Statistical hypothesis testing
The parametric and nonparametric cases

Madalina Olteanu, Université Paris 1

2016-2017

Contents

Chapter 1

Parametric hypothesis testing

1.1 An introduction on statistical hypothesis testing

A statistical test allows to decide between two possible hypothesis, given the observed values of an n-sample. [1].

Let us consider two statistical hypothesis H_0 and H_1, among which one and only one is true. The statistical decision consists in choosing between H_0 and H_1, while this decision may be right (true) or wrong (false). Then, four possible situations occur, two of them leading to errors, and which may be resumed as in the following table :

Reality Decision	H_0 true	H_0 false
Accept H_0	OK	β
Reject H_0	α	OK

The two possible errors may be now defined as :

1. The **1st type error** is the probability of mistakenly rejecting H_0:

$$\alpha = \mathbb{P}\left(\text{reject } H_0 | H_0 \text{ true}\right)$$

2. The **2nd type error** is the probability of mistakenly accepting H_0:

$$\beta = \mathbb{P}\left(\text{accept } H_0 | H_0 \text{ false}\right)$$

In practice, these errors correspond to different risks, as one may see in the following example.

<u>**Example**</u> : One wishes to know if introducing a speed limit on a given road reduces the number of accidents. In order to reach a decision upon this issue, the local authorities make a trial and introduce a speed restriction during three months. At the end of the period, they compare the results before and after having introduced the limitation. The two hypothesis being tested are :

H_0 : "the speed restriction has no impact on the number of accidents".
H_1 : "the speed restriction allows to decrease the number of accidents".

The two possible decisions are then :

1. If one decides H_0 is true, the speed restriction is not set up.
2. If one decides H_0 is wrong, then the speed restriction is set up.

Let us now look at the interpretation of the two associated errors (risks) :

1. α is the error committed by rejecting H_0 when H_0 is true, that is the error which consists in deciding to set up a speed limitation, while this limitation does not diminish the number of accidents.

[1]These lecture notes are greatly inspired from the online lecture notes and the books [2, 5, 1, 3, 4]

2. β is the error committed by accepting H_0 when H_0 is wrong, that is the error which consists in deciding not to set up a speed limitation, while this limitation would decrease the number of accidents.

Hypothesis selection
Normally, the two errors associated to a statistical test should be low. However, in practice, it is impossible to minimize (or at least control) both α and β. Consequently, when performing a statistical test, α **is generally fixed** (usual values are 0.01, 0.05 or 0.10), while β is computed ex post, provided that the probability distributions under H_1 are completely known. For this reason, the manner of selecting the hypothesis is crucial for the result and also for the trustworthiness of the test.

In practice, the choice of H_0 is dictated by various reasons :

- since one would not want to drop H_0 too often, H_0 must be an hypothesis solidly established and which hasn't been contradicted so far by the experiment;
- H_0 is an hypothesis considered as particularly important, for reasons which may be subjective (in the previous example, the choice of the hypothesis may be different according to the organization requesting the test);
- H_0 may correspond a cautious hypothesis (for instance, when testing the innocuousness of a vaccine, it is wiser to start with an hypothesis unfavorable to the new product);
- H_0 is the only hypothesis easy to express (for example, when testing $m = m_0$ against $m \neq m_0$, it is obvious that $H_0 : m = m_0$ only allows to perform easy computations).

Remark : One should also know that β varies oppositely to α. When one wants to decrease the 1st type error α, the it automatically increases $1 - \alpha$, the probability of accepting H_0 when H_0 is true. In this case, the decision rule becomes more strict since it would hardly ever drop H_0, hence it would also occasionally keep H_0 mistakenly. By seeking not to drop H_0, one ends keeping it almost every time, hence β, the risk of mistakenly keeping H_0, becomes larger.

Definition : $1 - \beta$, the probability of rightly rejecting H_0, is called **the power of the test**. Between two tests having the same level α, one should always prefer the one having the highest power.

Test statistic :
Once α is fixed, the next step consists in elaborating a test statistic from the n-sample, that is a measurable map of the sample $T = T(X_1, ..., X_n)$. This statistic should carry the maximum of information on the considered problem and should verify some conditions:

- the probability distribution of T under H_0 must be exactly known;
- the behavior of T under H_1 must be known, at least qualitatively, for determining the form of the rejection region.

