

Sample Size calculations

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR

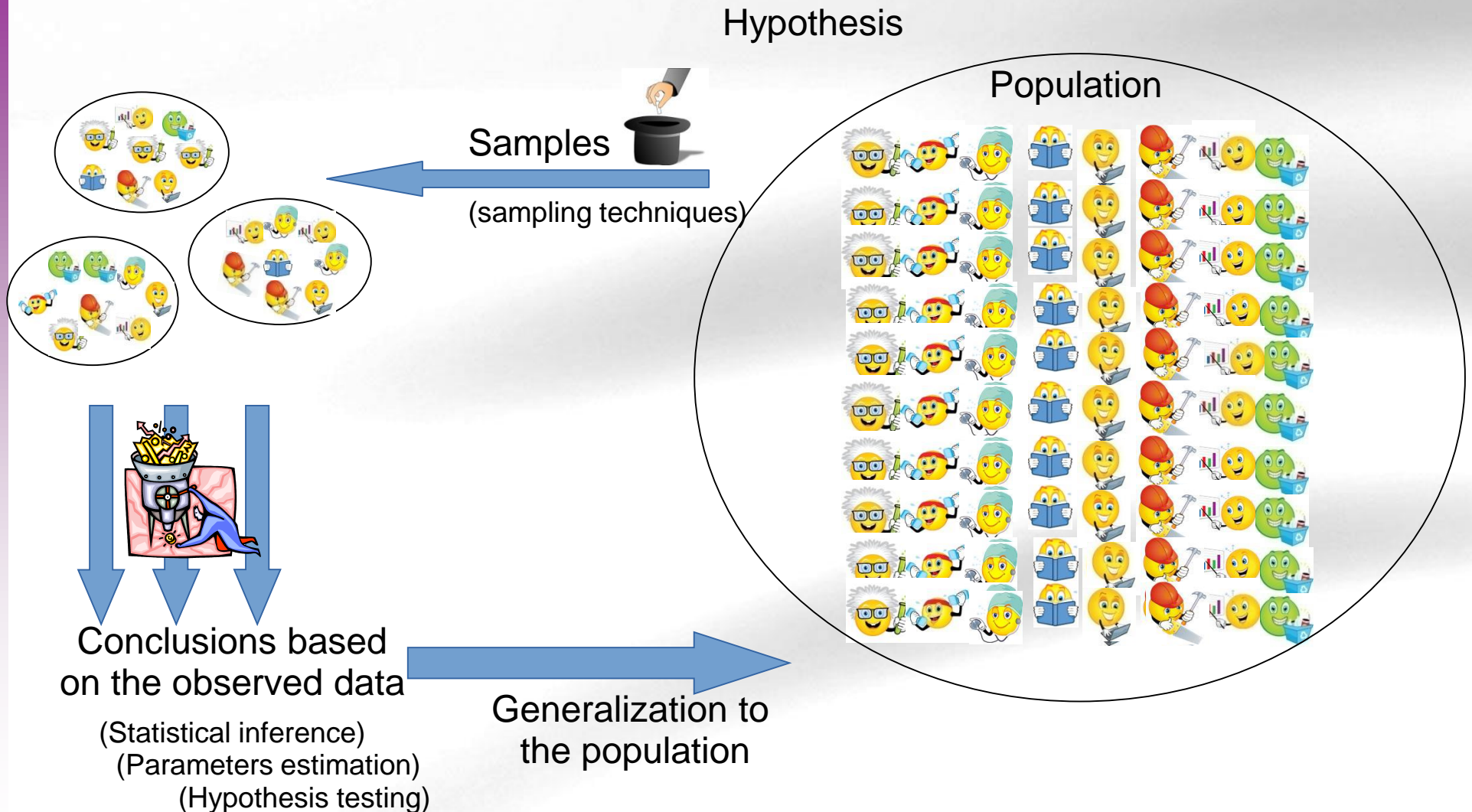
Santiago Pérez-Hoyos i Alex Sánchez-Pla

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) accuracy.
- This can be made only if
 - Some assumptions hold (e.g. Normality)
 - The sample size is ***big enough*** as to warrant the desired accuracy.



