

Introduction to Statistical Inference

Curs d'Estadística Bàsica per a la Recerca Biomèdica

UEB – VHIR
Alex Sánchez)

Alex.sanchez@vhir.org

Table of contents

- **Case study problem**
- **The objectives of statistical inference**
- **Sampling issues**
- **Estimators**
- **Precision of estimators. Standard errors**
- **Sample size computations.**
- **Confidence intervals.**

Our data set: ***osteoporosis.csv***

Bone density ("bua") is a key variable to determine if a person is suffering osteoporosis.

But, how do other variables affect or determine the level of bone density?

Have these other variables differences regarding their distribution and properties?

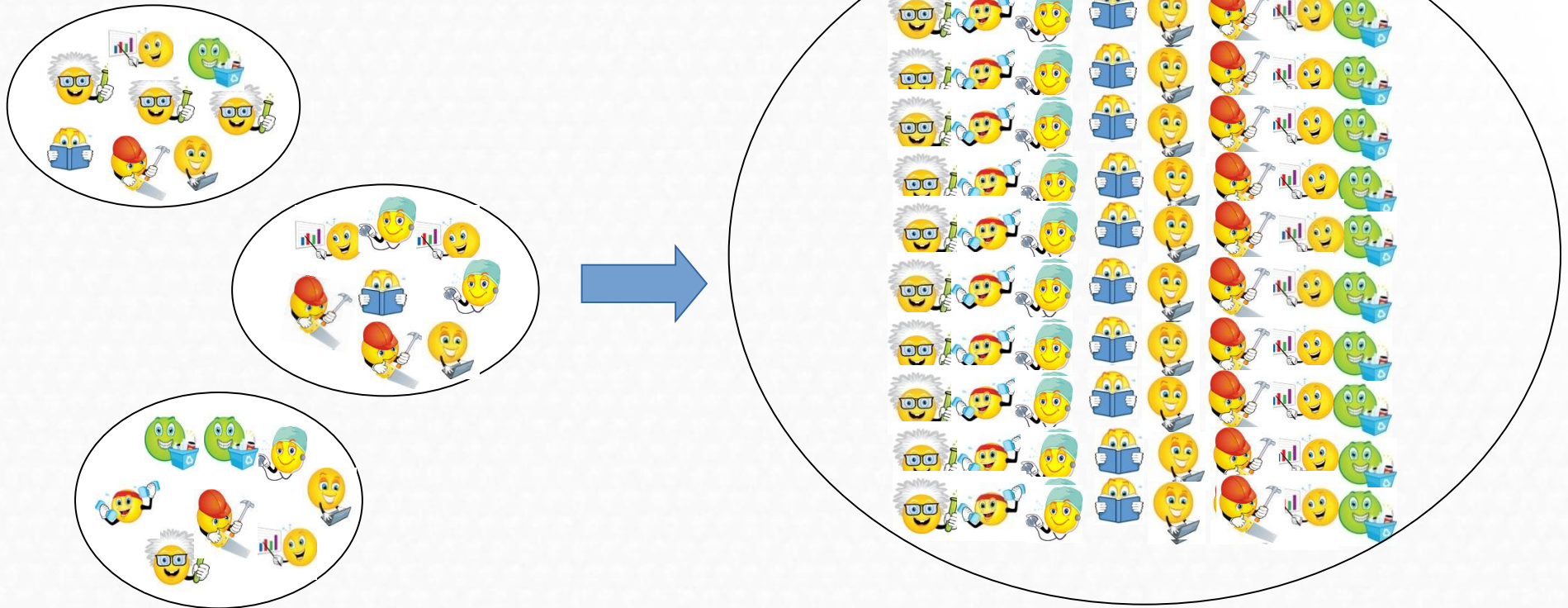
In our case study, we may want, for instance, to...

- **Explore the distribution of the bone density variable.**
- **Estimate the % of women with an unusually low bone density**
- **Compare it between menopausal and non-menopausal women.**

grupedad	peso	talla	imc	bua	clasific	menarqui	edad_men	menop	tipo_men	nivel_ed	imc_cat
57	55 - 59	70	168	24.8	69	OSTEOPENIA	12	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
46	45 - 49	53	152	22.94	73	OSTEOPENIA	13	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
45	45 - 49	64	158	25.64	81	NORMAL	14	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
53	50 - 54	78	161	30.09	58	OSTEOPENIA	10	50	SI	NATURAL	PRIMARIOS
46	45 - 49	56	157	22.72	89	NORMAL	13	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
45	45 - 49	63.5	170	21.97	76	NORMAL	14	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
48	45 - 49	86	161	33.18	87	NORMAL	11	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
50	50 - 54	61.5	164	22.87	74	NORMAL	10	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
51	50 - 54	60.5	158	24.23	58	OSTEOPENIA	14	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
57	55 - 59	64	149	28.83	61	OSTEOPENIA	13	50	SI	AMBAS	PRIMARIOS
48	45 - 49	70.3	160	27.46	67	OSTEOPENIA	12	48	SI	OVARECTOMIA	SECUNDARIOS
55	55 - 59	74.4	160	29.06	68	OSTEOPENIA	14	50	SI	NATURAL	PRIMARIOS
50	50 - 54	55.5	154.5	23.25	73	OSTEOPENIA	11	48	SI	NATURAL	PRIMARIOS
56	55 - 59	89	166	32.3	61	OSTEOPENIA	14	47	SI	NATURAL	PRIMARIOS
49	45 - 49	50.6	157	20.53	68	OSTEOPENIA	14	40	SI	NATURAL	PRIMARIOS
50	50 - 54	71.4	152	30.9	74	NORMAL	14	48	SI	AMBAS	PRIMARIOS
49	45 - 49	78	157	31.64	62	OSTEOPENIA	12	46	SI	NATURAL	PRIMARIOS
58	55 - 59	72	162	27.43	65	OSTEOPENIA	11	54	SI	NATURAL	PRIMARIOS
61	60 - 64	68	155.5	28.12	65	OSTEOPENIA	14	50	SI	NATURAL	PRIMARIOS
55	55 - 59	75	161	28.93	92	NORMAL	13	50	SI	NATURAL	PRIMARIOS
48	45 - 49	66.5	153	28.41	11	OSTEOPOROSIS	11	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
59	55 - 59	101	156	41.5	82	NORMAL	12	45	SI	AMBAS	SIN ESTUDIOS
68	65 - 69	66.5	145	31.63	57	OSTEOPENIA	13	50	SI	NATURAL	PRIMARIOS SIN
69	65 - 69	70	168	24.8	48	OSTEOPOROSIS	13	45	SI	NATURAL	PRIMARIOS
48	45 - 49	60.1	153	25.67	86	NORMAL	14	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
50	50 - 54	67	159	26.5	105	NORMAL	12	45	SI	NATURAL	PRIMARIOS

