Vall d'Hebron
Institut de Recerca

Principles of Hypothesis testing

UEB – VHIR

Santiago Pérez-Hoyos, Alex Sánchez-Pla, Miriam Mota and Ricardo Gonzalo

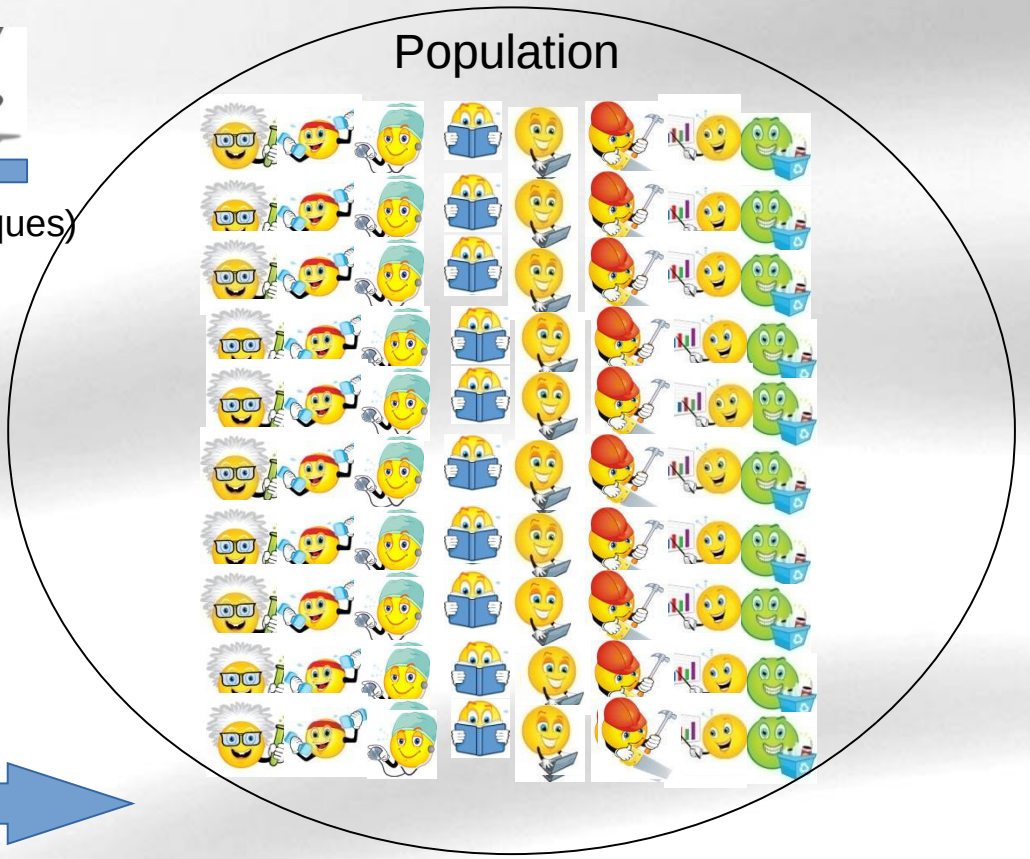
santi.perezhoyos@vhir.org

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · ISC Instituto de Salud Carlos III

The objectives of statistical inference

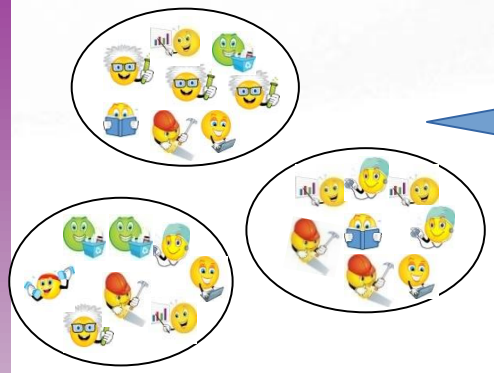
Hypothesis

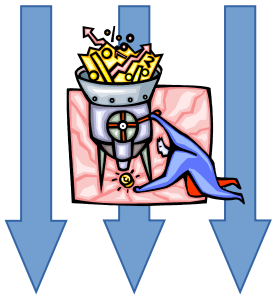
Population



Samples

(sampling techniques)





Conclusions based on the observed data

(Statistical inference)
(Parameters estimation)
(Hypothesis testing)

Generalization to the population

Statistical Inference Questions

Parameter estimation:

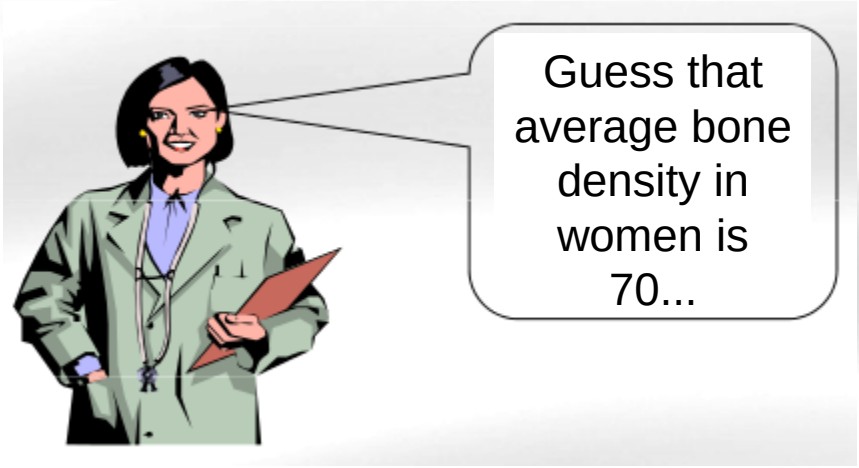
- After assuming population data follow a certain probability distribution (normal. Binomial, Poisson, etc) which are the value of the parameters that better fit the sample data.

Hypothesis testing:

- We have an assumption about the population data parameters
  - Population mean is equal to 10
  - The mean in population A is equal to the mean in population B

- Hypothesis testing tries to verify if sample data are compatible with that hypothesis

Hypothesis testing: Making decisions about populations



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

Case study problem I

Our guess:

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

Exercise 1):

- Explore osteoporosis data in order to get an idea about our first guess
- Do de same for the second question
- What other things you can figure out about the bone density in our population?

The average ''bua'' value in our population is 70.

```
library(dplyr)
# Read data
osteoporosis <- read.delim2("datasets/osteoporosis.csv", string
# Take subsample
# mean bone density
buaMean <- mean(osteoporosis$bua)
print(buaMean)
```

```
## [1] 73.297
```

```
t.test(osteoporosis[["bua"]])$conf.int
```

```
## [1] 72.2539 74.3401
## attr(,"conf.level")
## [1] 0.95
```

The ''bua'' mean value in menopausic and non-menopausic women is not the same

```
# Mean bone density ny groups
osteoporosis %>%
  group_by(menop) %>%
  summarize(m = mean(bua))
```

```
## # A tibble: 2 x 2
##   menop     m
##   <fct> <dbl>
## 1 NO     79.3
## 2 SI     70.7
```

Case study problem II

- Cohort study with new lung cancer cases after 12 years of follow up
  (The NHANES Epidemiologic Follow-up Study. Am J Epidemiol 1997;146:231-243)

Lung Cancer

		Yes	No	Total	% Disease
Fruit Consumption	High	44	2473	2517	**1.8%**
	Low	88	2429	2517	**3.5%**

Total lung cancer rate =132/ 5034=2.6%

- After this outcome can we say that fruit consumption is a protective factor for lung cancer?
- How can I say this results is not caused by sampling or random?

A framework for hypothesis testing

The "Null" and the "Alternative"

- We can establish two basic hypothesis

Null Hypothesis (H_0)

- Outcome is observed ***by chance***.
- ***No relationship*** among exposure and disease.

Alternative Hypothesis (H_1)

- ***There is relationship*** among exposure and disease.

How to decide which hypothesis?

