Applied Statistics

MSc Digital Marketing & Data Science
Summary of Courses

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Abstract

This document summarizes the most important points to remember in the course. Only those elements that are essential for practical use are mentioned, but there are no illustrations. This document is not a course in itself, as it does not provide any explanations or details of the concepts. You will therefore need to refer to your course to find the origin of the quantities manipulated, their meaning and any examples of illustrations or explanations.

Keep in mind that it is important to do and redo the examples and exercises you've seen in class, so that you can memorize and practice the points you have learned.

Each Section of the document provide the most important content to keep in mind for associated lecture and also a little summary of the used functions in Excel.

1 Generalities about Statistics and Normal Distribution

This first lecture deals with some generalities in Probabilities and Statistics.

Important to remember

Theory

Two types of Random Variables :

- **Discrete law:** such as the Binomial distribution $\mathcal{B}(n, p)$.
  For this type of variable, let us say X, we can compute the probability that it takes a given value x, *i.e.* $\mathbb{P}[X = x]$.
- **Continuous law** as the Normal distribution $\mathcal{N}(\mu, \sigma)$.
  In this case, **the probability of such a law taking a specific value is always zero!**

  On the other hand, we can always calculate the probability that a random variable X, distributed according to a normal distribution for example, takes its values in an interval $[t_1, t_2]$:

  $$\mathbb{P}[t_1 < X < t_2].$$

  This **probability** is then the **area under the probability density function** which represent the density of the function.

The function $F(t) = \mathbb{P}[X \leq t]$ is called**continuous density function** of X. For all t, $F(t)$ gives the probability that the random variable X takes values lower or equal than t.

Pratice

If you want to compute $\mathbb{P}[X \leq t]$ where X is normally distributed with the parameters μ et σ, *i.e.* $X \sim \mathcal{N}(\mu, \sigma)$ we always to transform this distribution to **normal law which is centered and reduced,** Z applying

$$Z = \frac{X - \mu}{\sigma}.$$

We then have the following equality

$$\mathbb{P}[X \leq t] = \mathbb{P}\left[\frac{X - \mu}{\sigma} \leq \frac{t - \mu}{\sigma}\right] = \mathbb{P}\left[Z \leq \frac{t - \mu}{\sigma}\right].$$

It remains to look for the probability in the Z-table.
In any other situation, if Z is centered and reduced (*i.e.* $\mathcal{N}(0, 1)$), we briefly recall that:

- $\mathbb{P}[Z \geq t] = 1 - \mathbb{P}[Z \leq t]$,
- $\mathbb{P}[Z \geq t] = \mathbb{P}[Z \leq -t]$,

- $\mathbb{P}[t_1 \leq Z \leq t_2] = \mathbb{P}[Z \leq t_2] - \mathbb{P}[Z \leq t_1]$.

(**These relations are always true, not only for the Z-distribution** except for the second one which holds only if the distribution is symmetric.)

With Excel

With Excel, we can compute the quantiles $F^{-1}(p)$, where p is probability, associated to a gaussian distribution using the following formula:

Version	Function
ENGLISH	NORM.DIST(t, μ, σ, CUMULATIVE)
FRENCH	LOI.NORMALE.N(t; μ; σ; CUMULATIVE)

where μ and σ are respectively the mean and the standard deviation of the normal distribution. The last parameter is called Cumulative, if:

- **TRUE**: it computes $\mathbb{P}[X \leq t] = F(t)$, otherwise
- **FALSE**: it computes $f(t)$ the value of the density for the given t.

We can compute the quantiles $F^{-1}(p)$, where p is probability, associated to a gaussian distribution using the following formula:

Version	Function
ENGLISH	NORM.INV(p, μ, σ)
FRENCH	LOI.NORMALE.INVERSE.N(p; μ; σ)

where μ and σ are respectively the mean and the standard deviation of the normal distribution and $p \in [0, 1]$ is the level of the quantile. The quantile of order $p \in (0, 1)$ is the value z_p such that $\mathbb{P}[Z \leq z_p] = p$.

2 Sampling Estimation and Confidence Regions

The aim is first to build a confidence region on an unknown parameter μ which is the mean value of data distribution.
We have to consider two different cases whether σ is known or not.

