Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2023/2024

Theoretical-Practical Form

Parametric Hypothesis Tests

Goals:

1. Familiarization with the R tool in support of parametric hypothesis tests.
2. Analysis and discussion of results.

Proposed Exercises

1. 15 portable computers of a given brand were randomly chosen, and the following measurements were obtained for their thicknesses (in *mm*):

30, 30, 30, 30, 31, 32, 32, 32, 32, 33, 33, 34, 34, 34, 35

Considering that the computer thickness is a normal random variable, test the hypothesis H_0: $\mu=32.5$ against H_1:$\mu\neq32.5$ (admit that the significance level is 5%, $\alpha=0.05$).

2. An online store indicated on its website that the delivery is carried out, on average, in 5 days. A regular customer made a complaint, stating that the average delivery time was higher than the amount indicated by the store. To find out if the customer is right, 36 purchases made on their website were randomly analyzed, and the delivery times were recorded, in days, below:

5	4	4	5	5	5	6	5	4	4	3	4	4	5	5	7	6	5
6	4	6	5	5	6	6	6	4	4	5	5	5	3	6	3	6	5

Considering a significance level of 1%:

a) Formulate the null hypothesis and its alternative hypothesis for the problem. Is this a one-sided or bilateral test? Indicate the observed value of the test statistic.
b) What can be concluded about the customer's complaint?
c) Indicate the observed value of the test statistic.

3. The maximum velocities (in dpi) of 24 computer mouses of a given brand were recorded, divided into 2 groups: **wired and wireless**. The results were:

Com fio	**2300**	**2000**	**2800**	**2000**	**2400**	**2200**	**2000**	**1800**	**1900**	**2100**	**2200**	**2400**
Sem fio	**2400**	**2200**	**1800**	**1900**	**1800**	**1900**	**2100**	**2050**	**2200**	**2000**	**1900**	**2000**

a) Indicate whether the samples are independent or paired.

b) For a significance level of 1%, assuming that speed is a normal random variable, test the hypothesis of equality of average velocities in wired and wireless rats.

4. A compression rupture test carried out on 12 cubic concrete specimens resulted in the following values of the rupture stress (kgf / cm^2).

263 254 261 236 228 253 249 262 250 252 257 258

Admit that the variable under study follows a normal distribution.

a) An engineer wants to know if the expected breakage voltage is not less than 255 kgf/cm^2. What evidence does the data on this issue provide if a significance level of less than or equal to 5 is admitted?

b) Knowing that the characteristic value of the rupture stress is defined as the value of the variable that has a probability of 95 being exceeded, calculate an estimate of the characteristic value of the concrete rupture stress under study.

5. To test a new medicine to treat attention deficit hyperactivity disorder (PDAH) a random sample of 24 children with PDAH was chosen. Each child was medicated with the new medicine (D60) and placebo (DO). After each treatment, each child answered 60 questions. The number of correct answers are stored in **the Data_C.csv**. Assuming that the number of correct answers is a variable with normal distribution, are there reasons to assume that the new treatment increases the average number of correct answers? Use a significance index of 1%.

6. Two groups of 36 students were randomly selected to participate in an experiment that consists of learning the meaning of words in a language they do not know. For 30 minutes the students tried to learn the greatest number of words. In group I the students worked in isolation. In group II the students worked in pairs to make sure they were learning the words. A test was then performed to determine the number of words learned by each student. The results can be found in **Data_A.csv**. Check whether the second learning method can be considered significantly better than the first method of learning.

7. Tests were carried out on 10 and 8 random samples of 1kg crude oil from holes in two fields to determine the amount (in grams) of sulfur. The obtained results are indicated in the following table:

Field A:	111	114	105	112	107	109	112	110	110	106
Field B:	109	103	101	105	106	108	110	104		

Assuming that the amount of sulphur is a variable with normal distribution, you can claim that, on average, the amount of sulphur per kilogram of field A is:

a) higher than field B? **b)** lower than that of field B?

8. The director of a hotel decided to invest in some works during 2014, with the objective of modernizing the enterprise, improving the quality of its equipment, and thus attracting a greater number of customers. To verify that this refurbishment had a

positive effect, monthly occupancy rates were recorded in %, 2013 (before construction) and 2015 (after construction). The data obtained are presented in the following table.

Month Average occupancy rates, in %	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Before works (2013)	20	35	40	55	60	75	95	100	90	80	45	25
After the works (2013)	25	30	45	75	80	100	100	100	100	85	65	30

Assuming that the sample data comes from a normal distribution:

a) Indicate whether the samples are independent or paired.
b) Formulate the null hypothesis and its alternative hypothesis for the problem. This is a one-sided or bilateral test?
c) Indicate the p-value and interpret (use a significance level α=0.05).
d) What can we conclude about the investment made?

9. One company has 4 suppliers of carbon fiber boards. The control division wants to find out if the quality of the plates is the same in all suppliers. For this purpose, 10 plates from each supplier were randomly selected, and tests were performed on the resistance of the plates. The test results can be found in the **Data_D.csv**. There are reasons to admit that there are significant differences in the quality of the plates in the 4 suppliers?

The following table shows the quotients between the final cost and the originally anticipated cost of R&D projects carried out in 4 large companies. Assuming the normality of the data, it is intended to investigate whether the factor "Company" has an effect on the worsening of projects. Consider a significance level (α=0.05).

Enterprise	Final Cost/Estimated Cost					
A	1.0	0.8	1.9	1.1	2.7	
B	1.7	2.5	3.0	2.2	3.7	1.9
C	1.0	1.3	3.2	1.4	1.3	2.0
D	3.8	2.8	1.9	3.0	2.5	

Consolidation Exercises

1. According to a 2019 study, Portugal is among the OECD countries with the highest percentage of low birth weight babies. In the **nutri.csv** are the data collected as part of a project with data on newborns in several maternity hospitals in the country in a given year, randomly selected from the total number of babies born that year. Check:

a) With a significance level of 0.01, if the average weight of newborns is lower than the average weight of newborns in 2011 which was 3400 g.

b) if the data agree with the hypothesis that girls (sex=2) are on average lighter than boys (sex=1) at birth.

2. As part of a project to investigate the impact of regular physical exercise on triglyceride levels, the concentration of triglycerides (mmol/L) in the blood serum of 34 male volunteers before and after participation in a 10-week physical exercise program was measured.

Using an appropriate test, check for statistical evidence at the level of 1%, based on the data in **the tri.csv** file, that regular exercise reduces the level of triglycerides in the blood.

3. In a clinical trial it is intended to study the way antiviral drug X is used to patients with influenza. Patients were randomly distributed into 3 groups: those who received inhaled X (group 1), those who received inhaled and intranasal X (group 2) and those who received placebo (group 3). The time interval (days) until relief of the main flu symptoms was measured and the data are found in the **antiviral file.csv**.

a) Establish the appropriate null hypothesis and alternative to the problem.
b) Represent the data graphically and test the null hypothesis with the test you consider appropriate.

4. For each of the following situations suppose that you are testing $H_0:\mu_1=\mu_2$ **vs** $H_1:\mu_1\neq\mu_2$. Indicate whether there is statistically significant evidence of H_1 in the following situations:

a) Value p = 0.078 for a significance level = 0.10. There is evidence of H_1 because $p < \alpha$
b) Value p = 0.055 for a significance level = 0.05. There is no evidence of H_1 because $p > \alpha$
c) t(observed) = 3.75 with 19 degrees of freedom for alfa=0.01