- Select a test statistic that measures the discrepancy between data and null.
- Two approaches
  1. Look for a "cut-off" or "critical-value" such that values greater than this cut-off lead to rejecting the null hypothesis.
  2. Compute the probability to observe values greater than what you have observed *if the null were true.*

Hypothesis Testing

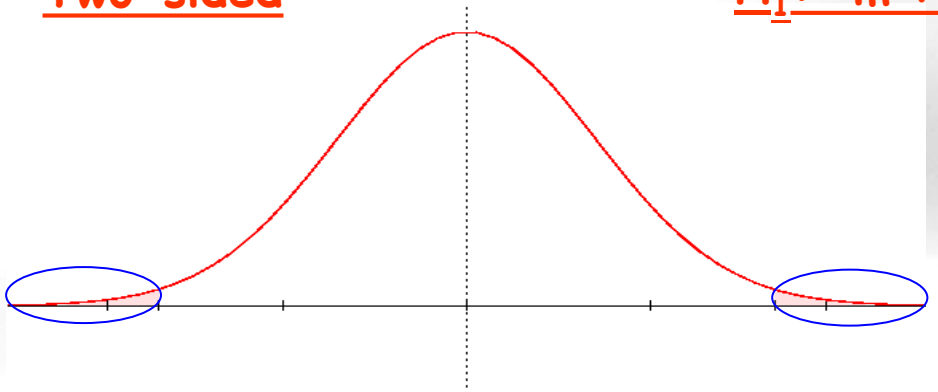
- Test population hypothesis from samples
  - Stablish Null Hypothesis(H_0)
  - Stablish Alternative Hypothesis (H_a)
  - Select statistical test to calculate probability ***under Null Hypothesis***
  - Decide after comparing test value with a critical value or probability under null hypothesis.

One-sided vs Two-sided

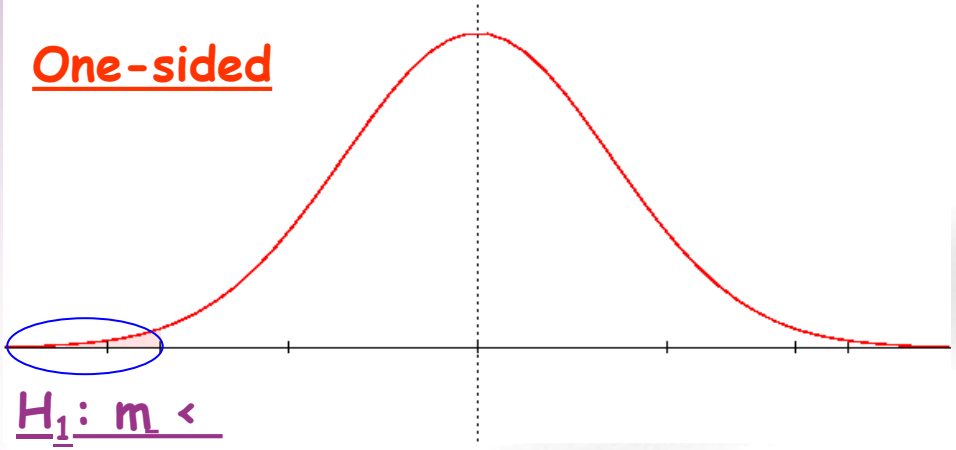
Critical value depends on the type of alternative Hypothesis

Two-sided

$H_1: m \neq$

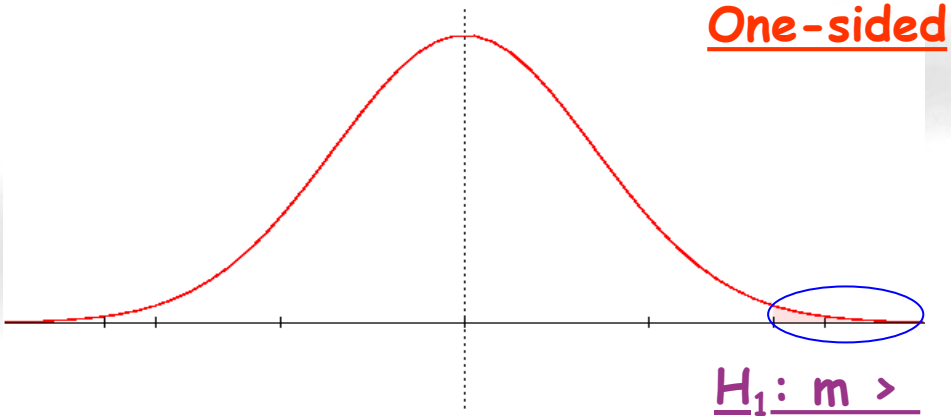


One-sided



$H_1: m <$

One-sided



$H_1: m >$

Example

Null hypothesis (H_0):

H_0: The mean of BUA values is 70.0

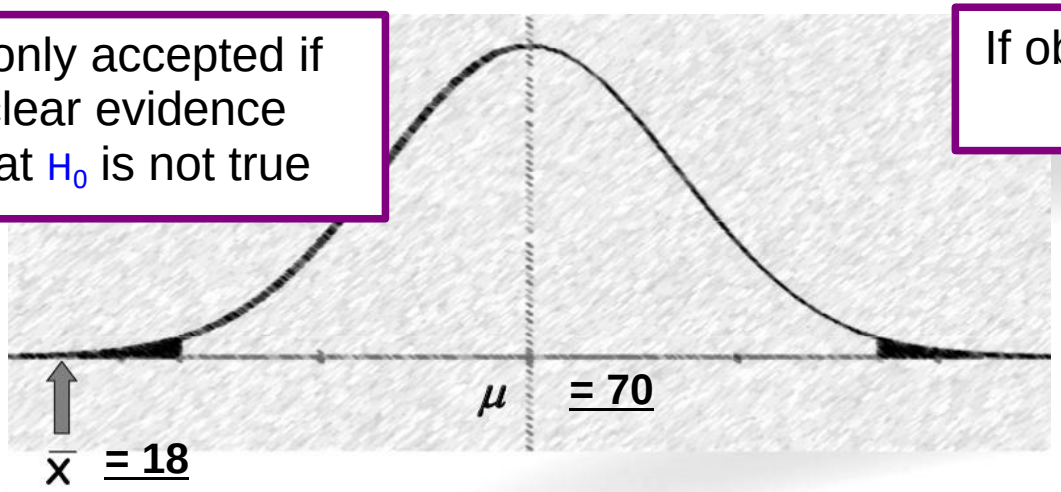
Alternative hypothesis ($H_\alpha = H_a = H_1$): the opposite idea

- H1: The mean of the bua values is not equal to 70.0 (Bilateral)
- H1: The mean of the bua values is higher(lower) than 70.0 (Unilateral)

Under the null hypothesis, *if all the samples of a given size could be selected and their sample means could be computed* the sampling distribuion would be:

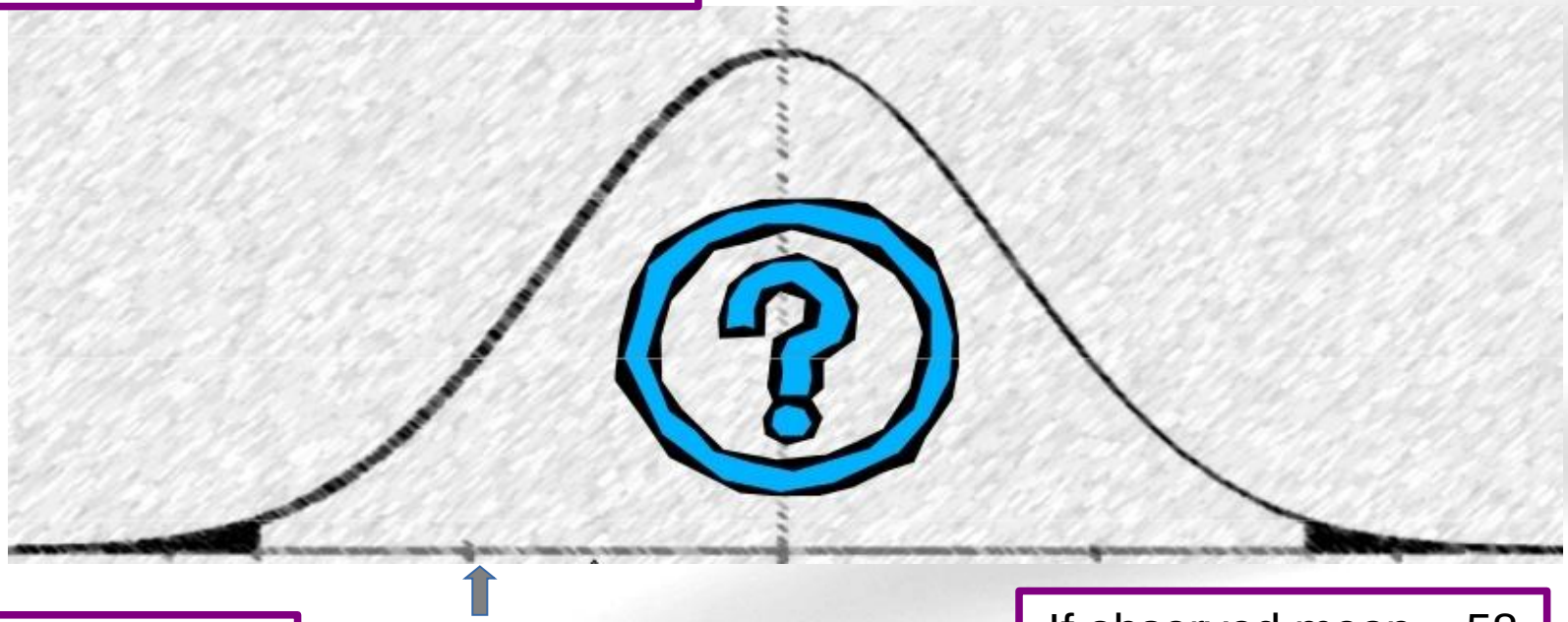
H_1 only accepted if clear evidence that H_0 is not true