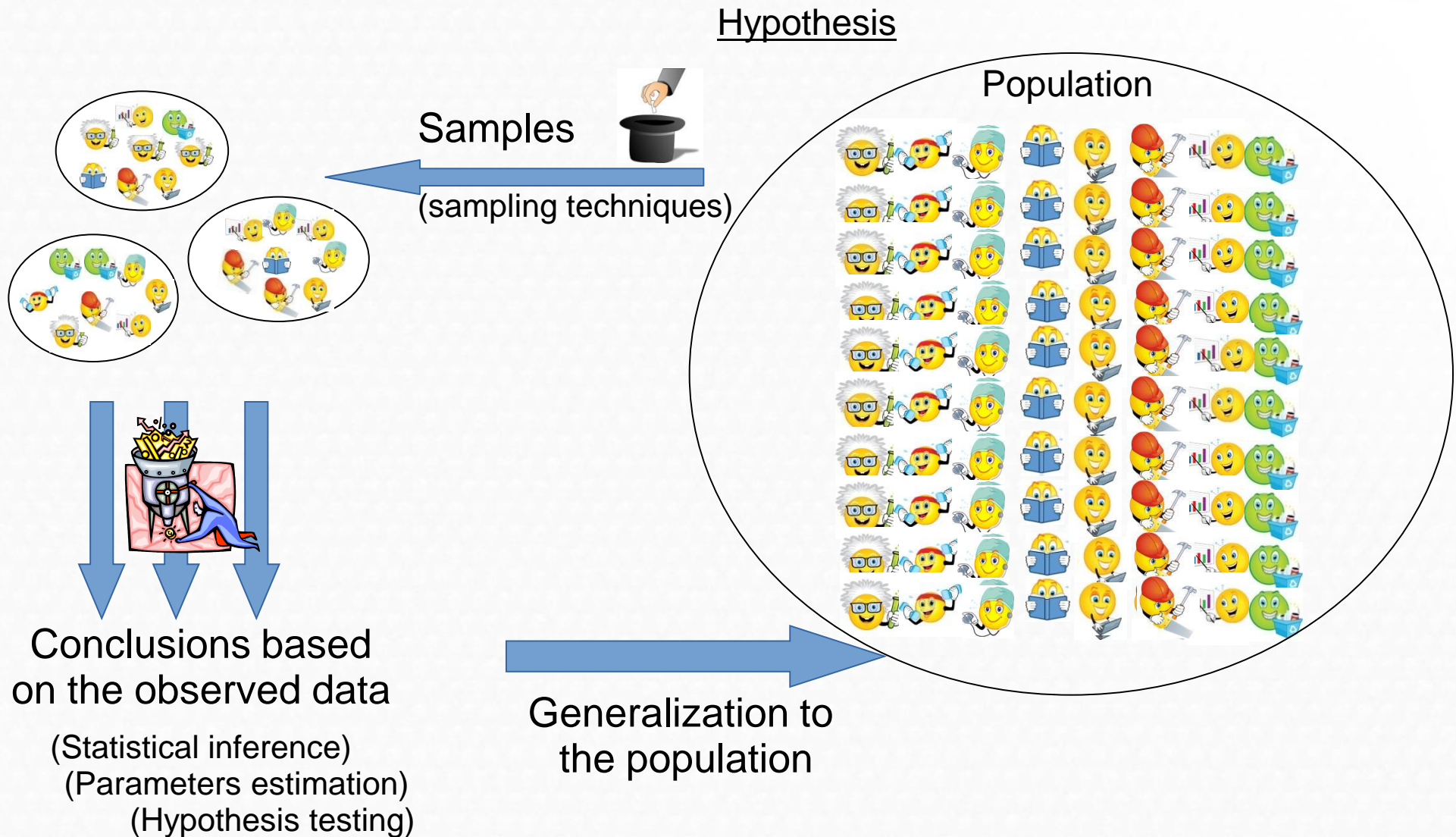
The objective of statistical inference

Taking the observed (measured)
values of a group of samples...



... determine the properties of the entire population.

The objective of statistical inference



- Sample(s) should be representative of the population where they come from.
- Sample size and sampling methods must follow the statistical criteria depending on the the experiment.
- It is **very common** to have sampling limitations (samples not randomly obtained, desired sample size unaffordable...)

