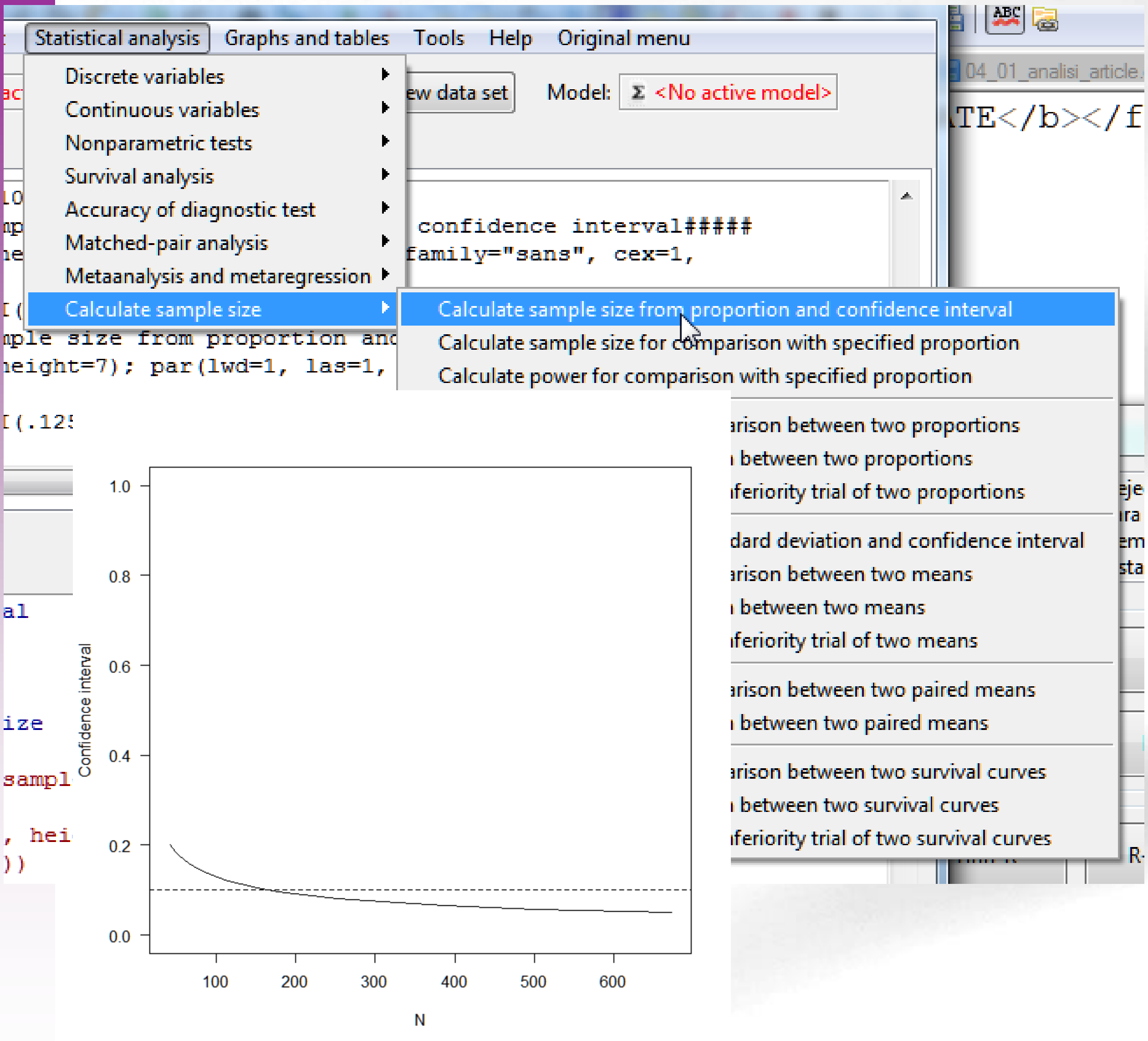
- If it is known that p is around 12.5%

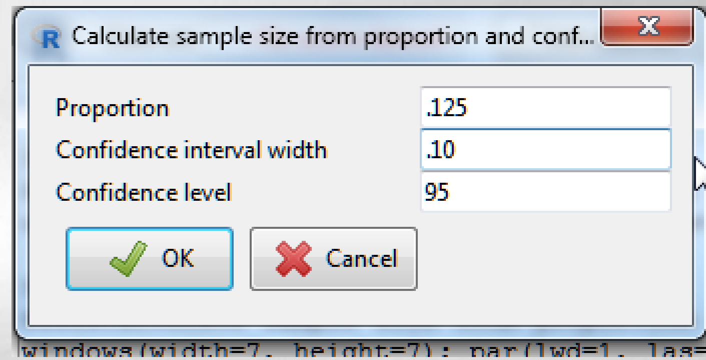
  $n = 1.96^2\ .125\ (1-.125)/.05^2 = 168$

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

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

Computing Sample Size with eZR





Assumptions	
P	0.125
Confidence interval	0.1
Confidence level	0.95

	Estimated
Required sample size	169