When we have access to the standard deviation σ of the data distribution

Important to remember

Theory

Let us consider a sample of size n denoted $x_1, \ldots, x_n$,then, the estimator of the mean$\bar{x}_n$ est donné par

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n}.$$

This estimator of the mean $\bar{X}_n$ is a random variable whose distribution depends on the context. In the case where the standard deviation σ of the distribution is known and the data are from a normal distribution or our sample size is greater than 30, then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \simeq Z \sim \mathcal{N}(0,1),$$

where μ is the mean parameter we aim to estimate.

Confidence interval (symmetrical!, but non-symmetrical confidence intervals are also possible) of level $1-\alpha$ for the mean μ.

$$\left[\bar{x}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[\bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

is the quantile of order α of the normal distribution, *i.e.* it is the value for which a random variable Z following a normal distribution verifies :

$$\mathbb{P}[Z \leq z_\alpha] = \alpha.$$

We can also say that a proportion $1-\alpha$ of estimates of the mean $\bar{x}_n$ fall within the interval

$$\left[\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Pratice

To give a confidence interval of level $1-\alpha$ on an unknown parameter such as the mean μ **in the case where the standard deviation of the distribution σ is known**, we must

1. estimate the mean value $\bar{x}_n$ from the data
2. check the size n of our sample
3. determine the value of $z_{1-\alpha/2}$
4. calculate the bounds of the confidence interval from the above information

$$\left[\bar{x}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

If you want to check whether a machine is up to standard (you know the reference value μ), you can check whether $\bar{x}_n$ lies in the interval

$$\left[\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

We will proceed in the same way for the construction of this interval.

When we do not have access to the standard deviation σ of the data distribution

Student Distribution It is another distribution which is close the normal distribution and which depends on only one parameter that is the number of degree of freedom p. We denote this distribution $\mathcal{T}_p$.

Important to remember

Theory

Considering a sample of n measurements denoted $x_1, \ldots, x_n$, then the estimators of the mean $\bar{x}_n$ and variance s^2 are given by

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n}.$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

This estimator of the mean $\bar{X}_n$ is a **random variable** whose distribution depends on the context. In the case where we don't know the standard deviation σ of the distribution and the data come from a normal distribution or our sample size n is greater than 30, then

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \simeq T \sim \mathcal{T}_{n-1},$$

where μ is the unknown parameter to be estimated.
Confidence interval (symmetrical!, but non-symmetrical confidence intervals are also possible) of level $1-\alpha$ for the mean μ.

$$\left[\bar{x}_n - t_{1-\alpha/2}\sqrt{s^2/n}; \bar{x_n} + t_{1-\alpha/2}\sqrt{s^2/n}\right] = \left[\bar{x}_n + t_{\alpha/2}\sqrt{s^2/n}; \bar{x}_n + t_{1-\alpha/2}\sqrt{s^2/n}\right],$$

where t_{alpha} is the quantile of order α of the Student's law with $n-1$ degrees of freedom, *i.e.* is the value for which a random variable T following a Student's law with $n-1$ degrees of freedom verifies :

$$\mathbb{P}[T \leq t_\alpha] = \alpha.$$

We can also say that a proportion $1-\alpha$ of estimates of the mean $\bar{x}$ fall within the interval

$$\left[\mu - t_{1-\alpha/2}\sqrt{s^2/n}; \mu + t_{1-\alpha/2}\sqrt{s^2/n}\right].$$

Practice

To give a confidence interval of level $1-\alpha$ on an unknown parameter such as the mean μ **in the case where the standard deviation of the distribution σ is unknown**, we must

1. give an estimate of the mean value $\bar{x}$ from the data
2. estimate the standard deviation s from the data
3. check the size n of our sample
4. determine the value of $t_{1-\alpha/2}$
5. calculate the bounds of the confidence interval from the above information

$$\left[\bar{x}_n - t_{1-\alpha/2}\sqrt{s^2/n}; \bar{x}_n + t_{1-\alpha/2}\sqrt{s^2/n}\right]$$

If you want to check whether a machine is up to standard (you know the reference value μ), you can check whether $\bar{x}$ lies in the interval

$$\left[\mu - t_{1-\alpha/2}\sqrt{s^2/n}; \mu + t_{1-\alpha/2}\sqrt{s^2/n}\right].$$

Proceed in the same way for the construction of this interval.

