Applied Statistics

MSc Digital Marketing & Data Science
Mock Exam - Correction

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

Rules: there are two types of questions in this exam:

- Multiple choice questions but only one answer is possible.
- Questions where you have to provide the right figure in order to complete a table.

For MCQ: if a good answer if selected, you will have 1 point, if you do not answer to the question, you will zero point. In case you provide a wrong answer, you will have -0.5 points.

For the others questions, where you have to complete a table, there is no negative point.

Preliminaries

Financial market prices are subject to wide variations over time, and it is assumed that the values taken by the latter over time are distributed according to a centered normal distribution whose mean is equal to 7900 points. We know that 30% of the values taken by the CAC 40 price are below 7700 points and that 10% of the values taken are between 7850 and 7950. In what proportion does the share price take values between 7700 and 7850?

- ☐ 15%
- ☐ 20%
- ☐ 25%
- ☐ 30%

In a study of salaries in the finance sector, the organization found that the distribution of salaries followed a normal distribution, with a mean gross salary equal to 100,000 and a standard deviation equal to 15,000.

Approximately what proportion of salaries lie in the interval [70,000; 130,000]?

- ☐ 90%.
- ☐ 95%.
- ☐ 97.5%.
- ☐ 99%.

It is customary to use a linear model to try and estimate the profits generated by a company on the basis of various criteria, such as sales, investment in research, investment in marketing, etc. In such a situation, what is the dependent variable?

1. In such a situation, what is the dependent variable?
   - ☐ the generated profits.
   - ☐ investment in marketing.
   - ☐ investment in research.
   - ☐ sales.
2. What are the assumptions related to the linear models?

- ☐ the data are assumed to be independant and indentically distributed according to a normal distribution, the errors of the model are dependant to the others with different variances.
- ☐ the data are assumed to be indentically distributed according to a normal distribution, the errors of the model are dependant to the others with equal variance.
- ☐ the data are assumed to be independant and indentically distributed according to a normal distribution, the errors of the model are independant with equal variance.
- ☐ the data are assumed to be independant and identically distributed according to a normal distribution, the errors of the model are independant with unequal variance.

Hypothesis testing are mainly used to take decisions but also to lead to several conclusions of statiscal analysis, such as comparing the behavior. For this purpose, we usually introduce a coefficient α, called an *error rate* to conclude to our hypothesis testing.

1. Let us imagine that we want to test if two means μ_1 and μ_2 are equal or not using a two tailed test, what are the hypothesis H_0 and H_1 :

   - ☐ $H_0 : \mu_1 \neq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$.
   - ☐ $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 > \mu_2$
   - ☐ $H_0 : \mu_1 \neq \mu_2$ vs. $H_1 : \mu_1 = \mu_2$
   - ☐ $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

2. Assuming that we are performing a two tailed test. If we denote U a random variable and u_{test} the value of the statistical test. How is defined the p-value?

   - ☐ $2\mathbb{P}[U \leq u_{test}]$.
   - ☐ $2\mathbb{P}[U \geq |u_{test}|]$.
   - ☐ $\mathbb{P}[U \leq t]$.
   - ☐ $\mathbb{P}[U \geq t]$.

3. What is the meaning of this value α ?

   - ☐ This is a risk of the first kind. It represents the greatest risk we are prepared to take for the rejection of the null hypothesis.
   - ☐ This is a risk of the second kind. It represents the greatest risk we are prepared to take for the rejection of the null hypothesis.
   - ☐ This is a risk of the first kind. It represents the greatest risk we are prepared to take for the acceptation of the null hypothesis.
   - ☐ This is a risk of the second kind. It represents the greatest risk we are prepared to take for the acceptation of the null hypothesis.

A case study

A company specializing in finance is conducting an equal pay study. The human resources department and union representatives have been asked to conduct the study, which covers a number of points. To this end, a survey was carried out among a number of employees: workers and technicians, men and women in three different cities: Bordeaux, Marseille and Lyon. The study is made on two different years: 2020 and 2021 of data for a selected sample of employees. The study is conducted on a sample of 45 instances.
The test conducted in this part are two tail test using $\alpha = 0.025$.

1. First of all, we want to know what is the distribution of employee salaries. How can the team in charge of this study verify that this distribution is indeed a gaussian distribution?

   - ☐ We can only check if the distribution is symmetric.
   - ☐ It is enough to see if there is only positive values.
   - ☐ We can use a quantile-quantile or a normal probability plot to check this assumption.
   - ☐ We can check if the the mode is equal to the mean and equal to the median of the distribution.