Sample size calculations for testing

- Similarly to sample size calculations for estimation, several points need to be considered so that the right question is:

- *What is the sample size needed to detect at least a difference* Δ *with the null hypothesis with a power* β *and a confidence (1-*α*)*

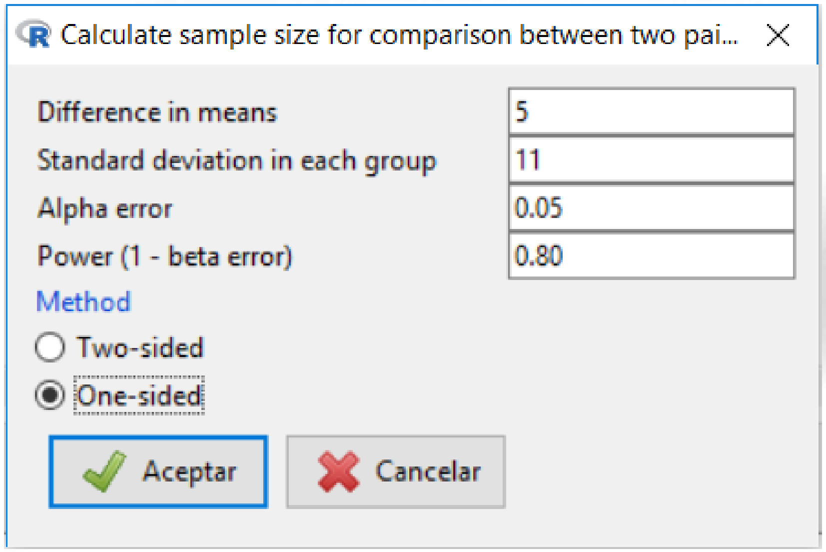
  - The computations also need to know or estimate parameters such as standard deviation or the percentatge.