If observed mean = 18
H_0 is unlikely



Accepting or rejecting the NULL

H_1 only accepted if clear evidence that H_0 is not true



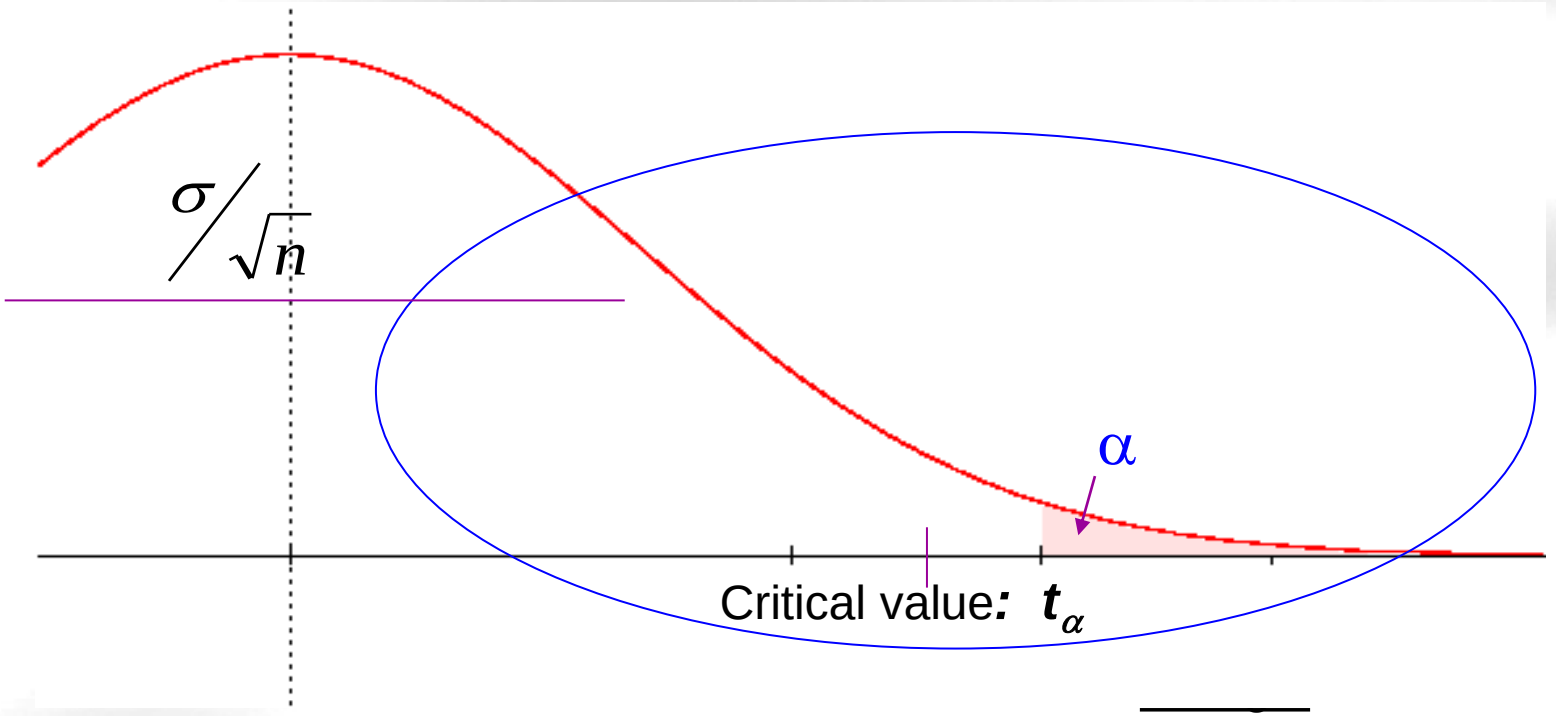
If observed mean = 18
H_0 is rejected

$\mu = 70$

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

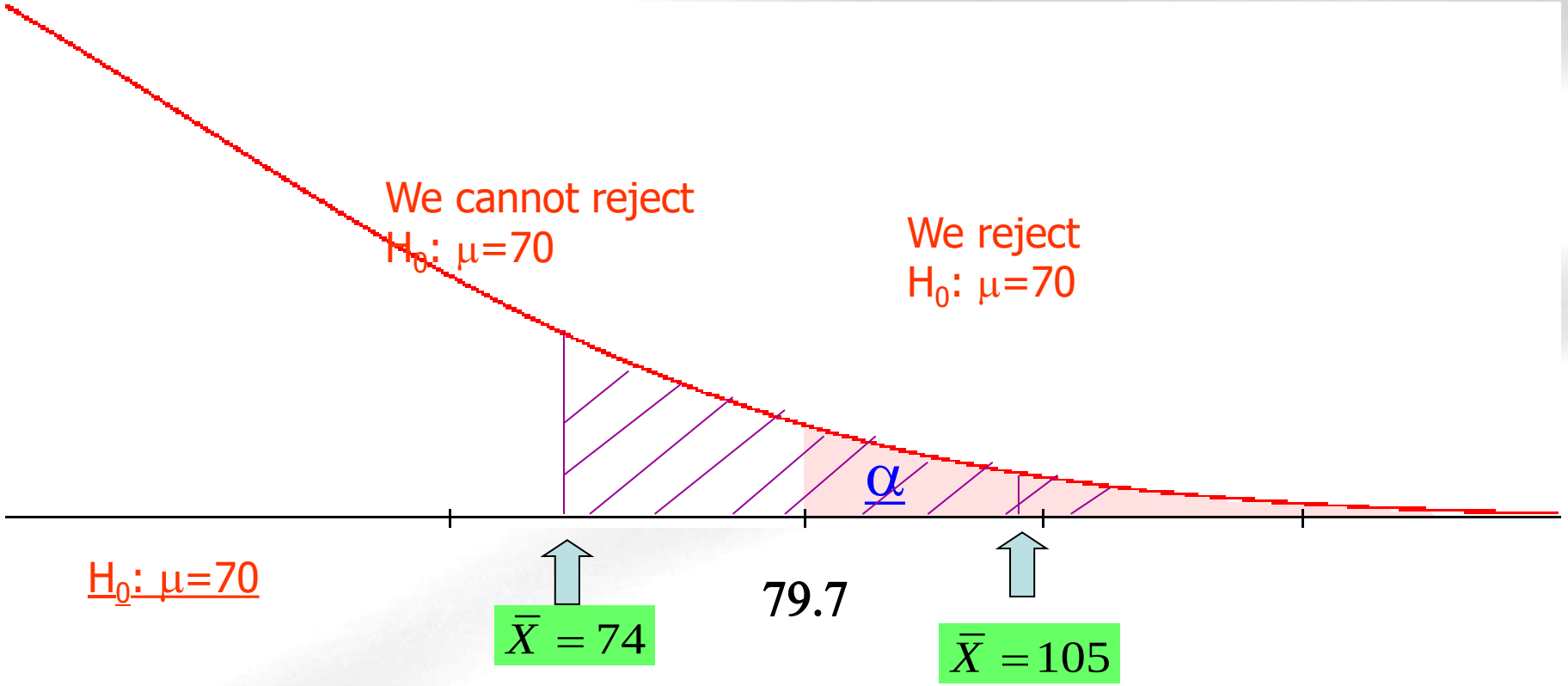
Critical Value

- At which value of the sample mean does one change from non-rejecting to rejecting the null hypothesis?
  - A value is selected such that the probability that the sample mean exceeds it, if the null hypothesis is true, is “small”, (for example 5%).
  - This value is called “Critical Value” t_α and
  - the probability is called “significance level (α)” and it is set to be small.



