Applied Statistics

Ms 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

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 :

$$P[Z \leq z_\alpha] = \alpha.$$

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 $x_1, \dots, x_n$, 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 :

$$P[T \leq t_\alpha] = \alpha.$$

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 :

$$P[Z \leq z_\alpha] = \alpha.$$

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.

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

Talk about the different settings (independent vs related) and the different tests + F-test

ANOVA

One Way ANOVA : variance decomposition and so on ...

5 Simple and Multiple Linear Regression

Generalities about the linear regression.