Much better to know our limitations than the sampling techniques!

We are usually faced with one of the following types of questions

ESTIMATION

- When **we wish to learn some characteristics of our population**, such as, but not limited to, the **mean** or the **proportion**
 - *The percentage of non osteopenic or menopausal women*
 - *The mean bone density in each of these groups*

HYPOTHESIS TESTING

- When we wish to check about some statement on some characteristic of the population or we wish to make some comparisons
 - *Is it true that the mean bone density is smaller than 75 in menopausal women?*
 - *Can we state that non menopausal women have a higher bone density than menopausal?*

- An estimator is a numeric value calculated on a sample that we *hope* to be a good approximation of a certain parameter in the population.
- Intuitively, we work with many estimators, such as the mean or a computed percentage of a given sample, that we assume that are somehow characterizing a population.
 - **It is not always obvious to decide which is the best estimator for each parameter.**
 - **There are methods to define the precision of the estimators and quantify the underlying error(s).**

Example 1-Estimating the level of triglicerides

- Two investigations aiming at studying the level of triglicerides in the blood of obese mice have collected two samples
- One consisting of 4 individuals
 - 23.88, 14.26, 9.77, 17.30,
- Another consisting of 9 individuals
 - 23.88, 14.26, 9.77, 17.30, 12.28, 14.93, 21.32, 20.30, 15.19.
- The mean value of both variables is *very similar*
 - $mean(X_4) = 16.30$
 - $mean(X_9) = 16.58$

Example 1-Doing the computations

Using R

Store values in a vector

```
triglic <- c(23.88,14.26,  
            9.77, 17.30,12.28,14.93,  
            21.32, 20.30, 15.19)
```

Compute the mean

```
> mean (triglic)  
[1] 16.57974
```

With R-commander

Data

Nuevo conjunto de datos

[Entrar les dades]

[Nombrar la variable]

Statistics

Resúmenes

Resúmenes numéricos

- An obvious question when we choose an estimator is: *how precise is it to approximate the value of the population parameter.*
- This can be answered using **the standard error of the estimator**
- The standard error is a great quantity because it:
 - **Informs** about the **precision** of our estimates
 - Helps building another type of estimators: **confidence intervals**
 - Helps finding formulae to compute **sample size for estimation**

La primera pregunta (“què tan precisa és l’estimació”) es pot respondre calculant
l’error estàndar de la mitjana

$$EEM = SEM = \frac{\hat{S}}{\sqrt{n}}$$

*Precisió de la mitjana mostral per
estimar la poblacional*

A més dispersió de les dades → major EEM →
→ L’estimació de la mitjana és menys precisa.

Més mida mostral → menor EEM →
→ L’estimació de la mitjana és més precisa.

$$EEM = SEM = \frac{\hat{S}}{\sqrt{n}}$$

$$EEP = SEP = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Standard error of the mean

Standard error of the proportion

The more disperse are the data (the higher "S")

→ The **higher** is the SEM →

→ The estimation of the mean is **less precise**

The higher is the sample size ("n")

→ The **smaller** is the SEM →

→ The estimation of the mean is **more precise**

- Sometimes we wish to have an estimation that
 - Provides a unique approximation to a certain population characteristic
 - Accounts for the precision of the estimation.
- This can be done by combining
 - A point estimator such as the mean or proportion
 - With their standard error
- The combination is known as a ***confidence interval***

Confidence intervals are based on standard errors

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

$$\hat{p} \pm z_{\varepsilon/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

**Confidence interval for the mean
of a normal population**

**Confidence interval for the
proportion**

*The coefficient that multiplies the standard error ($t_{\varepsilon/2}$ or $z_{\varepsilon/2}$) to build a coefficient interval changes its value but it **is usually not far from two***

- In general R does not compute (has no functions) for the direct calculation of confidence intervals
- This can be done by calling the corresponding tests functions such as `t.test` or `prop.test`
- Some R commander plugins such as EZR allow this computations directly

Examples – CI for the mean using R or Rcmdr

```
> buaMenop <- bua[menop=="SI"]
```

```
> t.test(buaMenop)
```

```
One Sample t-test
data: buaMenop
t = 44.582, df = 67,
p-value < 2.2e-16
alternative hypothesis: true mean
is not equal to 0.95 percent
confidence interval:
67.30146 73.61031
mean of x 70.45588
```

In Rcmdr the menu options below will yield the same result as in the left pannel

Estadísticos

→ *Medias*

→ *Test-t para la media*

In EZR this can be done by simply providing the mean the standard deviation and the sample size

Examples – CI for the proportion

```
> table(menop)
```

```
> prop.test(x= 32, n= 100)
```

```
1-sample proportions test with  
continuity correction  
data: 32 out of 100,  
null probability 0.5  
X-squared = 12.25, df = 1,  
p-value = 0.0004653  
alternative hypothesis: true p is  
not equal to 0.5  
95 percent confidence  
interval: 0.2322385 0.4217920  
sample estimates: p 0.32
```

In RCmdr the menu options below will yield the same result as in the left pannel

Estadísticos

→ *Medias*

→ *Test-t para la media*

In EZR this can be done by simply providing the mean the standard deviation and the sample size

Examples – CI for the mean using EzR in R commander

- Proceed as you would do if doing the computation manually

- Pre-compute the mean and sd:

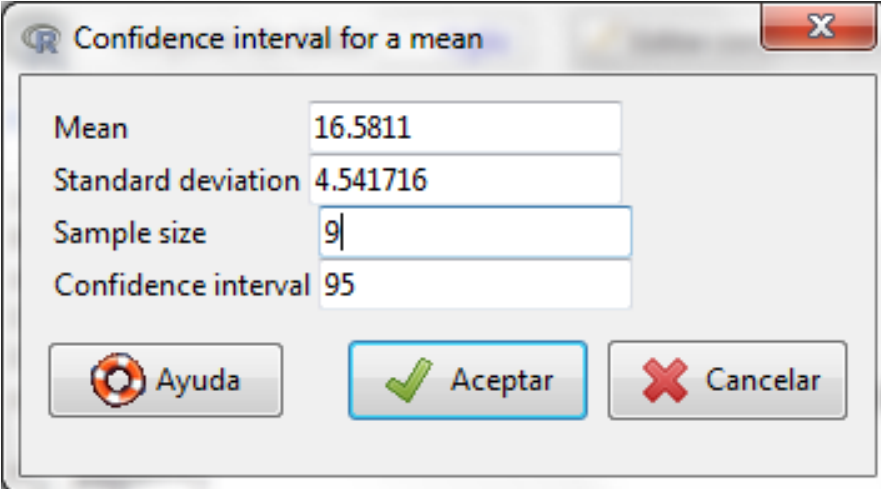
mean	sd	n
16.58111	4.541716	9

- Go to EzR and select:

Estadísticos

→ *Medias*

→ *Intervalo de confianza para la media*



Confidence interval for a mean

Mean 16.5811

Standard deviation 4.541716

Sample size 9

Confidence interval 95

Ayuda Aceptar Cancelar

95 %CI 13.09-20.072

- The standard error informs of how precise an estimation is if one knows the variability and the sample size

$$EE = \hat{\sigma} / \sqrt{n}$$

- We can proceed in the opposite sense: assuming we know the variability (e.g. from a pilot study) and the highest precision we wish to attain ("arm length" of a confidence interval: $\Delta = z_{\alpha/2} \cdot EE$)
- The sample size needed to attain this precision can be isolated from the formula

$$n = \left(\hat{\sigma} / \Delta \right)^2$$

$$n = \frac{t_{\varepsilon/2}^2 \hat{S}^2}{d^2}$$

Sample size needed to estimate the mean of a normal population with a given precision ***d***

$$n = \frac{z_{\varepsilon/2}^2 \hat{p}\hat{q}}{d^2}$$

Sample size formula to estimate a proportion

Left: assuming p is known from a pilot study

Right: assuming $p=q=0.5$

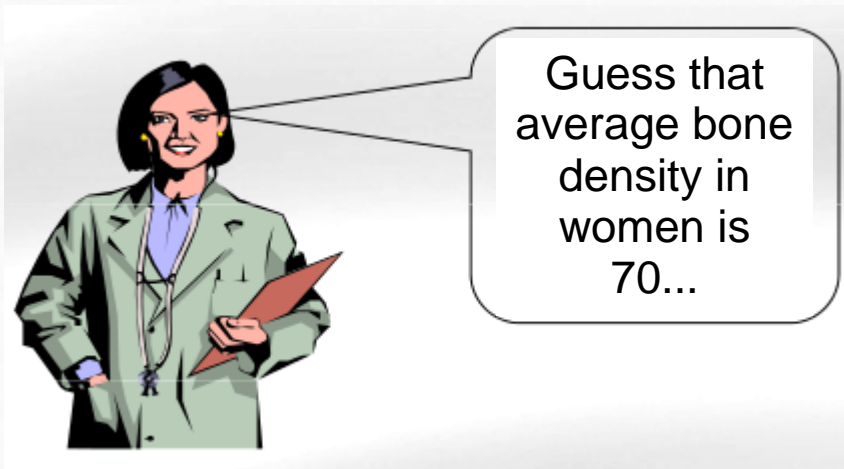
$$n = \frac{z_{\varepsilon/2}^2}{4d^2}$$

- In general R does not compute (has no functions) for sample size calculations for estimation purposes.
- One can use web sites such as:
<http://www.imim.cat/ofertadeserveis/software-public/granmo/index.html>
- If the goal is to do a hypothesis test this can be done using the `power` package
- Some R commander plugins such as EZR allow this computations directly

1. Extract a random sample of 100 patients from the osteoporosis database
2. Provide point estimates and confidence intervals for
 1. The percentage of menopausal women.
 2. The mean bone density for both menopausal and non-menopausal.
3. How many individuals should our sample contain if we wish (*"What sample size do we need to..."?*)
 1. To estimate the percentage of menopausal women with an error less than 2%, 5%?
 2. To estimate the mean bone density with an error smaller than 5?

Hypothesis testing

Hypothesis testing: Making decisions about populations



But... why not to check median, mode or other estimators?