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

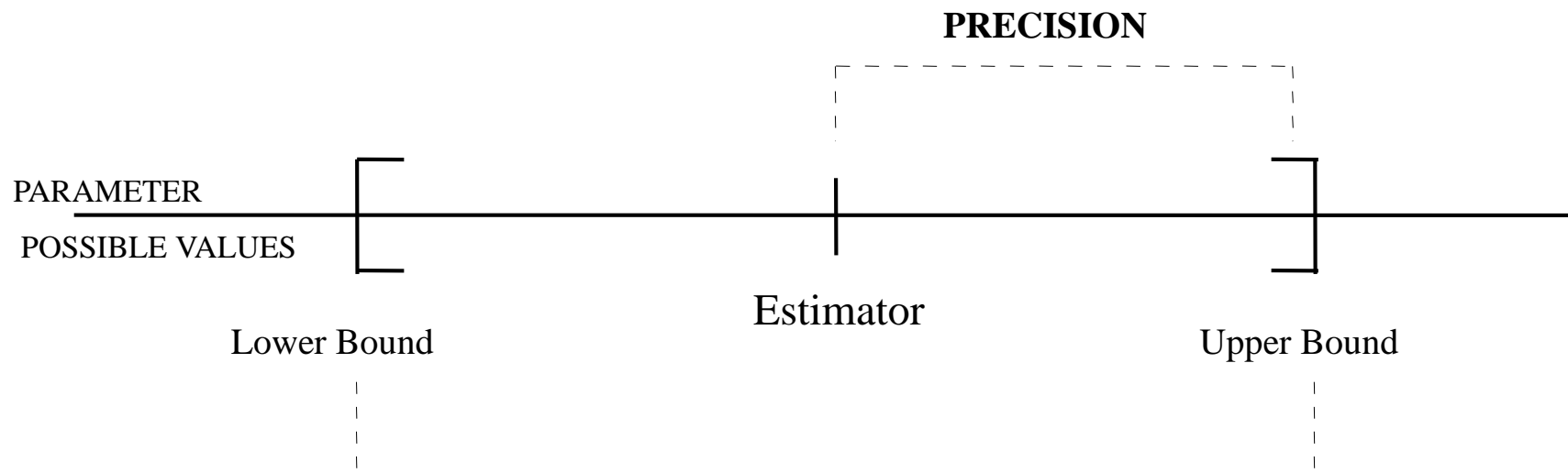
$$\text{standard error} = \frac{\sigma}{\sqrt{n}} \cong \frac{s}{\sqrt{n}}$$

Standard error of a proportion

- The standard error of a proportion is computed similarly to the SEM.
 - Instead of the standard deviation it uses the population proportion in the formula.
 - Because p is usual unknown it is substituted by its estimator.
 - It is common to put $q=1-p$

$$\text{standard error} = \frac{p \cdot (1 - p)}{\sqrt{n}} \cong \frac{\hat{p} \cdot (1 - \hat{p})}{\sqrt{n}}$$

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)
 - Population variance unknown, estimated by sample variance
- Data: Counts of presence or absence of an event
 - Sample must be “big enough”

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

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

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

Example: Confidence interval for the mean

- Goal: Estimate ureic nitrogen concentration 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 genetic profile the AGx allele for the gene related with the disease.
- With a confidence of 99% which is the estimation for the percentage 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?

Sample Size

Sample Size Calculation

- Some questions must be answered before we can compute “sample size”
 - 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