Examples

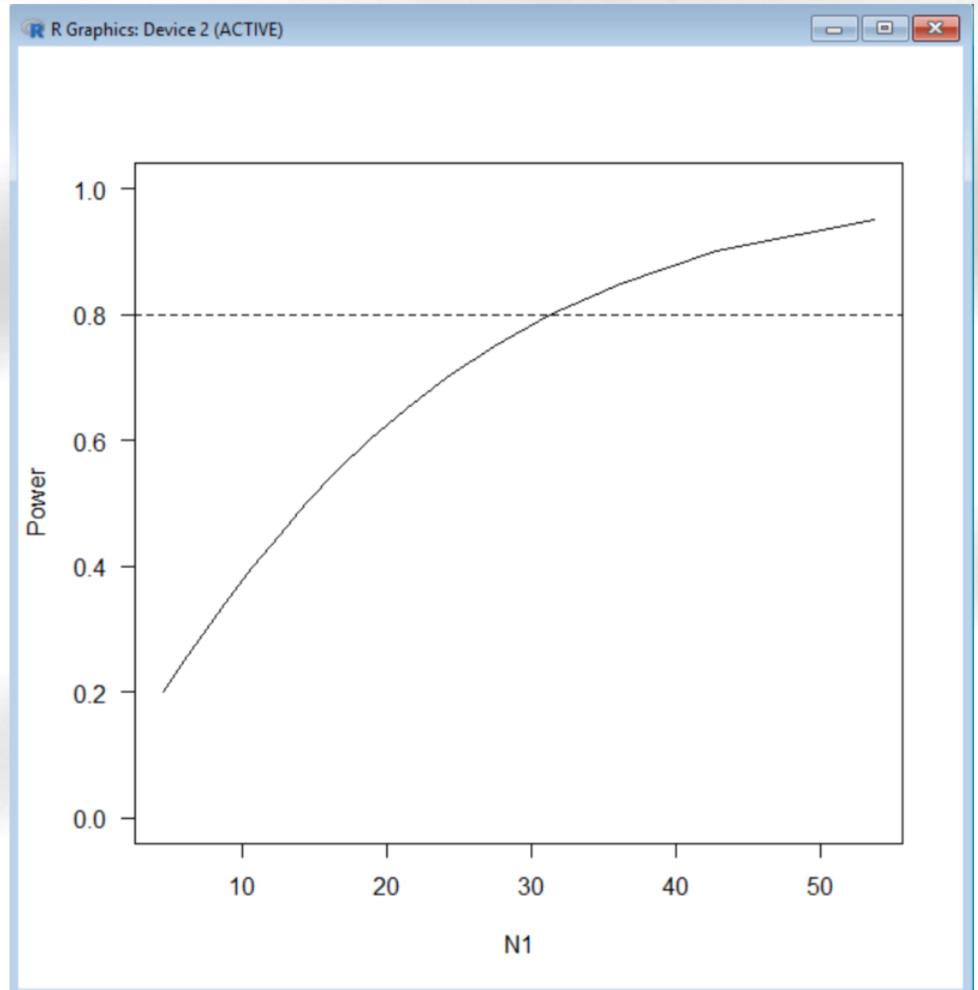
- The question:
  - What sample size is needed to test the belief that systolic pressure in a hypertense population is 90 or bigger than that
- Needs to be re-stated as:
  - *What sample size is needed to test the belief that systolic pressure in a hypertense population is 90 or bigger than that with a difference of at least 5 units, a power of 80% and a confidence of 95% assumint that the standard deviation is 11?*

Computing sample size with eZR

- eZR can only compute sample size for paired differences, so we cheat a little:




```
> SampleMeanPaired(5, 11, 0.05, 0.80, 1)
                              Assumptions
Difference in means                     5
Standard deviation                     11
Alpha                                0.05
                                one.sided
Power                                 0.8

Required sample size    Estimated
N                              32
```



Recall: Truth, Decision, Errors

TRUTH → DECISION ↓	Null Hypothesis **True**	Null Hypothesis **False**
Test does not reject null hypothesis	Significance level ✓	Type II Error β
Test rejects null hypothesis	Type I Error α	✓ Power (1- β)

Concept review: Power

- The power of a test describes the probability of correctly rejecting the null hypothesis that is, rejecting H_0 when it is false.
- A good test "controls" the probability of type I error and has a power "as big as possible".
  - Control of type I error is warranted by the way the test is built (with a given high confidence).
  - Power cannot be warranted but it depends on
    - The minimum difference to be detected by the test
    - The sample size
    - The population variability

Factors affecting power

- Power cannot be warranted simultaneously with type I error but it depends on:
  - The minimum difference to be detected by the test
    - *The bigger the minimum difference → the bigger the power*
  - The sample size
    - *The bigger the sample size → The bigger the power*
  - The population variability
    - *The bigger the variability → The smaller the power*
- Usually three of the previous four are set and the fourth is computed.
  - This is called **"power analysis"**

Vall d'Hebron Institut de Recerca

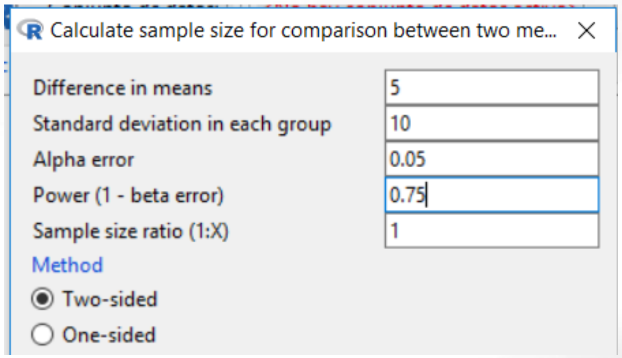
UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Some examples using R