With Excel

With Excel, we can compute the probabilities and the quantiles associated to the Student distribution with p degree of freedom using the functions below: (i) the probabilities and (ii) the quantiles

Version	Function
ENGLISH	T.DIST(t, p, CUMULATIVE)
FRENCH	LOI.STUDENT.N(t; p, CUMULATIVE)

where p is the number of degree of freedom and $t \in \mathbb{R}$. The last parameter is called Cumulative, if:

- **TRUE**: it computes $\mathbb{P}[T \leq t] = F(t)$, otherwise
- **FALSE**: it computes $f(t)$ the value of the density for the given t.

Version	Function
ENGLISH	T.INV(α, p)
FRENCH	LOI.STUDENT.INVERSE.N(α; p)

where p is the number of degree of freedom and $\alpha \in [0, 1]$.

Confidence region on an unknown proportion p

Important to remember

Theory

Considering a sample of n measurements denoted $x_1, \cdots, x_n$, an estimator of the proportion $\bar{p}$ is given by

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n},$$

where $x_i, i \in [\![1, n]\!]$ take the values 0 or 1.
This estimator of the proportion $\bar{p}$ is a **random variable** again. Asymptotically, when the sample size is large enough to verify $n\bar{p} \geq 5$ and $n(1-\bar{p}) \geq 5$

$$\frac{\bar{p} - p}{\sqrt{\bar{p}(1-\bar{p})/n}} \simeq Z \sim \mathcal{N}(0, 1),$$

where p is the unknown parameter to be estimated.
Confidence interval (symmetrical!, but non-symmetrical confidence intervals are also possible) of level $1-\alpha$ for proportion p.

$$\left[\bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \bar{p} + z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right].$$

Which is the same as:

$$\left[\bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \bar{p} + z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right],$$

where z_{alpha} is the quantile of order α of the centered-reduced Normal distribution, *i.e.* is the value for which a random variable Z following a centered-reduced Normal distribution verifies :

$$\mathbb{P}[Z \leq z_\alpha] = \alpha.$$

We can also say that a proportion $1-\alpha$ of the estimates of the proportion $\bar{p}$ fall within the interval

$$\left[p - t_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}; p + t_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}\right].$$

Practice

To give a confidence interval of level $1-\alpha$ on an unknown parameter such as the proportion p, we must

1. estimate the proportion $\bar{p}$ from the data
2. check the size n of our sample
3. determine the value of $z_{1-\alpha/2}$
4. calculate the bounds of the confidence interval from the above information

$$\left[\bar{p} - z_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}; \bar{p} + z_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}\right]$$

If you want to check whether a given proportion of the population has the studied characteristic (you know the reference value p), you can check if $\bar{p}$ lies in the interval

$$\left[p - z_{1-\alpha/2}\sqrt{p(1-p)/n}; p + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right].$$

We will proceed in the same way for the construction of this interval.

3 Hypothesis testing: Generalities on One sample test

In the following we consider U a random variable that follows any distribution (it could be the Normal distribution or the Student distribution for instance).

Important to remember

Theory: Type of test and rejection region

Let μ be the statistical quantity on which the test is conducted (it is similar for the proportion p), the mean value of the data distribution for instance. We briefly restate the definition of the **rejection regions** of H_0 according to the **formulation of the assumption** H_1.

$$H_0 : \mu = \mu_0 \in \mathbb{R} \quad \text{v.s.} \quad H_1 : \mu \neq \mu_0 \in \mathbb{R}.$$

Two tail test for which the rejection regions are defined as:

$$\left[-\infty; u_{\alpha/2}\right] \cup \left[u_{\alpha/2}; \infty\right]$$

and the $p-$value is given by $2\mathbb{P}[U \geq |u_{\text{test}}|]$, where U is random variable that has the same distribution as test, and test is the statistical test computed using our data.

$$H_0 : \mu = \mu_0 \in \mathbb{R} \text{ or } \mu \leq \mu_0 \in \mathbb{R} \quad \text{v.s.} \quad H_1 : \mu > \mu_0 \in \mathbb{R}.$$

Right tail test or **Upper tail test** for which the rejection region is defined by:

$$[u_{1-\alpha}; \infty]$$

and the $p-$value is given by $\mathbb{P}[U \geq u]$, where U is random variable that has the same distribution as test, and test is the statistical test computed using our data.

$$H_0 : \mu = \mu_0 \in \mathbb{R} \text{ or } \mu \geq \mu_0 \quad \text{v.s.} \quad H_1 : \mu < \mu_0.$$

Left tail test or **Lower tail test** for which the rejection region is defined by:

$$[-\infty; u_{\alpha}]$$

and the $p-$value is given by $\mathbb{P}[U \leq u]$, where U is random variable that has the same distribution as test, and test is the statistical test computed using our data.