Critical region or rejection region :
The critical region W is the set of values for the test statistic T which lead to rejecting H_0 in favor of H_1. The form of the critical region W is determined by the behavior of T under H_1 (the event $\{T \in W\}$ must be "dubious" or of "low probability" if H_0 is true). The exact form of W is derived by writing :

$$\mathbb{P}_{H_0}(T \in W) = \mathbb{P}(T \in W | H_0) = \alpha$$

1.1.1 The steps in the construction of a test

In summary, the steps of a statistical test are the following :

1. Select the hypothesis H_0 and H_1;
2. Fix the level of the test or the 1st type error equal to α;
3. Select the test statistic, T;
4. Determine the form of the rejection region W, depending on the behavior of T under H_1;
5. Explicitly compute the rejection region W according to α;

6. (Optional) Compute the 2nd type error and/or the power of the test;
7. Compute the observed value, t, for the test statistic T;
8. According to t, decide whether to accept or not H_0.

<u>**Remark**</u> :
Statistical tests may be classified into two large categories:

1. **Parametric hypothesis testing**: the family of probability distributions P_θ is known and the test concerns the value of the parameter θ only ;
2. **Nonparametric hypothesis testing** : the family of probability distributions P_θ is unknown and the test is about issues such as: the probability distribution belongs to the Gaussian family?; the probability distribution is uniform?; are two samples issued from the same probability distribution or not?; ...

1.1.2 The p-value

In most statistical software, hypothesis testing is slightly different from the general framework introduced above. Indeed, rather than fixing a level α for the test and deciding whether the hypothesis H_0 should be rejected or not, a p-value is directly computed from the observed values of the sample.

<u>**Definition**</u> : The p-value is equal to the probability of obtaining a value for the test statistic T at least as "extreme" or as "awkward" as the observed t, under the hypothesis H_0.

The hypothesis H_0 will then be rejected whenever the p-value is lower than some fixed threshold (0.01, 0.05 or 0.10). A very low p-value means that the observed test statistic t is highly unlikely under the hypothesis H_0 and leads to rejecting H_0.

1.2 Some general considerations on parametric tests

In this section, we shall place ourselves in the parametric framework. Let us consider a dominated parametric model $(E^n, \mathcal{E}_n, \mathbb{P}_\theta, \theta \in \Theta)$, where $\Theta \in \mathbb{R}^d$, and let θ be the "true" value of the parameter. Hypothesis testing consists in this case in deciding one of the hypothesis

$$\begin{cases} H_0 & : & \theta \in \Theta_0 & \text{(null hypothesis)} \\ H_1 & : & \theta \in \Theta_1 & \text{(alternative hypothesis)} \end{cases},$$

where $\Theta_0, \Theta_1 \subset \mathbb{R}^d$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

Within this framework, one may specify two categories of parametric tests, according to the content of Θ_0 and Θ_1.

<u>**Definition**</u>: A hypothesis (H_0 of H_1) is said to be **simple** if it is associated to a singleton (Θ_0 or Θ_1). Otherwise, it is said to be **composite**. In dimension one ($\Theta \in \mathbb{R}$), if H_0 is simple of the form $\theta = \theta_0$ and if H_1 is composite of the form $\theta > \theta_0$ or $\theta < \theta_0$, one shall speak of **unilateral** test; if H_1 is composite of the form $\theta \neq \theta_0$, we shall speak of **bilateral** test.

<u>**Definition**</u>: Consider $\hat{T}$ a measurable function of an n-sample $(X_1, ..., X_n)$, issued from the model $(E^n, \mathcal{E}_n, \mathbb{P}_\theta, \theta \in \Theta)$, valued in $\mathbb{R}^d$. $\hat{T}$ is called a **test statistic**. The test is then defined by the decision function $\hat{\phi} = \mathbb{1}_{\{\hat{T} \in W\}}$, where $W \in \mathbb{R}^d$ is called **critical region** or **rejection region**. Its complement in $\mathbb{R}^d$, $\overline{W}$, is called acceptance region. If $\hat{\phi} = 1$, then H_0 is rejected, otherwise it is accepted.