Example: Critical value and Sample mean

- *If σ=17, n =25 and α=0.05 the critical value is 79.7*
  - With a sample mean of 74 we will not reject H_0
  - With a sample mean of 83 we will t reject H_0



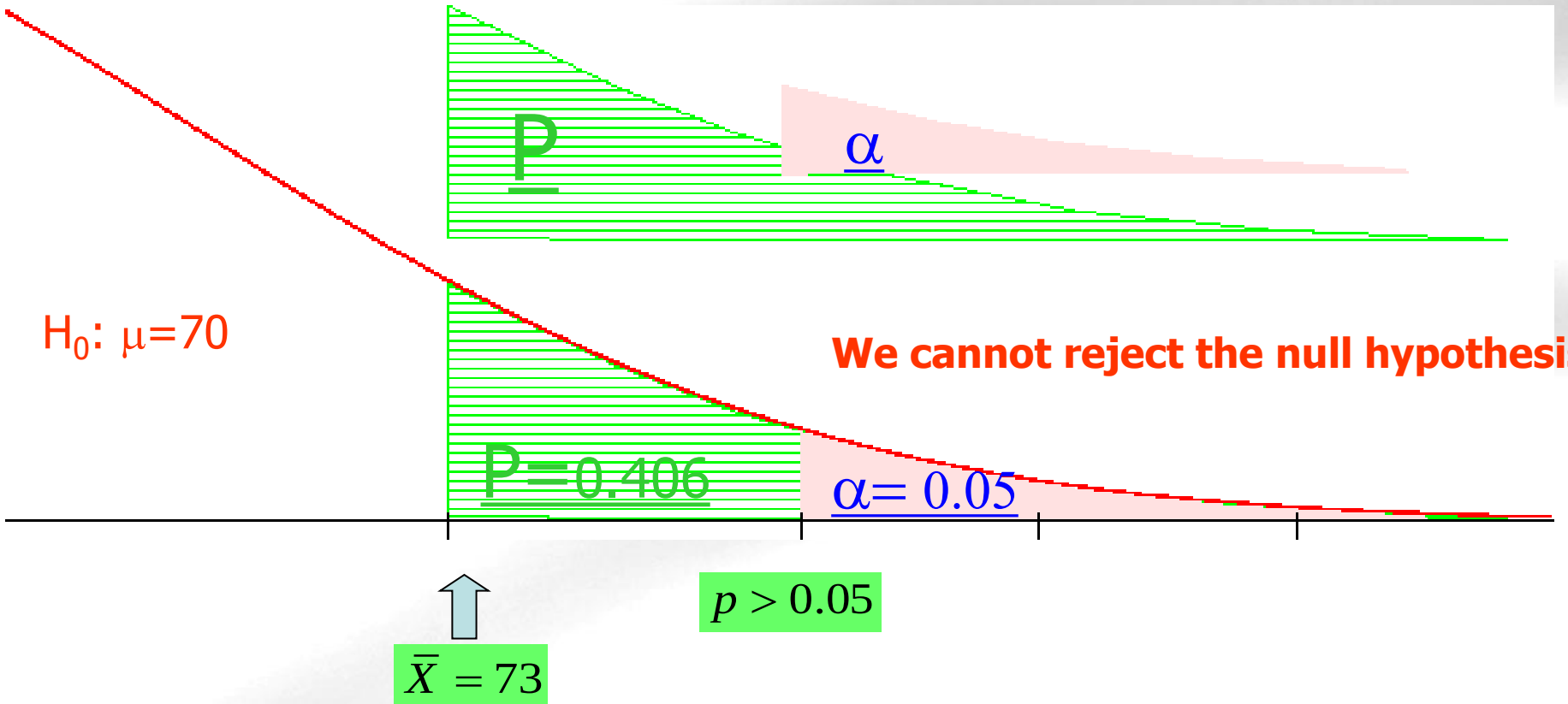
P values: The alternative

- We have based our decision about rejecting H_0 on comparing sample mean (i.e. 73) with the critical value (i.e. 79.7)
- Instead we can compare the probability of observing at least that sample mean (p value) with the significance level (α) (which is the probability of observing at least the critical value),
  - The probability is smaller than alfa if (and only if) the sample mean is bigger than the critical value.
    - In such situation we decide to reject H0
  - The probabiliy is bigger than alfa if (and only if) the sample mean is smaller than the critical value.
    - In such situation we cannot reject H0 so we accept it
- Both criteria (critical value and p-value) are valid for testing hypotheses.

Example: P-value vs critical value

- ***If σ=17, n =25 and the sample mean is 73 then***
- The probability that assuming that H0 is true, that is m=70, we can observe by chance a sample greater than 73 is: 0.406

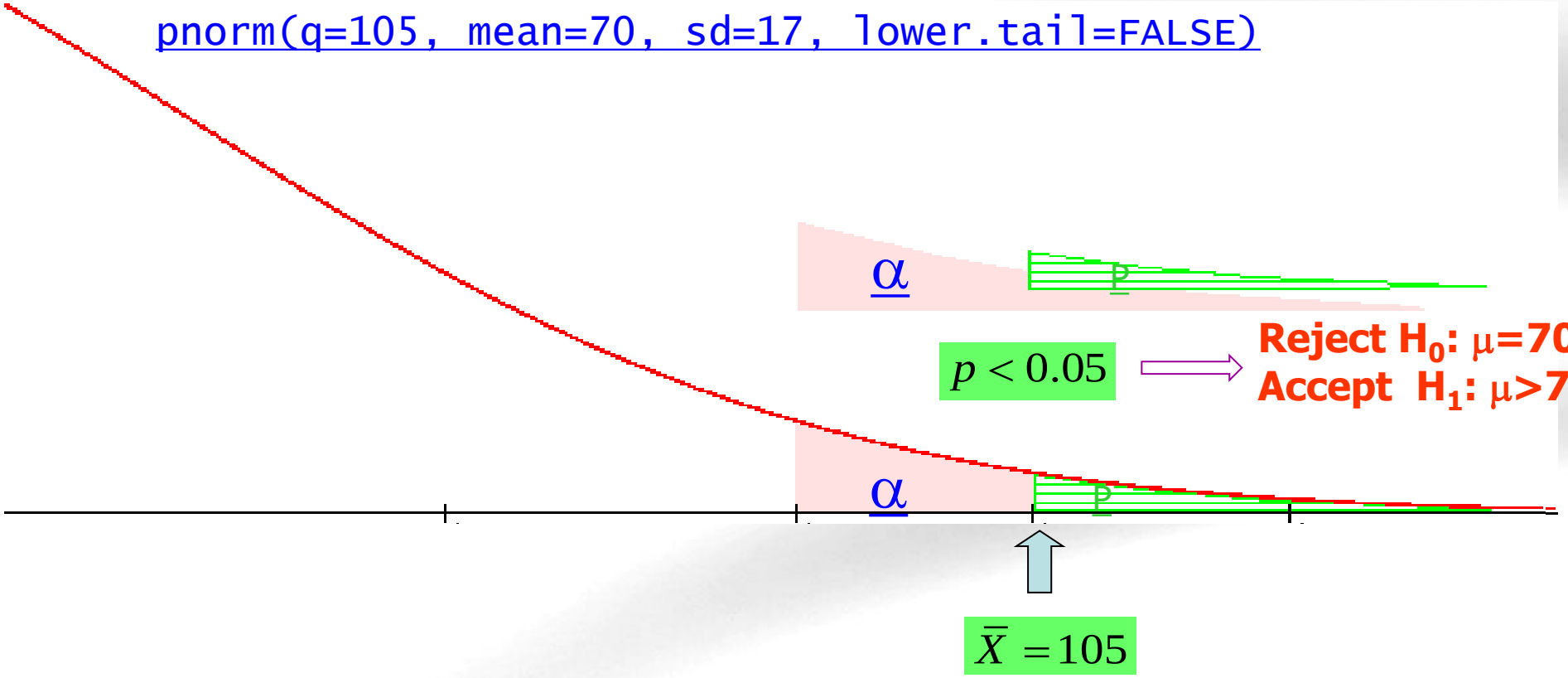
```
pnorm(q=74, mean=70, sd=17, lower.tail=FALSE)
```



Example: P-value vs critical value

- ***If σ=17, n =25 and the sample mean is 105 then***
- The probability that assuming that H_0 is true, that is μ=70, it can be obtained by chance a sample with a mean greater than 70 is: 0.019

```
pnorm(q=105, mean=70, sd=17, lower.tail=FALSE)
```



We usually say the test is statistically significant if $p < \alpha$

Summary: α vs p

α and P are related but they are not the same …

- About α
  - It is prefixed before experiment
  - Usually low (0.05)
  - Linked with critical value ("knowing one, the other is automatically known)
  - Unaffected by the sampling process.
- About p
  - It is calculated after the experiment
  - Can take any values in (0,1)
  - After calculation one can know the *achieved significance level*.
  - Depends on the sampling process

Type of Hypothesis

Confirmation Hypothesis

Aim is to confirm hypothesis about parameters or distributions.