Null hypothesis (H_0): depends on the question we want to answer

- Is the mean of the bua values equal to 70.0? H_0 : The mean of the bua values is 70.0
- Is the mean bua of the menop. and non-menop groups equal? H_0 : Both means are equal

Alternative hypothesis ($H_\alpha = H_a = H_1$): the opposite idea

- H_1 : The mean of the bua values is not equal to 70.0
- H_1 : Both means are different

H_1 only accepted if
clear evidence
that H_0 is not true

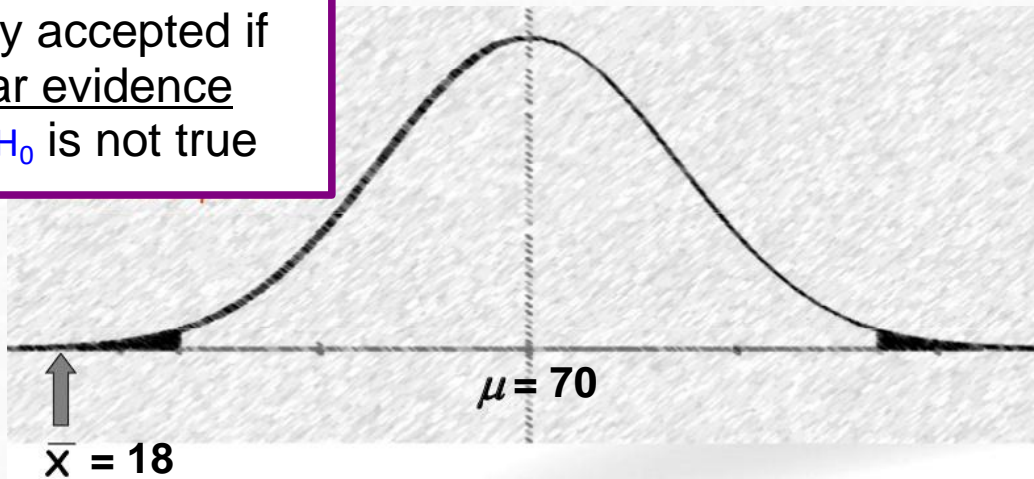
Null hypothesis (H_0): depends on the question we want to answer

- Is the mean of the bua values equal to 70.0? H_0 : The mean of the bua values is 70.0
- Is the mean bua of the menop. and non-menop groups equal? H_0 : Both means are equal

Alternative hypothesis ($H_\alpha = H_a = H_1$): the opposite idea

- H_1 : The mean of the bua values is not equal to 70.0
- H_1 : Both means are different

H_1 only accepted if
clear evidence
that H_0 is not true



If observed mean = 18
 H_0 is rejected

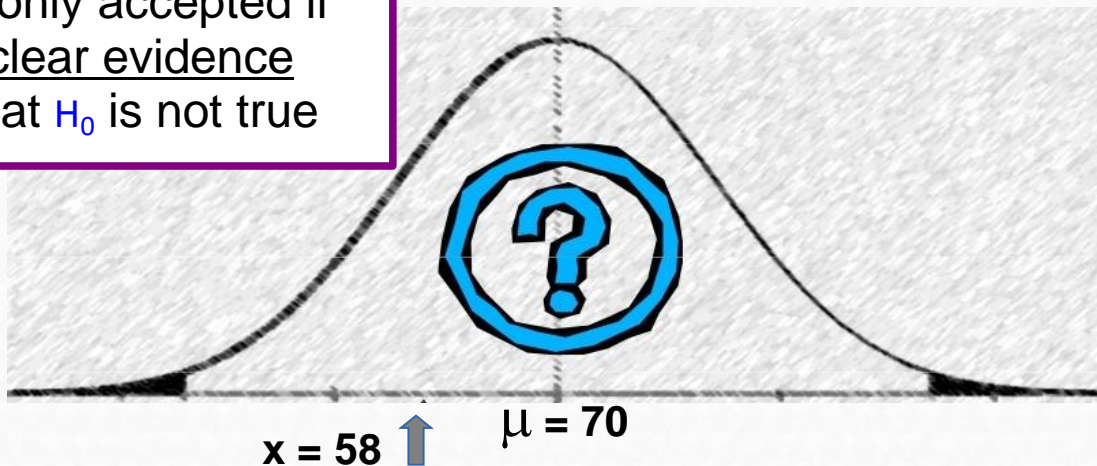
Null hypothesis (H_0): depends on the question we want to answer

- Is the mean of the bua values equal to 70.0? H_0 : The mean of the bua values is 70.0
- Is the mean bua of the menop. and non-menop groups equal? H_0 : Both means are equal

Alternative hypothesis ($H_\alpha = H_a = H_1$): the opposite idea

- H_1 : The mean of the bua values is not equal to 70.0
- H_1 : Both means are different

H_1 only accepted if
clear evidence
that H_0 is not true



If observed mean = 18
 H_0 is rejected

If observed mean = 58
 H_0 can not be rejected
(it not means H_0
can be accepted!!)

Null hypothesis (H_0): depends on the question we want to answer

- Is the mean of the bua values equal to 70.0? H_0 : The mean of the bua values is 70.0
- Is the mean bua of the menop. and non-menop groups equal? H_0 : Both means are equal