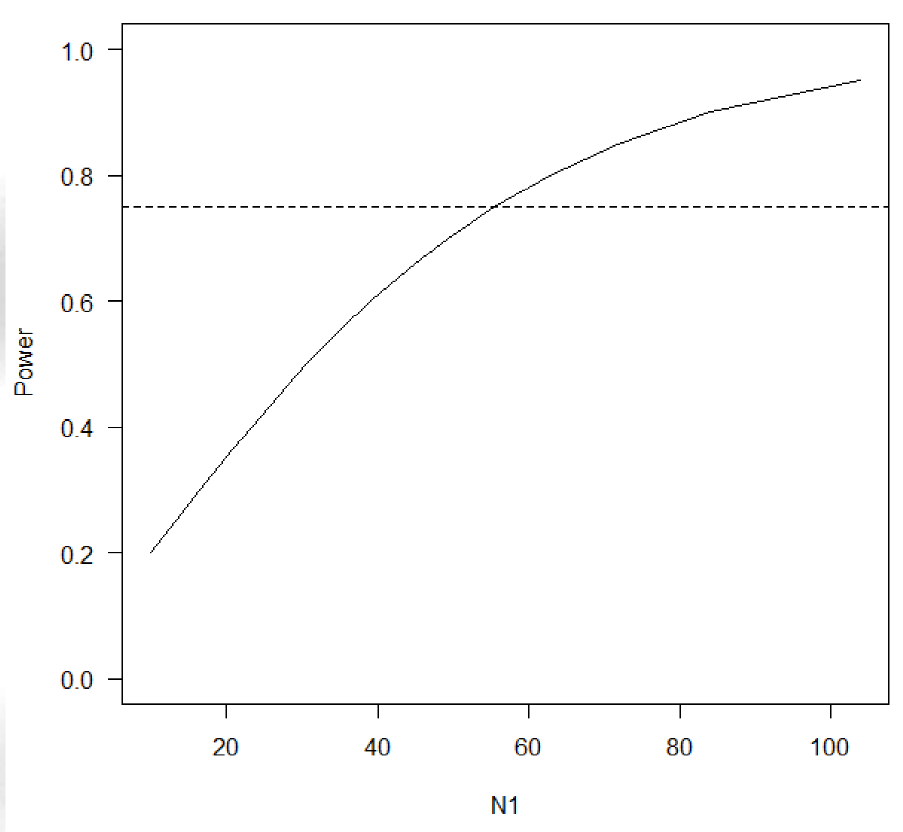
- **What is the minimum sample size needed** to detect *a difference of at least 5* among two groups whose *standard deviation is 10* if one wishes *to attain a power of 0.75*?
- **What is the power attained** if one uses a *sample size of 20* (per group) to detect a *minimum difference of 5* between two groups assuming that *the standard deviation (in both groups) is 10*.

Some examples using R

Sample size from power

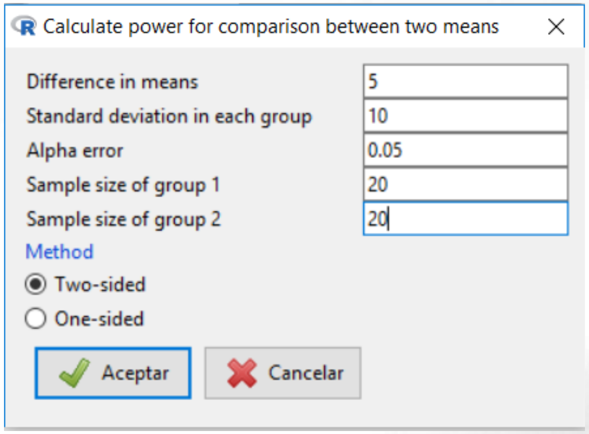



```
Assumptions
Difference in means              5
Standard deviation              10
Alpha                         0.05
                         two-sided

Power                         0.75
N2/N1                            1

Required sample size     Estimated
N1                              56
N2                              56
```



Some examples using R

Power from sample size




```
Assumptions
Difference in means           5
Standard deviation           10
Alpha                      0.05
                      two-sided

Sample size
N1                           20
N2                           20

                      Estimated
Power                     0.352
```