Goodness of fit test to verify hipothesis about the distribution of variable in population

- Does populational blood pressure adjust to a normal distribution?

Test to verify values about a parameter.

- Is the average "bua" value in our population equal to 70?
- Is the proportion of lung cancer cases equal to 2.6%?

Type of Hypothesis

Independence Hypothesis

Aim is to test hypothesis for relation of variables in a population or no differences of a variable in two or more populations

- Is the average "bua" value the same in menopausic and in no menopausic population?
- Is the proportion of lung cancer cases the same in people with high or low fruit consumption?
- Is CD4 lymphocytes count related with CD8 count in HIV positive?

Type of Test

Parametric Test

It is assumed that the variable under study follows a particular distribution and values about its parameters are tested

- Distribution of proportion of lung cancer is binomial

$$H_0: p = 3\%$$

- BUA is the same in menopausic and non menopausic and variable is normal or symmetric

$$H_0: \mu_{Menopausic} = \mu_{Non\ \ menopausic}$$

- Distribution is binomial and proportion of lung cancer is the same in high and low fruit consummers

$$H_0: p_{High\ fruit} = p_{Low\ \ fruit}$$

Type of Test

Non Parametric Test

No distribution is assumed and test are related to distribution not to values about parameters

- Distribution of bua follow a normal distribution
- Bua is the same in menopausic and non menopausic and variable is normal or symmetric

H_0: distribution in Menopausic= Distribution in non menopausic

- Lung cancer is not related to fruit consumption. They are independent.

Hypothesis testing examples

- Imagine we have the following beliefs (our reference documents state that this is "TRUE")
  1. *The average ''bua'' value in our population is 70.*
  2. *The ''bua'' mean value in menopausic and non-menopausic women is not the same.*
- These beliefs can be checked through the corresponding hypothesis tests
  - H_0: $\mu_{bua} = 70$, H_1: $\mu_{bua} <> 70$,
  - H_0: $\mu_{[bua,\ Menop]} = \mu_{[bua,\ NoMenop]}$, H_1: $\mu_{[bua,\ Menop]} <> \mu_{[bua,\ NoMenop]}$

The average "bua" value in our population is 70

```
# osteoAll is the dataset obtained reading osteoporosis.csv


t.test(osteoData$bua,  alternative="less")

        One Sample t-test

data:  osteoData$bua
t = 23.161, df = 24, p-value = 1
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
    -Inf 79.6381
sample estimates:
mean of x
    74.16
```

Haga clic para agregar texto

The average "bua" is the same in "MENOP" and "Not MENOP"

```
t.test (bua~menop, alternative="two.sided", data=osteoData))


Welch Two Sample t-test

data: bua by menop t = 1.797,
df = 90.184, p-value = 0.07568
alternative hypothesis: true
difference in means is not equal to 0

95 percent confidence interval:
-0.5399937 10.7761048

sample estimates:
mean in group NO     mean in group SI
     76.05556           70.93750
```

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 after testing

		True	
		Innocent	Guilty
verdict	Innocent	OK	Error
	Guilty	Error	OK

Errors and Right Decisions

	Null Hypothesis True	**Null Hypothesis False**
Test *does not reject null* hypothesis	**Right decision (*1- α*)**	**Type II Error β**
Test *rejects* null hypothesis	**Type I Error α**	**Right decision Power (*1- β*)**

Power and Sample Size

- In an ideal situation one might want to control for both probabilities of error.
  - Select a test such that, for example:
    P(Type I err) < 0.05 **AND**
    P(Type II err) < 0.3
- In practice it is usually not possible and *decreasing the probability of one error type increases the probability of the other*.

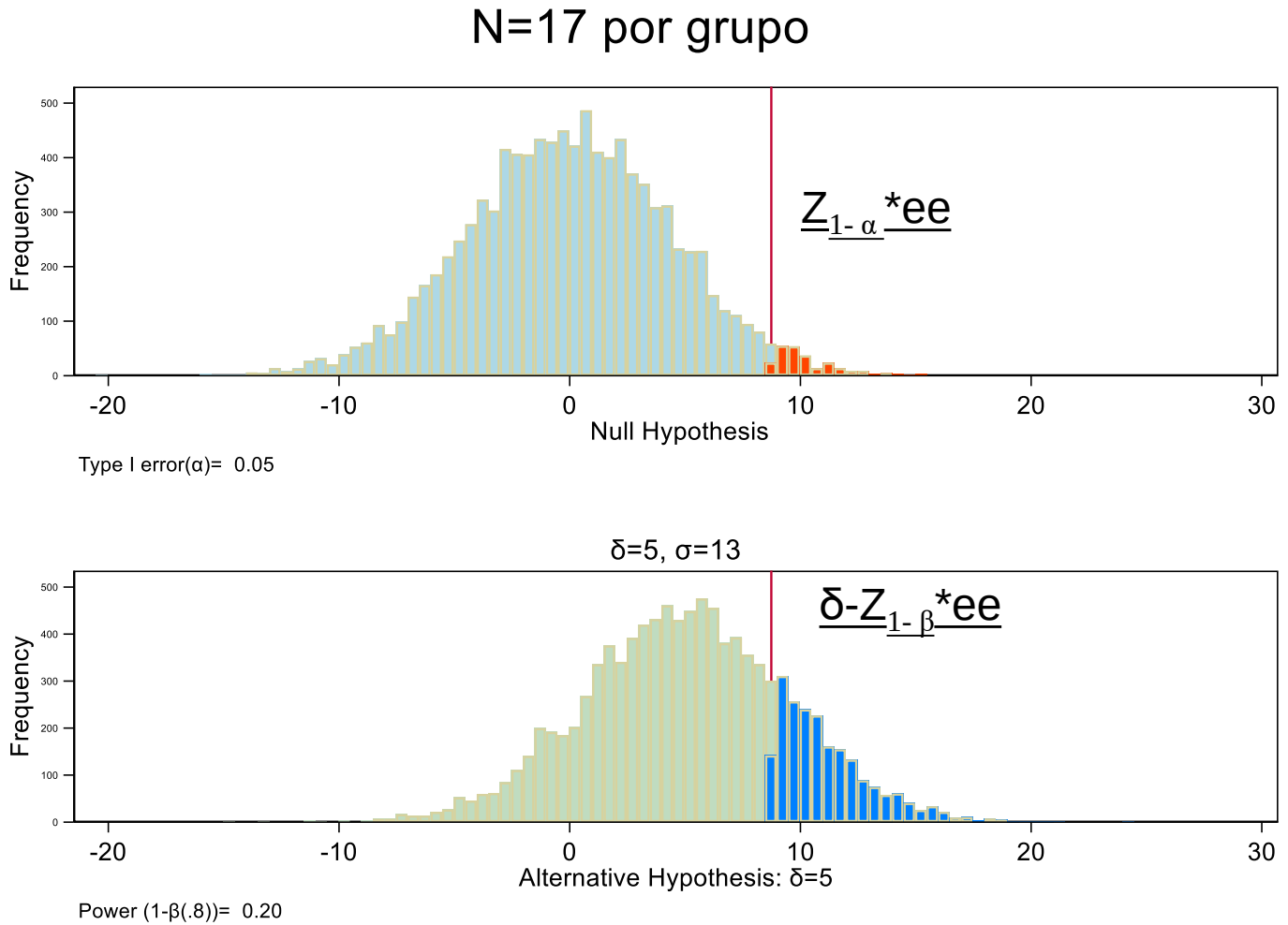
Power and Sample Size (II)

- In practice, there are only two ways yo increase power in a test
  - Increase sample size
  - Change the test so that, the effect to be detected, is bigger.
- This can be reversed and in practice we aim at computing the sample size reuired to attain a certain power given a desired effect size to be detected.