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

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

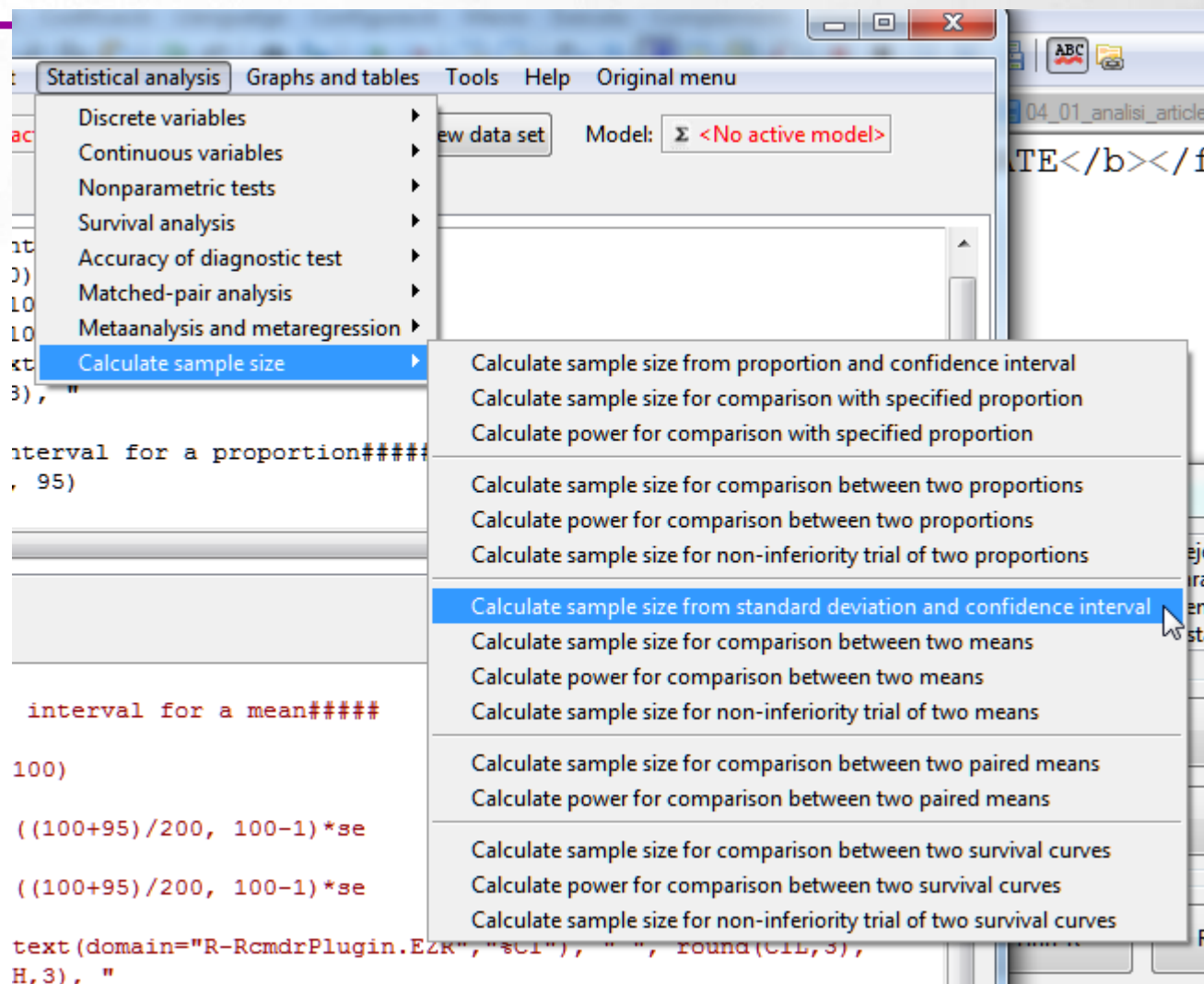
$$n = \frac{z_{\varepsilon/2}^2 \sigma^2}{\textit{precision}^2}$$

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:

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

Computing Sample Size with eZR





Computing Sample Size with eZR

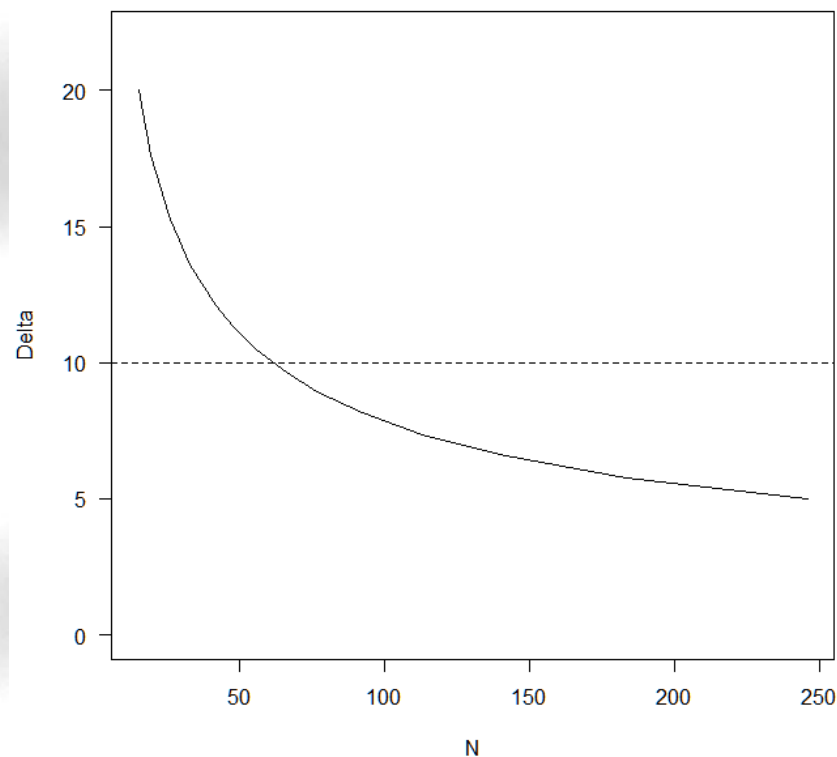
```
> SampleMeanCI(20, 10, 95)
```

Assumptions

Standard deviation	20
Confidence interval	10
Confidence level	0.95

Estimated

Required sample size	62
----------------------	----



Sample size for proportion

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

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

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

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

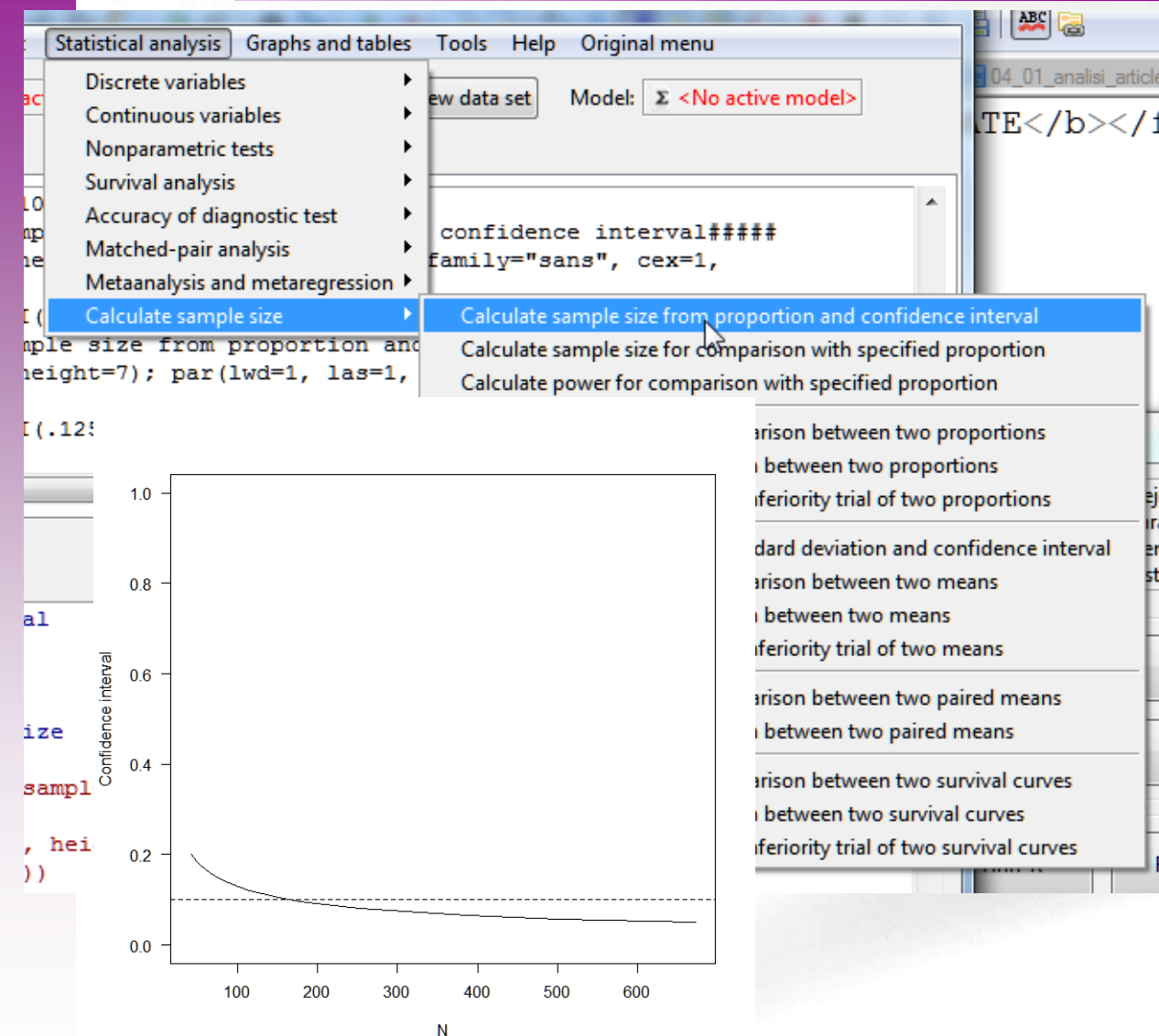
- If it is known that p is around 12.5%

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

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

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

Computing Sample Size with eZR



The dialog box 'Calculate sample size from proportion and conf...' is shown. It contains the following fields:

- Proportion: .125
- Confidence interval width: .10
- Confidence level: 95

There are 'OK' and 'Cancel' buttons at the bottom.

Assumptions

P	0.125
Confidence interval	0.1
Confidence level	0.95

Estimated
Required sample size 169