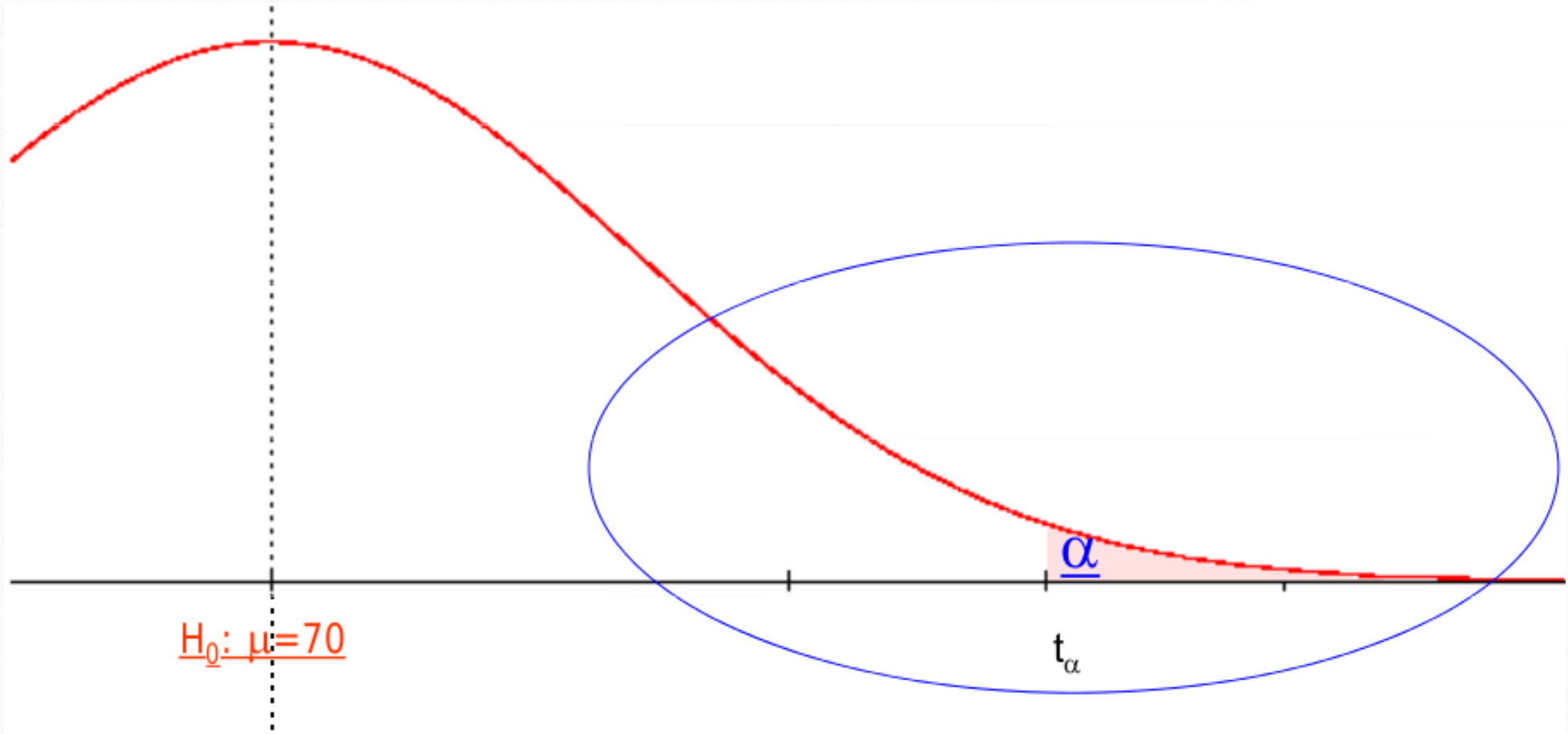
Alternative hypothesis ($H_\alpha = H_a = H_1$): the opposite idea

Only accepted if clear evidence
that H_0 is not true

- H_1 : The mean of the bua values is not equal to 70.0
- H_1 : Both means are different

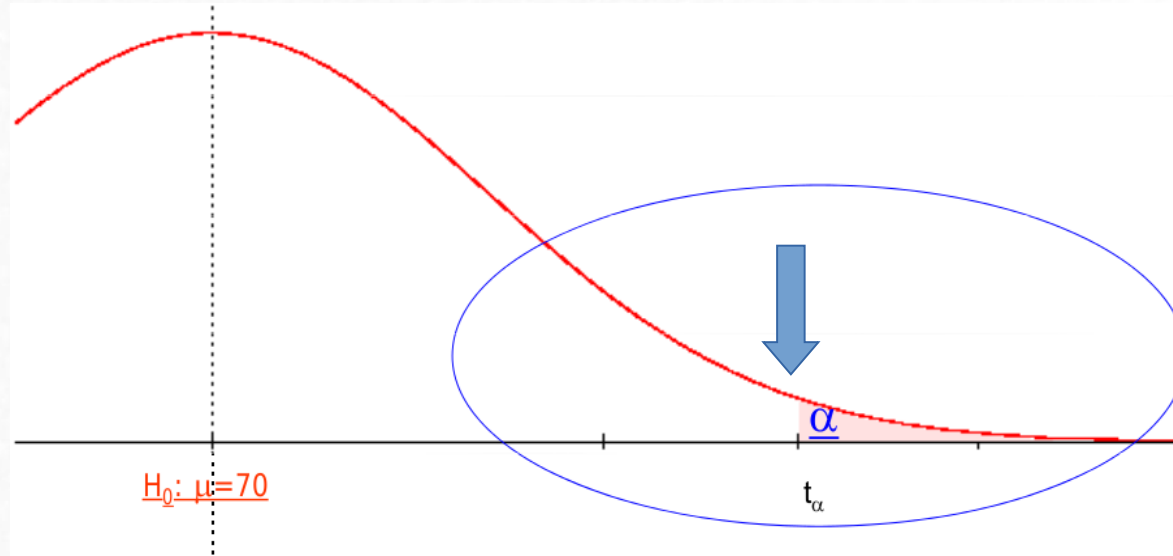
Test: Depends on the sample data (sample size, distribution of the variable of interest...). It is a measure of discrepancy between H_0 and the observed data.