If we use the p-value approach for our test, we reject the assumption H_0 if the p-value is lower than α.
Keep in mind that the p-value represents the risk you face by rejecting H_0.

Which test to use?

- For a test on **the mean value μ when the standard deviation σ is known**, we are going to use the **Z-distribution**. We also consider the following the following statistical test:

$$z_{\text{test}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \underset{H_0}{=} \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

- For a test on **the mean value μ when the standard deviation σ is known**, we are going to use the **t-distribution** with a number of degrees of freedom equal to the sample size n minus one. We also consider the following the following statistical test:

$$t_{\text{test}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \underset{H_0}{=} \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}},$$

where s denotes the standard deviation estimated on the sample.

- For a test on a **proportion**, we are going to use the **Z-distribution**. We also consider the following the following statistical test:

$$z_{\text{test}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{H_0}{=} \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where p_0 denotes a reference for the proportion.

Pratice

Steps to follow to conduct your test

1) Define the *Null Hypothesis* H_0 and the *Alternative Hypothesis* H_1
2) Fix an error rate α, which will be used to take our final decision.
3) Determine the right law of U under H_0, *i.e.* the statistical test and the distribution of the associated random variable (see above). These quantities are used (with α to draw the conclusion of your hypothesis testing.
4) Compute the the value of the statistical test u_{test} using your data.
5) Conclude using either you reject H_0 or not using *the critical value* approach or the *p-value* approach.

4 Hypothesis testing: Two sample test and ANOVA

Two sample test

Important to remember

Theory

You want to compare if the mean values,μ_1 and μ_2, of two studied populations are equal or not, *i.e.* perform the following test

$$H_0 : \mu_1 = \mu_2 \quad \text{v.s.} \quad H_1 : \mu_1 \neq \mu_2.$$

Remark: it is also possible to test if one mean value is greater or lower than the other one.

To proceed to this verification, we are going to distinguish three different scenarios.

- **The two studied populations are independent.**

  In this situation, we have to distinguish the two following scenarios:

  1. **The two populations have the same variance.**

     We first compute the mean values associated to the different samples $\bar{x}_1$ and $\bar{x}_2$. We also compute the variances of the two samples s_1^2 and s_2^2.
     First, we compute a common variance s_p^2 associated to the two groups. Such a variance is defined by:

     $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

     where n_1 and n_2 are the sample size of the two samples respectively.

     Then we consider the following statistical test which will be used to conclude:

     $$t_{\text{test}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

     This statistical test follows a Student distribution with $n_1 + n_2 - 2$ degrees of freedom.

  2. **The two populations have different variances.**

     We have not seen how to deal with such a situation. However, there is a function in Excel which will help you to conduct the test.

- **The two studied populations are related.**

  This correspond to the case where we the two studied samples correspond to the same population but at different moments (before and after a treatment for instance).

  Here, we need to remove the part variation due to the change in values before and after a treatment. To this aim, we first create another sample D defined by

  $$D_i = X_i^{(1)} - X_i(2),$$

where $X_i^{(j)}$ is used to denote the the i-th observation of group j.

Then we compute the needed statistics $\bar{x}_D$ and s_D, the sample mean and sample standard deviation respectively. Then we consider the following statistical test

$$t_{\text{test}} = \frac{\bar{x}_D}{\frac{s_D}{\sqrt{n}}},$$

where n is the sample size of D.

Two compare the variances of the two populations, you are going to perform a F-test, where the assumption are stated as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{v.s.} \quad H_1 : \sigma_1^2 \geq \sigma_2^2 \text{ (or } \sigma_2^2 \geq \sigma_1^2).$$

The statistical quantity F_{test} that is used is

$$F_{\text{test}} = \frac{s_+^2}{s_-^2},$$

where s_+^2 refers to the sample variance of the group who the largest one and s_-^2 refers to the variance of the group who has the smallest one.
And the assumption H_0 is rejected when this ratio is higher than $F_{1-\alpha}$ for a Fisher distribution with $n_+ - 1$ and $n_- - 1$ degrees of freedom, or using the p-value approach. We recall that n_+ refers to the sample size of the group who the largest variance and n_- refers to the sample size of the group who has the smallest variance.

