Statistical Inference

Hypothesis testing

We make a hypothesis H_0 about a distribution, and must decide if it is true (= we ACCEPT it) or false (= we REJECT it). For this, we collect a sample of data from the distribution, and measure the discrepancy between the data and H_0 by means of a **statistic** T. There are 4 possible scenarios:

	H_0 **true**	H_0 **false**
We ACCEPT H_0	OK	type 2 error
We REJECT H_0	type 1 error	OK

It's worse to make a type 1 error, so we bound the probability of committing a type 1 error with a **significance level** α . Usually, α will be small, at most 0.1 and usually it will be 0.05.
Out of all the decision rules for which P(type 1 error) $\leq \alpha$, we take the one minimizing P(type 2 error). We divide the set of samples in two groups:

- Critical Region (C. R.) $\rightarrow$ their discrepancy with H_0 is large, so we REJECT H_0
- Region of Acceptance (R. A.) $\rightarrow$ their discrepancy with H_0 is small, so we ACCEPT H_0

How to tell if a sample is in the critical region or not? (discrepancy is large or not):
To determine the boundary between the two regions, we use that P(being in the Critical Region | H_0 is true) = α

We can solve the hypothesis testing by means of the critical region or using the **p-value**

Procedure 1

1 - We fix the hypothesis H_0 and the significance level α
2 - We determine the Critical Region and the Region of Acceptance
3 - We collect the data
4 - If the dataset belongs to the Critical Region, we REJECT H_0. If the dataset belongs to the Region of Acceptance, we ACCEPT H_0

If we increase α, the Critical Region increases. We have two extreme cases:

- If α = 0, the Critical Region is empty. Therefore, we'll never reject
- If α = 1, the Region of Acceptance is empty and the Critical Region includes all the set. Therefore, we'll never accept

For a given dataset, there is a smallest significance level α for which the associated critical region includes the dataset.

Procedure 2

1 - We fix the hypothesis H_0 and the significance level
2 - We collect the data
3 - Apply a formula that gives the p-value
4 - Compare the p-value with the significance level

If p-value is < than the significance level we REJECT H_0, and if it's greater or equal we ACCEPT H_0.

Observations

1 - The two procedures are equivalent (= lead to the same decision)
2 - The only thing that will change from one hypothesis to another will be the formula of the statistic and its distribution when H_0 is true
3 - If they don't give me a value of α, use α = 0.05
4 - In some cases we will only tell which is the test to apply, but in four cases:

- the single sample **t** test
- the single sample **proportion** test
- the single sample **variance** test
- the **chi-square** test of independence

We will see the formula of the statistic and its distribution.

> **Important**
>
> One of those four test will be in the exam

Exercises

Ex 6

n = 50; 1-α = 0.95; Confidence interval for μ: [37, 43]

a) FALSE the confidence interval at 95% uses a formula that gives a "good" interval (it includes the parameter) 95% of times, but there's a chance that it does not

b) FALSE for the confidence interval being greater, that means that we have a bigger range so we leave less room for errors. The confidence interval range for 99% cannot be smaller than the one for 95%

c) TRUE with different data we can get different confidence intervals

Ex 7

If we are using the same data, the confidence interval at 99% should be wider than the one at 95%. Therefore, the first interval is at 99% [0.64, 0.84] and the second one is at 95% [0.67, 0.82].