Hypothesis testing: elements



Rejection value (critical value, T_α): Value obtained by the statistics test from which we reject H_0

Significance level

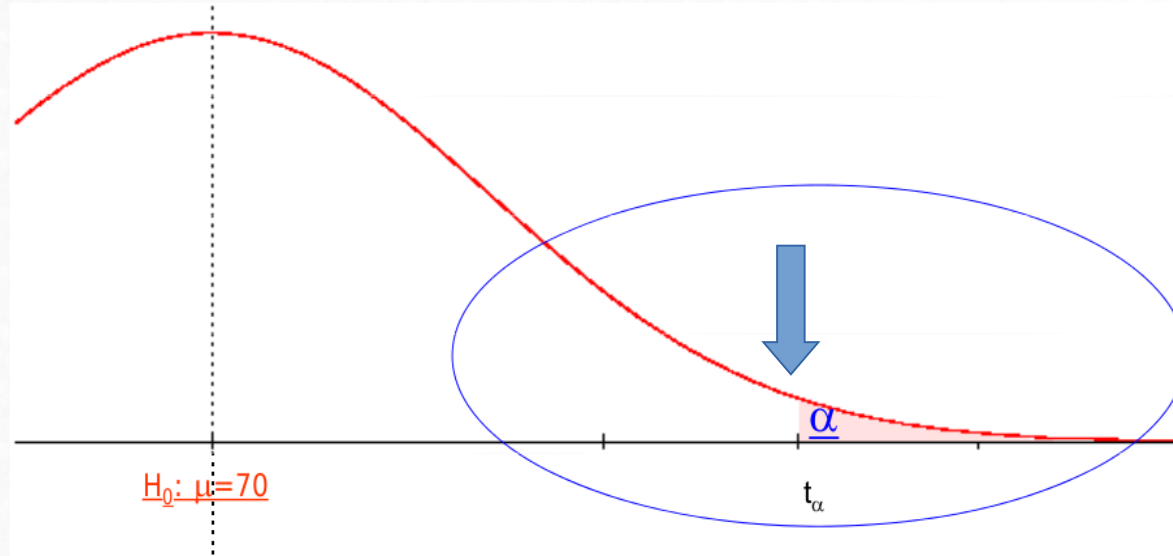


Significance level (α): Arbitrary level (that should be established *a priori*) to compare the obtained p-values in order to reject or not H_0

Is the probability of the critical region (the region left at one or two sides by the critical value)

Commonly set at (but not necessarily always) **0.05** (5%) or **0.01** (1%)

Significance level

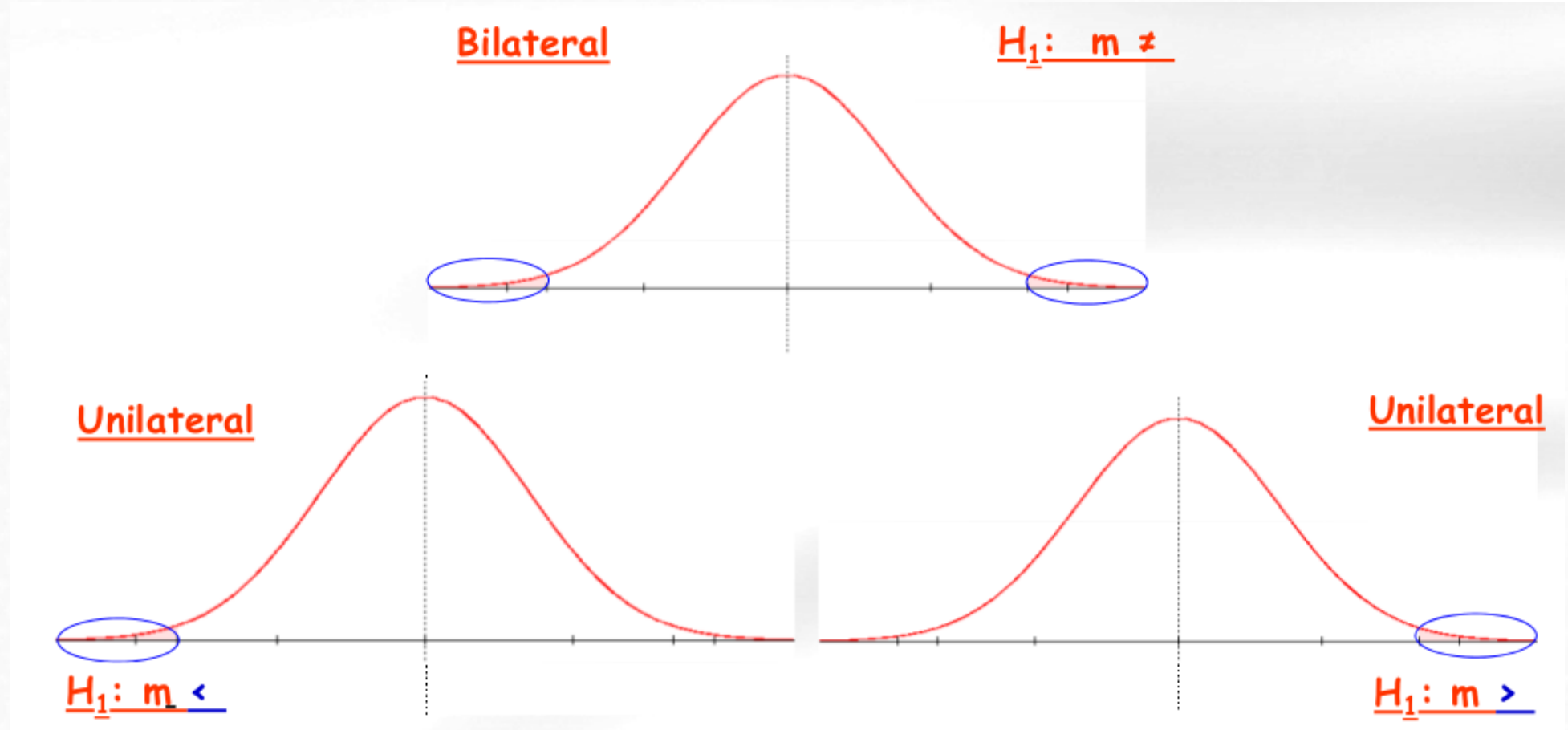


Significance level (α): Arbitrary level (that should be established *a priori*) to compare the obtained p-values in order to reject or not H_0