Pratice

In order to choose the right test to compare the two populations, you need to follow the following steps:

1. You need to determine if the two studied populations are **independent** or **related**.

2. If the two populations are **independent**:
   (a) you need first to check if they have the same variance or not, *i.e.* apply an F-test to compare the variances of the two populations.
   (b) choose the appropriate T-test according to the output of the F-test.
      i. if you **do not reject** the assumption which states that the variances of the wo groups are equal, you are going to apply a *Pooled* T-test.
      ii. if you **reject** the assumption which states that the variances of the wo groups are equal, you are going to apply a *Separate* T-test. This second test shall be done with Excel directly.

3. If the two studied populations are **related:**
   (a) you will create your new sample D and compute the different needed statistics.
   (b) perform a T-test on one population to conclude.

Important to remember

Theory

The aim is two study the relation between one **quantitative variable** and one **qualitative variable** (*e.g.* different marketing) strategies) and see, for instance, if the average turnover of different stores depends on the applied marketing strategies.

To conduct this study, we are going to test the following assumption:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_c \quad \text{v.s.} \quad H_1 : \mu_i \neq \mu_j \text{ for a couple } i \neq j.$$

To proceed we need to satisfy the following assumptions:

- data have to be normally distributed (or at least of size 30).
- data shall be independent.
- the different populations shall have the same variance.

We study the variance, which represents the quantity of information that you have in your data, and see the part of the information explained by the different groups (due to the values of the categorical variable) and the part not explained by this variable, *i.e.*

$$SST = SSA + SSW,$$

where :

- **SST:** *total sum of squares* or *total variation in the data*

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

  where $\bar{x}$ represents the mean of all the data.

- **SSA:** *sum of squares among groups*

$$SSA = \sum_{k=1}^{c} n_k (\bar{x}_k - \bar{x})^2,$$

  where c represents the number of groups, $\bar{x}_k$ and n_k represent the mean value and the number of examples in group c respectively.

- **SSW:** *sum of squares within groups*

$$SSW = \sum_{k=1}^{c} \sum_{i=1}^{n_k} (x_i - \bar{x}_k)^2,$$

  where c represents the number of groups, $\bar{x}_k$ and n_k represent the mean value and the number of examples in group k respectively.

We can summarize the ANOVA as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Among groups	SSA	$c-1$	$\text{MSA} = \dfrac{\text{SSA}}{c-1}$
Within groups	SSW	$n-c$	$\text{MSW} = \dfrac{\text{SSW}}{n-c}$
Total	SST	$n-1$	$\text{MST} = \dfrac{\text{SST}}{n-1}$

We use the following statistical quantity to perform our test:

$$F_{\text{test}} = \frac{\text{MSA}}{\text{MSW}}.$$

And the assumption H_0 is rejected when this ratio is higher than $F_{1-\alpha}$ for a Fisher distribution with $c-1$ and $n-c$ degrees of freedom, or using the p-value approach.
Saying differently, we reject the assumption H_0 when the part of the variance explained by the factor is much higher than the one that is not explained by the factor.

Pratice

In practice, there is nothing more to do than using the appropriate function in Excel:

Data $\longrightarrow$ Data Analysis $\longrightarrow$ One-Way Anova

With Excel

With Excel, we can compute the probabilities and the quantiles associated to the Fisher distribution with n_1 and n_2 degrees of freedom using the functions below: (i) the probabilities and (ii) the quantiles

Version	Function
ENGLISH	F.DIST(t, n_1, n_2, CUMULATIVE)
FRENCH	LOI.FISHER.N(t; n_1; n_2; CUMULATIVE)

where n_1 and n_2 are the numbers of degrees of freedom and $t \in \mathbb{R}_+$. The last parameter is called Cumulative, if:

- **TRUE**: it computes $\mathbb{P}[F \leq t] = F(t)$, otherwise
- **FALSE**: it computes $f(t)$ the value of the density for the given t.

Version	Function
ENGLISH	F.INV(α, n_1, n_2)
FRENCH	LOI.FISHER.INVERSE.N(α; n_1; n_2)

where n_1 and n_2 are the numbers of degree of freedom and $\alpha \in [0, 1]$.