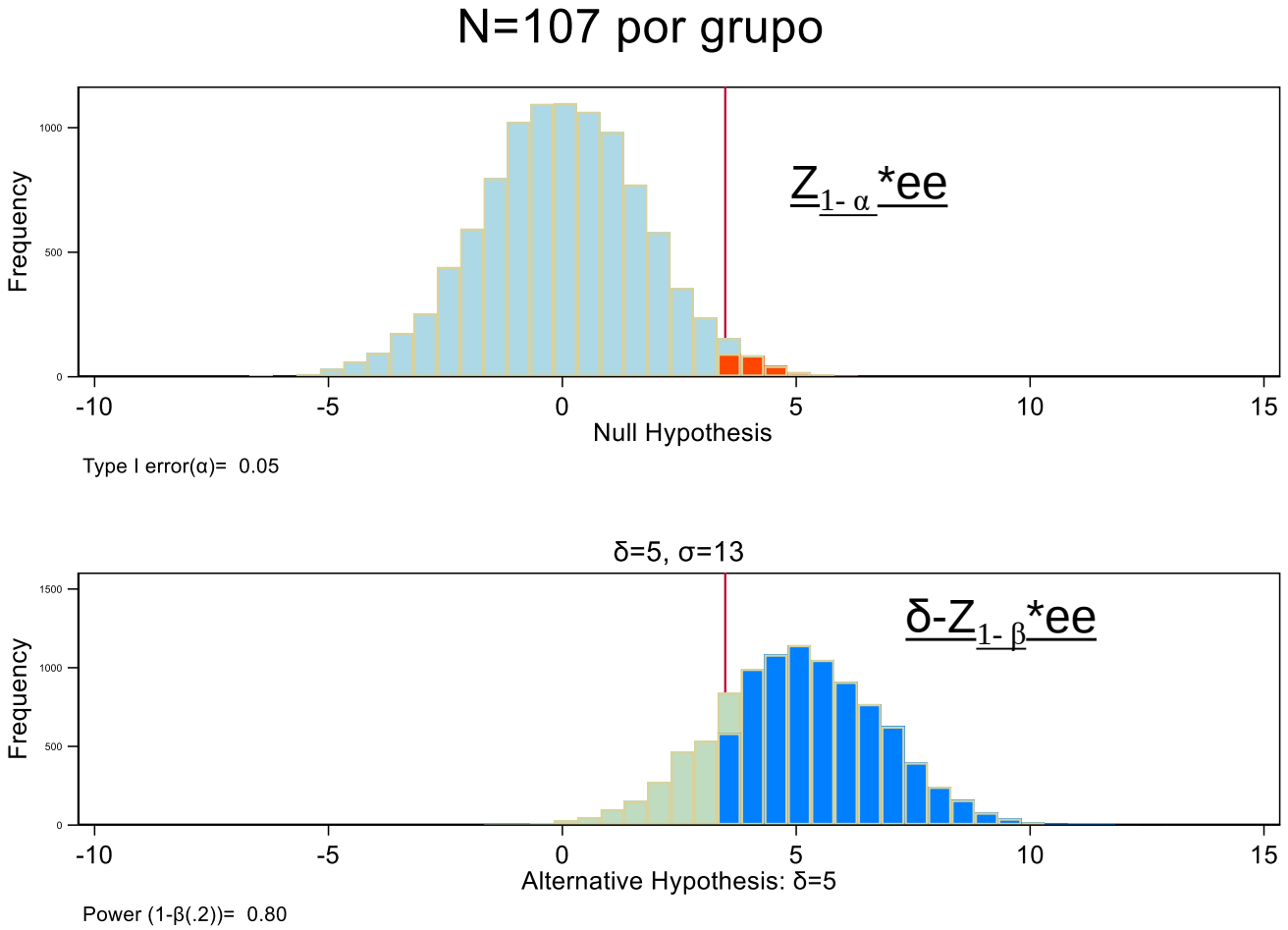
The factors affecting sample size

Type I Error α	Probability of rejecting the Null Hypothesis when its true (5%)
Power 1- β(Type II Error)	Probability of rejecting the Null Hypothesis when its false (80%,90%)
Variability of the data σ^2	Variance of the data
Effect Size δ	Minimum detectable difference betweeen the two groups to compare

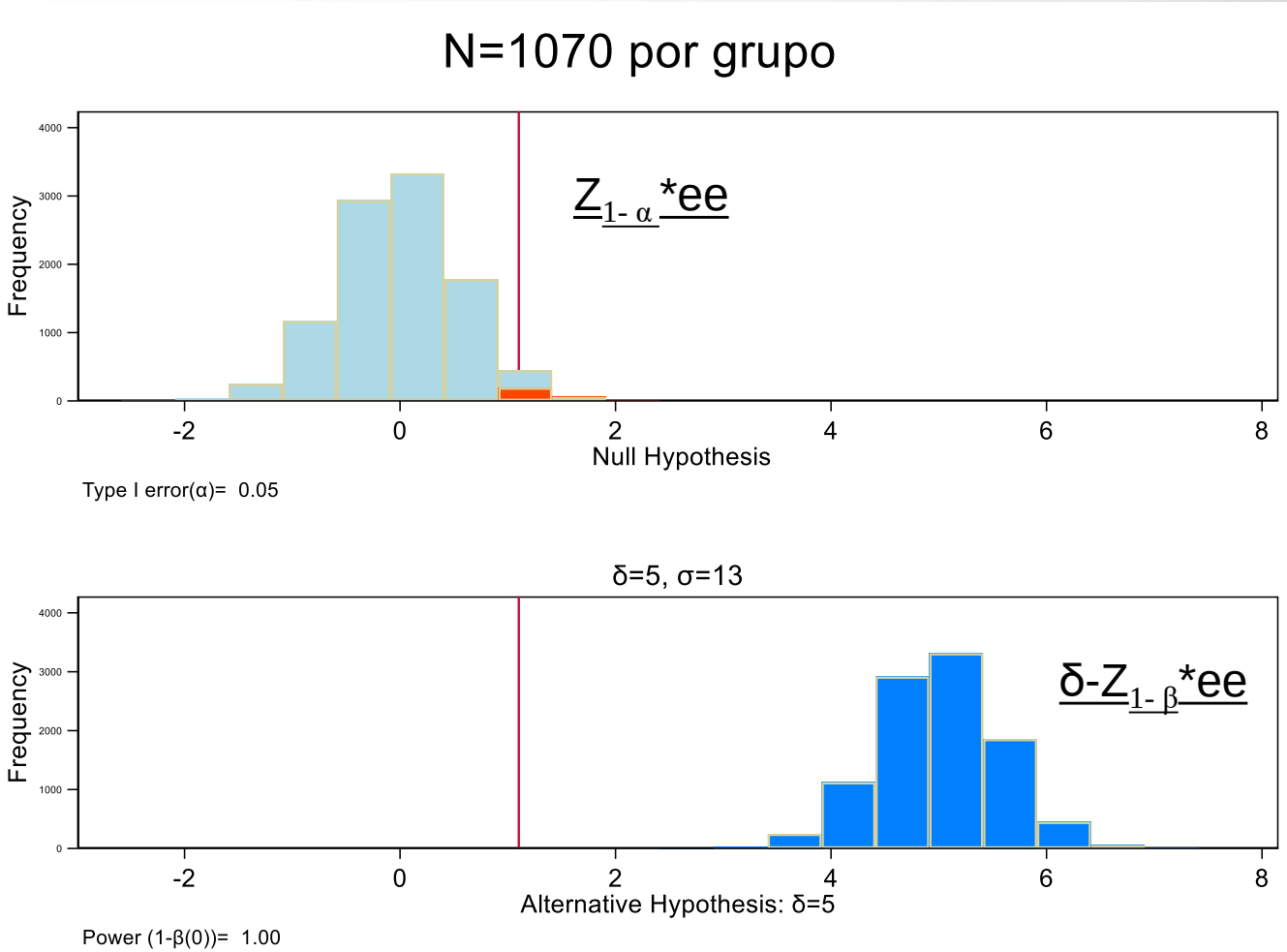
Effect size=5, Sample size = 17, Power = 0.20