Is the probability of the critical region (the region left at one or two sides by the critical value)

Commonly set at (but not necessarily always) **0.05** (5%) or **0.01** (1%)

Bilateral / Unilateral tests

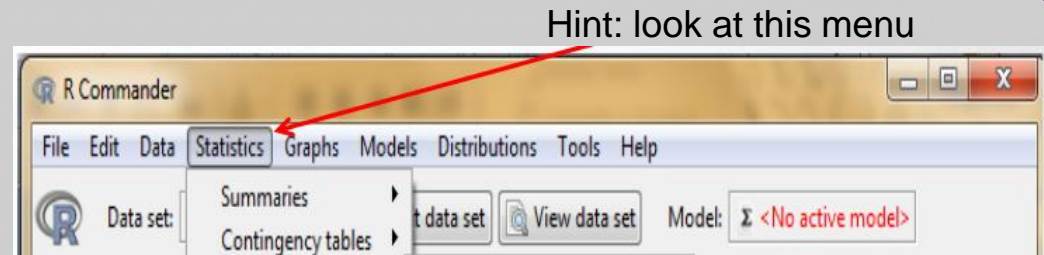


Case study problem

Our assumptions:

- The average "bua" value in our population is 70.
- The "bua" mean value in menopausal and non-menopausal women is not the same.

Exercise 2):



- Test if the mean bone density value is 70.0 or not
- Test if the mean bone density is equal or not between groups if we separate our observations by "menop" category

Exercise 2 – Solution (1)

➤ `t.test(bua, mu=75)`

One Sample t-test

data: bua t = -0.22669, df = 99,

p-value = 0.8211

alternative hypothesis: true mean
is not equal to 75

95 percent confidence interval:

71.68398 77.63602 sample

estimates: mean of x 74.66

In RCmdr the menu options below will yield
the same result as in the left pannel

Estadísticos

→ *Medias*

→ *Test-t para la media*

Exercise 2 – Solution (2)

> t.test (bua~menop)

```
Welch Two Sample t-test
data: bua by menop
t = 4.2203, df = 53.292,
p-value = 9.536e-05
alternative hypothesis: true
difference in means is not equal
to 0
95 percent confidence interval:
6.894761 19.380975 sample
estimates:
mean in group NO mean in group SI
83.59375 70.45588
```

In RCmdr the menu options below will yield the same result as in the left pannel

Estadísticos

→ Medias

→ Test-t para la media

Errors and power (in hypothesis testing)

H_0
(innocent)
(not speculative)

Data can lead to reject it

Accepted if data don't
show the contrary

Reject it by mistake (if it is true)
has severe consequences

H_1
(guilty)
(speculative)

Should not be accepted without
enough evidence

Reject it erroneously has less dramatic
consequences



Errors and power (in hypothesis testing)



Experiment
conclusions

Real population

	when H_0 is true	when H_1 is true
Do not Reject H_0	correct decision $p = 1 - \alpha$	Type II error $p = \beta$
Reject H_0	Type I error $p = \alpha$	correct decision $p = 1 - \beta$

Type I error = False positives

Type II error = False negatives

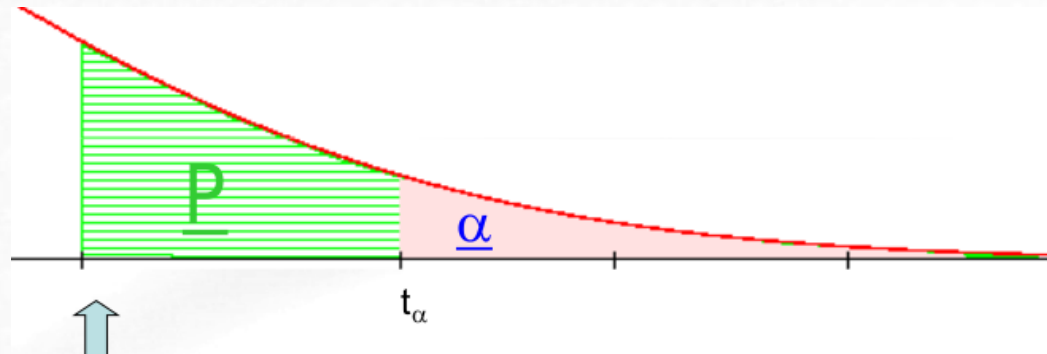
α = Prob [refusing H_0 | H_0 is true] = Proportion of **false positives** = **Significance level**

β = Prob [not refusing H_0 | H_0 is false] = Proportion of **false negatives**

$1 - \alpha$ = **confidence level** (95 - 99%) = Prob [refusing H_0 | H_0 is false]

$1 - \beta$ = **statistical power** = Prob [refusing H_0 | H_0 is false]

What does the obtained p-value mean?



- The probability that would correspond with the critical region starting exactly at the test value that we have obtained from our data, and
- the probability of having a new sample that differs more than our sample in comparison with H_0 , and
- the probability of finding, just by chance, a sample more rare than the one that we have used for the test.