2. Now, we want to to check if the wages have evolved between 2020 and 2021. What kind of test, the team shall perform for this purpose?

   - ☐ A t-test without anymore assumption.
   - ☐ A t-test with equal variance.
   - ☐ A paired t-test.
   - ☐ A t-test with unequal variance.

   The appropriate two tailed test leads to the following table of results

t_{test}	-6.8838
df	...
p-value	$1.69.10^{-8}$

3. What is the conclusion of this test?

   - ☐ We cannot conclude if the wages have evolved or not.
   - ☐ We can say that the wages have evolved but we do not know if they have increased or decreased.
   - ☐ We can say that the wages have evolved but we can say that they have increased.

☐ We can say that the wages have evolved but we can say they have decreased.

4. What is the number of degree of freedom related to this test.

   ☐ df = 42.
   ☐ df = 43.
   ☐ df = 44.
   ☐ df = 45.

5. To perform this test, what are the statistical quantities computed by the team?

   ☐ The mean of the two groups and the variance of the two groups
   ☐ The mean of the two groups and the pooled variance.
   ☐ The mean of the difference of the groups and the variance of the sample difference.
   ☐ The mean of the difference of the groups and the pooled variance.

Now we want to find out whether men and women in this company earn the same wages or not in 2021. To do this, we specify that the sample studied comprises 28 men and 17 women.
The conducted test leads to the following outputs.

Test of the equality of the variances

variance mean man	$5,297,043$
variance mean women	$2,304,340$
F-test	...
degrees of freedom	... and ...
p-value	0.085

1. Complete the previous table by indicated the value of the F-test.

   2,299

2. Complete the values on the number of degrees of freedom.

   27 and 16

3. What is the conclusion of this test?

☐ We reject the equality of the variances and to compare the wages we have to to proceed to a paired t-test.

☐ We do not reject the equality of the variance and to compare the wages we have to to proceed to a test test with equal variances.

☐ We reject the equality of the variances and to compare the wages we have to to proceed to a t-test with unequal variances.

☐ We do not reject the equality of the variances and to compare the wages we have to a t-test with unequal variances.

Test of the equality of the means

sample mean man	$23,389$
sample mean women	$20,503$
t-test	5.0632
p-value	0.085

What is the conclusion of this test which consists in comparing the mean values?

☐ Men and women have exactly the same salary.

☐ Men have a higher wage than women.

☐ We cannot conclude.

☐ Men and women wages are statistically different.

We are now interested in the equal treatment of employees in the different cities studied. The result of the study is provided in the following table

Results of the analysis

	Degrees of Freedom	Sum of Squares	Mean Squares	Statistical Test	p-value
City	2	$83,460,738$	41,730,389	9.5	0.0004
Residuals	42	$184,487,263$	4,392,554	X	X

1. Complete the number of degrees freedom in the previous table.

2. Complete the Mean Squares values in the above table.

3. Complete the value of the Statistical test in the previous test.

4. What is used test for this study.

   ☐ A t-test with equal variance.
   ☐ A paired t-test.
   ☐ A t-test with unequal variances.
   ☐ A F-test based on the fisher distribution.

In the last part of their study, the team want to study if it is possible to predict the wages in 2021, using the wages in 2020.

The results of the linear regression is provided in the following table.

Results of the linear regression: about the coefficients

	Estimation	Standard Error	t-value	p-value
Intercept	$20,067$	$3,513$	5.71	9.6×10^{-7}
Salary 2020	0.1164	0.1823	0.638	0.526

Results of the linear regression: study the model

Residual Standard Error	2485
Degree of freedom of Residuals	43
Degree of freedom of Regression	1
Residual Sum of Squares	265,534,675
Total Sum of Squares	$268,000,000$
Regression Sum of Squares	2,465,325
R^2	0.009
Adjusted R^2	-0.013
F-test	0.399
p-value	0.531

1. How many parameters do you have in this linear model

   ☐ 1.
   ☐ 2.
   ☐ 3.
   ☐ 4.

2. Complete the table **Results of the linear regression: about the coefficients**
3. Complete the number of degrees of freedom in the table **Results of the linear regression: study the model**
4. Complete the Residual and the Regression Sum of Squares in the **Results of the linear regression: study the model**
5. Complete the R^2 and the adjusted R^2 of the table **Results of the linear regression: study the model**
6. Complete the F-test of the table **Results of the linear regression: study the model**
7. Is it possi ble to predict the wages in 2021 using the wages in 2020.

   ☐ Yes

   ☐ No