Effect size=5, Sample size = 107, Power = 0.80



Effect size=5, Sample size = 1070, Power >0.999



An example sample size formula

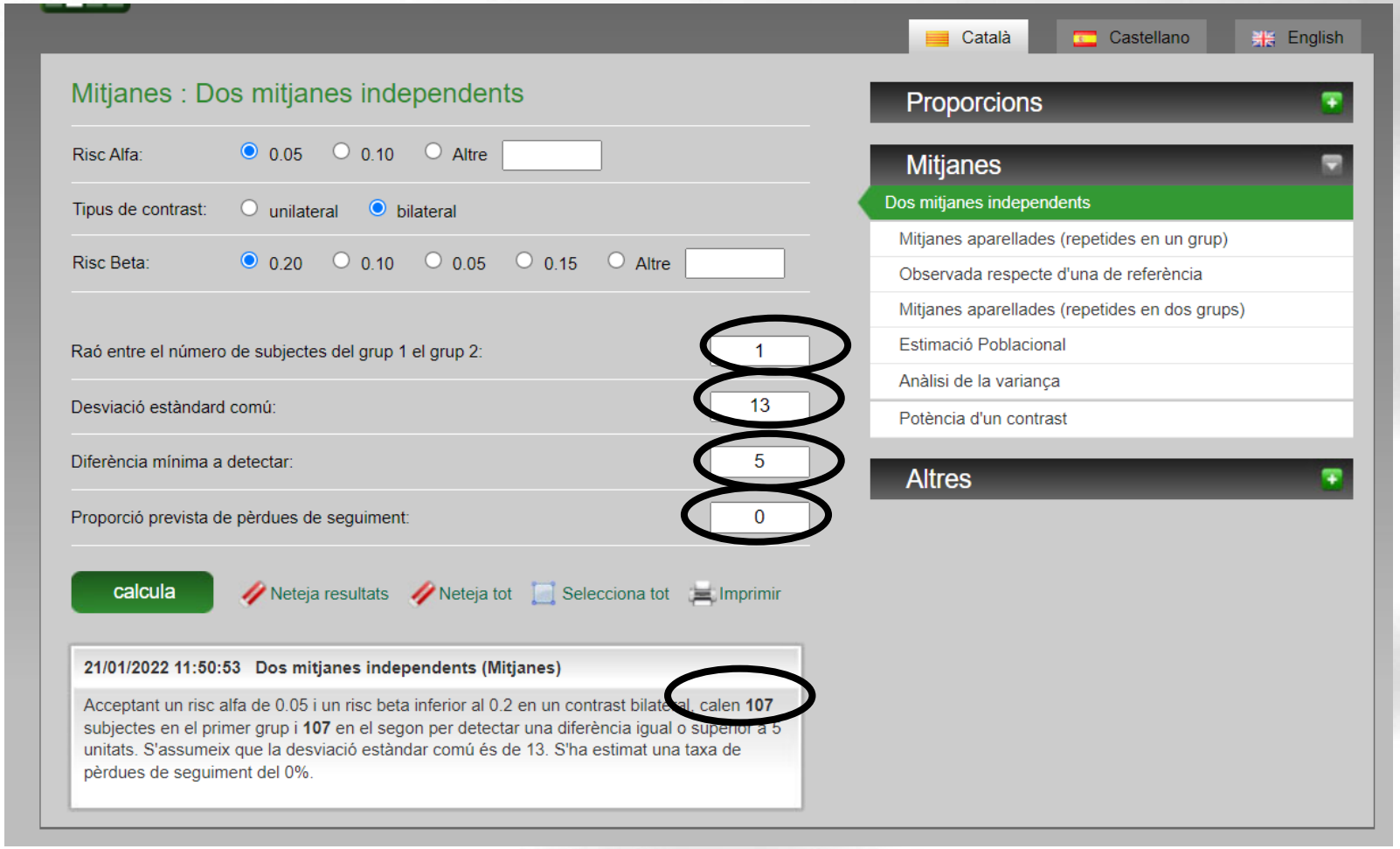
We want

$$Z_{1-\alpha} * ee = \delta - Z_{1-\beta} * ee$$

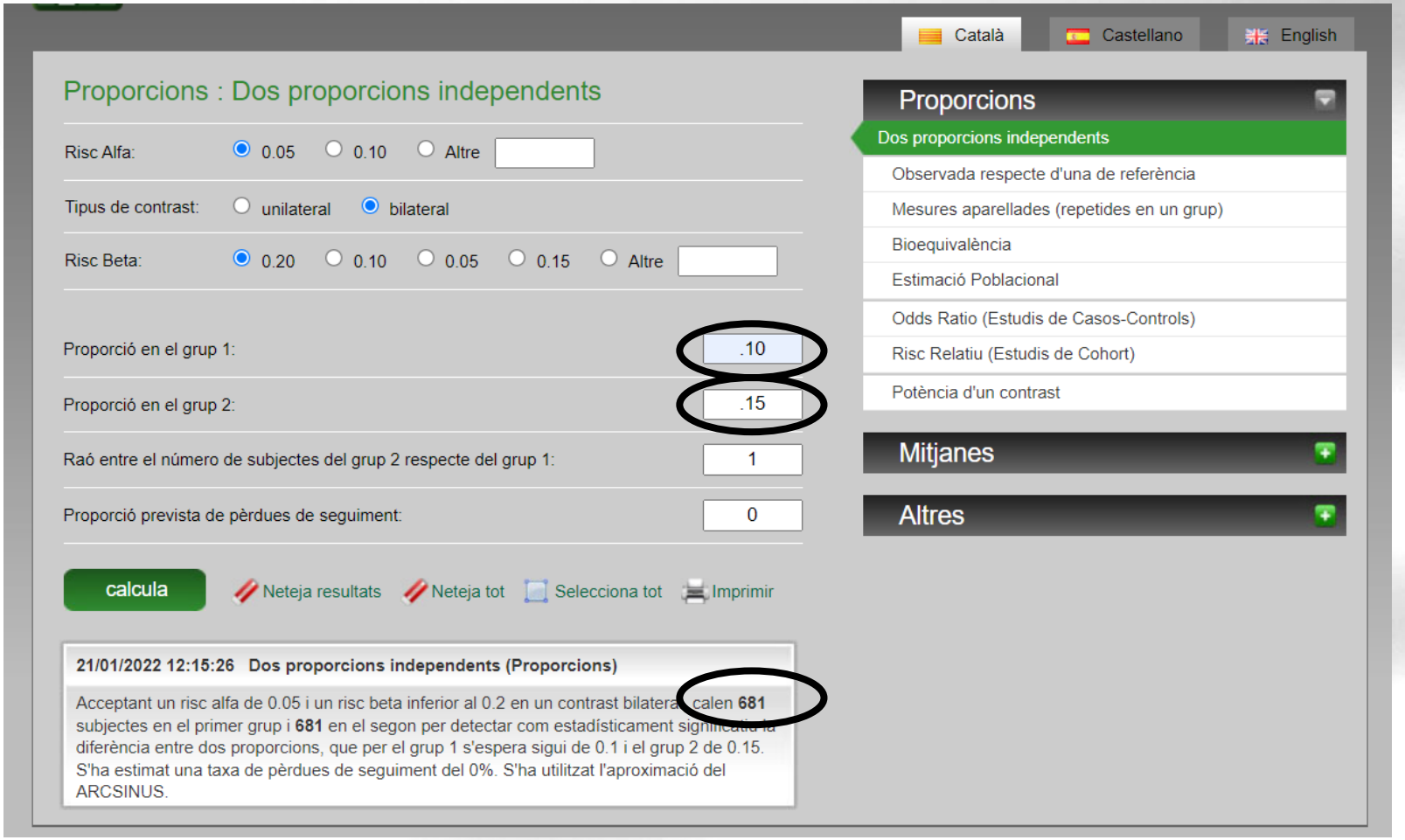
We know α, β and $ee = \sigma/\sqrt{n}$

$$n = \frac{2\sigma^2 (z_{1-\alpha} + z_{1-\beta})^2}{\delta^2}$$

Sample size for mean differences



Sample size for differences in proportions



Some tips for sample size calcuations

- Sample size goes up
  - for smaller α
  - For higher β
  - For smaller δ
  - For higher σ
  - For p closer to 50%
- Sample size is higher for proportions than means
- Sample size must be calculated a priori. Is not sensible to calculate power after
- SD can be calculated from 95% CI
- Upper-Lower limit of a CI is about 4 Standard Error and $SE=s/\sqrt{n}$
- Some % of survivors can be obtained from Kaplan-Meier survival curves and can be used for calculations
- Sample size is not an exact science and must be the product of calculations and reality

Common misunderstandings about the p-value

Common misunderstandings about the p-value

- The p-value is **not** the probability that the null hypothesis is true, nor it is the probability that the alternative hypothesis is false (it is not connected to either of these).
- The p-value **cannot** be used to figure out the probability of a hypothesis being true.
- The p-value is **not** the probability of wrongly rejecting the null hypothesis.
- The p-value is **not** the probability that replicating the experiment would yield the same conclusion.
- The p-value does **not** indicate the size or importance of the observed effect. The two do vary together however: the larger the effect (effect size), the smaller sample size will be required to get a significant p-value.