<u>**Definition**</u>: For the test statistic $\hat{T}$, let us define

1. **The 1st type error**: $\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\hat{T} \in W \right)$;
2. **The 2nd type error**: $\beta = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta \left(\hat{T} \notin W \right)$;

3. **The power function**: $\pi(\theta) = \mathbb{P}_\theta\left(\hat{T} \notin W\right)$, $\forall \theta \in \Theta$;

4. **The efficiency function**: $e(\theta) = \mathbb{P}_\theta\left(\hat{T} \in W\right)$, $\forall \theta \in \Theta$.

Definition: A statistical test $\hat{\phi}$ is **unbiased** if $\pi(\theta) \geq \alpha$, $\forall \theta \in \Theta_1$.

Definition: Let $\hat{\phi}_1$ and $\hat{\phi}_2$ two statistical tests of level $\leq \alpha$ for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. We say that $\hat{\phi}_1$ is **uniformly most powerful (UMP)** than $\hat{\phi}_2$ if $\pi_{\hat{\phi}_1}(\theta) \geq \pi_{\hat{\phi}_2}(\theta)$, $\forall \theta \in \Theta_1$.
When the alternative is simple, we shall only speak of the most powerful test.

1.2.1 Exercices

Exercice 1: One knows that a box contains either three red balls and five black balls, or five red balls and three black balls. Three balls are drawn at random from the box. If strictly less than three balls are red, then the decision is that the bow contains three red balls and five black balls. Compute α and β.

Exercice 2: One knows that a box contains one or two winning tickets and eight losers. Let H_0 and H_1 be the two possible hypothesis. In order to test H_0 against H_1, the tickets are drawn, without replacement, until one winning ticket is obtained. Consider X the random variable associating the number of drawn tickets. Compute the probability distribution of X under H_0 and under H_1. The rejection region is $W = \{X \leq 5\}$. Compute α and β.

Exercice 3: The number of accidents recorded each month on a given route may be considered as the realization of a random variable distributed according to a Poisson distribution with parameter $\lambda > 0$. We admit that λ can have two values only, $\lambda_1 = 4$ and $\lambda_2 = 6,25$, corresponding to the hypothesis H_1 and H_2 to be tested.
According to the value of λ, the decision to improve the route is taken or not. For this test, we dispose of a nine-month observed sample :

$$x_1 = 4\ ,\ x_2 = 9\ ,\ x_3 = 2\ ,\ x_4 = 5\ ,\ x_5 = 6\ ,\ x_6 = 2\ ,\ x_7 = 3\ ,\ x_8 = 7\ ,\ x_9 = 7$$

1. According to the test is requested by an automobile association or by the ministry for equipment and transport, which of the two hypothesis should be chosen as the null hypothesis by the statistician?

2. The following decision rule is adopted :

   $$\begin{array}{lll} \text{si } \overline{x} \geq c & , & \text{the route is improved;} \\ \text{si } \overline{x} < c & , & \text{the route is not improved.} \end{array}$$

   Using a Gaussian approximation for the probability distribution of $\overline{X}$, compute c_1 or c_2, depending on the hypothesis selected as null among H_1 or H_2, with a 1st type error α =1%, and then evaluate the associated powers π_1 and π_2.

3. According to the hypothesis chosen as null, which decision will be taken based upon the available data?

4. Is it possible to answer the following question:

   Is H_1 more probable than H_2?

   One should first examine in which framework this question makes sense, and then which supplementary (a priori) information would be necessary for answering it.

1.3 Neyman-Pearson lemma

Example: Consider $(X_1, ..., X_{16})$ an iid 16-sample, drawn from a Gaussian distribution, $\mathcal{N}(m, 1)$, $m \in \mathbb{R}$ is the unknown parameter. One has to decide between two possible values for m:

$$H_0 : m = 0 \text{ against } H_1 : m = 2\ .$$

Intuitively, since the parameter to test is the expected value of the variables in the sample, the empirical mean, $\overline{X} = \frac{1}{16}\sum_{i=1}^{16} X_i$, may be a good test statistic. Moreover, from the general properties of Gaussian vectors follows that

$$\overline{X} \sim_{H_0} \mathcal{N}(0, \frac{1}{16}) \ ; \ \overline{X} \sim_{H_1} \mathcal{N}(2, \frac{1}{16}) \ .$$

The test statistic $\overline{X}$ is a Gaussian distribution centered in 0 under H_0 and it shifts towards the right under H_1 (the variance remains unchanged). Hence, if $\alpha = 0.05$ is the level of the test, one would select as rejection region $W = \left\{\overline{X} > \frac{q_{1-\alpha}}{4}\right\}$, where $q_{1-\alpha} = 1.64$ is the $(1-\alpha)$-quantile of the $\mathcal{N}(0,1)$ distribution.