5 Simple and Multiple Linear Regression

Important to remember

Theory

The aim is two study the relation between two **quantitative variables** strategies) and see if there is a *correlation between these two random variables.*

Thus, we consider the following regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

where:

1. Y is the *dependent* variable, *i.e.*, the variable whose values we want to predict.
2. $X_j,\ j \in [\![0, p]\!]$ are the *independent* variables, *i.e.*, the variable that are used to estimate the values of the variable Y.
3. $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters of the regression. β_0 is an intercept and represent the theoretical value of Y when all the X_j are equal to 0.
4. ε represents a *random error*, also called, *residuals.*

We assume, for the regression model, that the errors are *independent and identically distributed*, they follow a *Normal distribution*, with a mean equal to 0 and a variance equal to σ^2.

$$\varepsilon \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Prediction

Suppose that we have estimated the parameters of the regression using our data. We denote by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ the estimated parameters.
Then the prediction of the y value, denoted by $\hat{y}$ for a set of value $(x_1, x_2, x_3, \ldots, x_p)$.

Overall significance of the model

To conduct this study, we are going to test the following assumption:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{v.s.} \quad H_1 : \beta_j \neq 0 \text{ for one value } j \in [\![1, n]\!].$$

To conduct this test we are going to study the part of variance *explained by the regression* (SSR) and the one that *is not explained by the regression* (SSE). Such as in the ANOVA, the total variance (SST) that we have in the data can written as the sum of these two terms.

$$SST = SSR + SSE,$$

where:

- **SST:** *total sum of squares* or *total variation in the data*

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

  where $\bar{y}$ represents the mean of all the data.

- **SSR:** *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

- **SSE:** *error sum of squares*

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$

We can summarize the ANOVA of regression as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Regression	SSR	p	$\text{MSR} = \dfrac{\text{SSR}}{p}$
Errors	SSE	$n-p-1$	$\text{MSE} = \dfrac{\text{SSE}}{n-p-1}$
Total	SST	$n-1$	$\text{MST} = \dfrac{\text{SST}}{n-1}$

We use the following statistical quantity to perform our test to see if the linear regression makes sense:

$$F_{\text{test}} = \frac{\text{MSR}}{\text{MSE}}.$$

And the assumption H_0 is rejected when this ratio is higher than $F_{1-\alpha}$ for a Fisher distribution with p and $n-p-1$ degrees of freedom, or using the p-value approach.
Saying differently, we reject the assumption H_0 when the part of the variance explained by the regression is much higher than the one that is not explained by the regression.

Estimation of the variance of the errors The variance of the errors σ^2, denoted by s^2 is given by:

$$s^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This quantity is nothing than MSE.

Quality of the model To measure the quality of the model, *i.e.* **how good is the link between Y and the set of independent variables**, we introduce the R^2 which is given by:

$$R^2 = \frac{SSR}{SST} \in [0, 1].$$

An R^2 value close to 1 means that there is a strong correlation between the Y and the various independent variables. But that does not mean the link is significant! We need to look at our Fisher test to judge this.

We also introduce the adjusted R^2_{adj} to compare the quality of different models that involve a different number of parameters. This quantity is defined by:

$$R^2_{\text{adj}} = 1 - (1 - R^2) \frac{n-1}{n-p-1}.$$

Regression parameters

If the estimated will be directly given by Excel, it is also important to look at the importance of each variable X_j and the significance of their associated slope β_j. To test if the slope is significant, *i.e.* if the variable X_j is important for the prediction task, we conduct the following hypothesis testing:

$$H_0 : \beta_j = 0 \quad \text{v.s.} \quad H_1 : \beta_j \neq 0.$$

The statistical test is computed as follows

$$t_{\text{test}} = \frac{\hat{\beta}_j}{s_j},$$

where s_j is the standard deviation of the $\hat{\beta}_j$ (computed by excel). This statistical test follows a Student distribution with $n - p - 1$ degrees of freedom. If you reject the assumption H_0, it means that the variable X_j is important for the regression.

Pratice

In practice, there is nothing more to do than using the appropriate function in Excel:

$$\text{Data} \longrightarrow \text{Data Analysis} \longrightarrow \text{Regression}$$

Then, you only need to look at the p-values and use the model for prediction.