Multiple Comparisons and múltiple testing

Testing hypothesis repeatedly

- Every time we do a test there is a chance to take the wrong decision by rejecting the null hypothesis while it is TRUE.
- If, instead, we do many tests simultaneously the probability that there is, by chance, at least one false positive increases and does not match the type I error probability anymore.
- This increase in the probability of type I error has to be compensated in some way → **multiple testing adjustments**

To cross or not to cross?

The previous situation can be better understood with the “bridge analogy”.

Imagine you are an adventurer that has the option of to cross a bridge in order to escape from danger, find a treasure...

and that there is a post in front of the bridge stating:

“This bridge has broken only one out of 100 times"

To cross or not to cross?

Imagine you are an adventurer that has the option of to cross a bridge in order to escape from danger, find a treasure...

and that there is a post in front of the bridge stating:

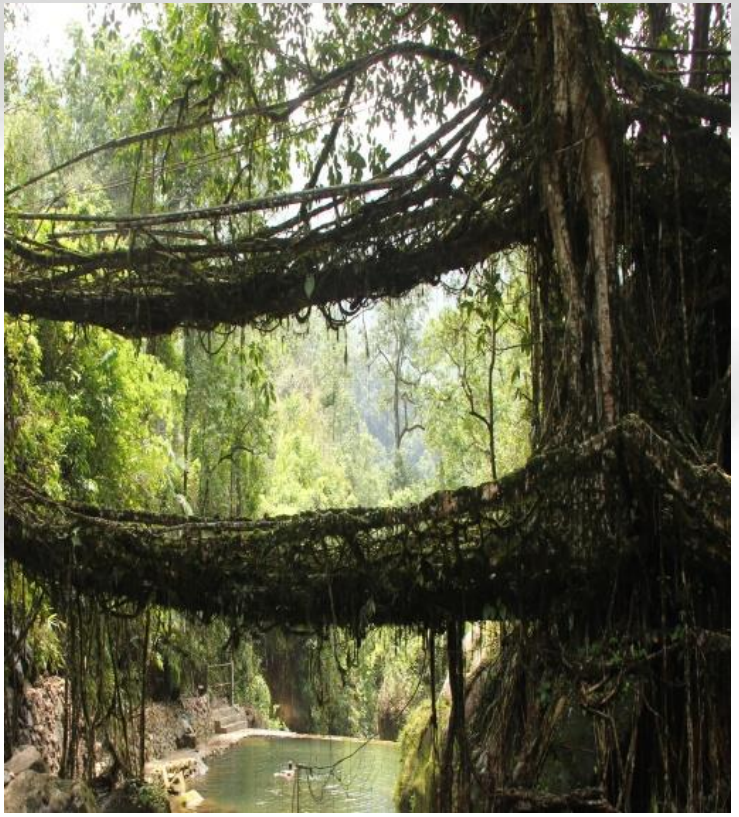
"This bridge has broken only one out of 100 times"

So, the p-value of our metaphor is 0.01

You could accept that **1%** is **a risk small** enough to **pass the bridge** and pursue your goal. OK

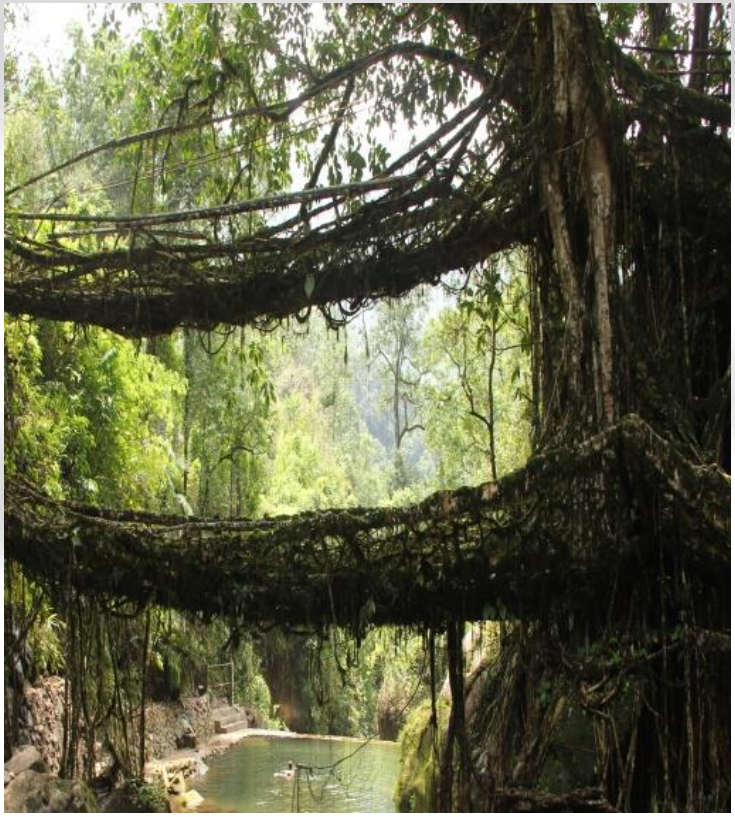
To cross or not to cross?

But... what do you decide if, in order to reach your goal, you have to cross hundreds of bridges of that kind?



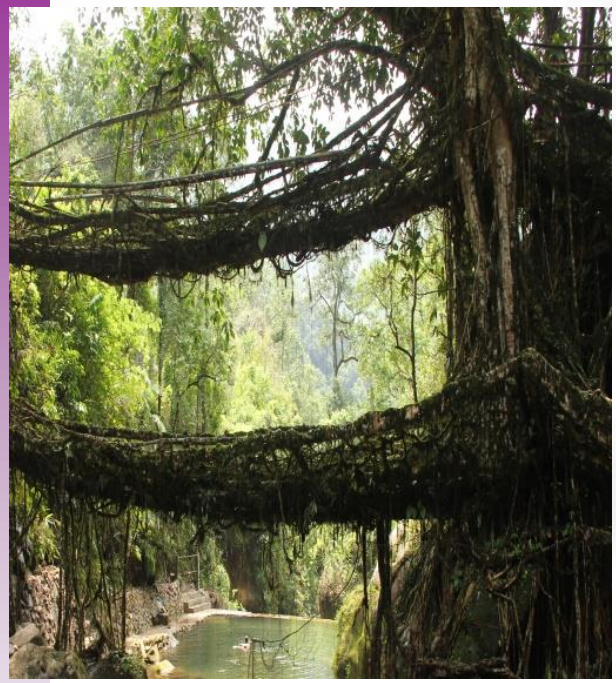
To cross or not to cross?

But... what do you decide if, in order to reach your goal, you have to cross hundreds of bridges of that kind?



•In this case, the probability of falling while crossing one of the bridges is obviously too high ('cause we have just one life).

To cross or not to cross?



Therefore, in this case (multiple testing), the p-value by itself is not a good reference for accepting or not statistical significance.

We must apply some type of adjustment to the p-values (allowing us to be safe in crossing all the bridges).

Some p-value adjustments

- Bonferroni (α/k)
- Post-Hoc test ANOVA (Tukey, Scheffe, Dunn-test)
- False Discovery rate
- Benjamini-Hochberg correction

Multiple comparisons vs multiple testing

- There are two distinct situations where p-value adjustment may be necessary:
  - Post-hoc tests in ANOVA:
    - This is usually called multiple comparisons and common methods of adjustment are Tukey, Fisher HSD.
  - Testing many variables in the same study
    - This is usually called multiple testing and common methods of adjustment are Bonferroni, Holm or Benjamini and Hochberg (False Discovery Rate).

Multiple testing

- When many variables are compared independently with the same test
  - Find differences between treated/untreated for a set of biomarkers such as cytokines.
    - Number of comparisons may be low ("dozens")
  - Find differentially expressed genes, i.e. genes whose expression may change between conditions.
    - Number of comparisons high ("hundreds" to "thousands")
- This is usually called multiple testing and common methods are Bonferroni, Holm or Benjamini and Hochberg (False Discovery Rate).

Post-hoc ANOVA tests

- If we wish to compare all means against all means he number of tests increases quickly (to compare all pairs of means if there are k groups (k*k-1)/2 tests are required).
- This is usually called **multiple comparisons** and common methods of adjustment are Tukey, Fisher HSD or Bonferroni.