The power associated to this test may then be computed

$$\pi = \mathbb{P}_{H_1}(W) = \mathbb{P}_{m=2}\left(\overline{X} > \frac{1.64}{4}\right) \simeq 1 \ .$$

The power of this "intuitive" or "reasonable" test is thus very close to 1, but is there a way to state, with certainty, that this test is the most powerful among all tests of level $\alpha = 0.05$? The following results will allow to answer this question.

1.3.1 Randomized tests

The framework of not-randomized tests is, in some sense, an ideal situation in which, based on the observation of a random variable, x, one decides whether to reject or not the null hypothesis H_0. Here, we place ourselves in a more general context where, for some values of x, reaching a statistical decision may turn out to be a difficult issue (for instance, when the probability distribution is discrete and there is no continuous approximation available). We are then led to use a **randomized** test, where we consider that for some values x, the decision function $\phi(x) \in]0,1[$, defining a probability of rejecting H_0 based on the observation x.

Definition: A **randomized test** is defined by a decision rule, $\phi : (E, \mathcal{E}) \to [0,1]$ such that

- $\phi(x) = 1$ implies to reject H_0 based on x;
- $\phi(x) = 0$ implies to accept H_0 based on x;
- $\phi(x) = c \in]0,1[$ implies to reject H_0 with probability c, based on x.

Remark: The interest of randomized tests is, above all, theoretical. In practice, they are used but very seldom.

Interpretation: One may conceive a randomized test as the first step of a decision process. Based on the observation x,

1. Define a probability $\phi(x)$ of rejecting H_0;
2. Draw an element in $\{0,1\}$ according to a Bernoulli distribution, $\mathcal{B}(\phi(x))$. Let y be the value observed for this Y variable.

Eventually, one decides to reject H_0 if $y = 1$ and accept H_0 if $y = 0$.

The probability distribution of Y, conditionally to $X = x$ (and which does not depend on the parameter θ) is $\mathcal{B}(\phi(x))$. Hence, $\mathbb{E}(Y|X = x) = \phi(x)$ and $\mathbb{E}(Y|X) = \phi(X)$. Furthermore,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(\phi(X)) \ \Rightarrow \ Y \sim \mathcal{B}(\phi(X)) \ .$$

Definition: The **1st type error** for a randomized test ϕ_c is defined as the mapping $\alpha : \Theta_0 \to [0,1]$ such that $\alpha(\theta) = \mathbb{E}_\theta(\phi(X))$, $\forall \theta \in \Theta_0$.

Definition: Consider $\alpha_0 \in [0,1]$ fixed. A randomized test ϕ_c is

- of **level** α_0, if $\sup_{\theta\in\Theta_0} \alpha(\theta) \leq \alpha_0$;
- of **size** α_0, if $\sup_{\theta\in\Theta_0} \alpha(\theta) = \alpha_0$.

Definition: The **2nd type error** for a randomized test ϕ_c is defined as the mapping $\beta : \Theta_1 \to [0,1]$ such that $\beta(\theta) = \mathbb{E}_\theta(1 - \phi(X))$, $\forall \theta \in \Theta_1$.

Definition: The **power function** of a randomized test ϕ_c is defined as the mapping $\pi : \Theta_1 \to [0,1]$ such that $\pi(\theta) = 1 - \beta(\theta) = \mathbb{E}_\theta(\phi(X))$, $\forall \theta \in \Theta_1$.

1.3.2 Simple hypothesis testing

In this section, we are in the parametric framework $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$, dominated by the probability measure μ. Let $\mathcal{L}(X, \theta)$ be the likelihood of X, a random vector defined on the above probability space.

Consider two simple hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. The model is then dominated, for instance, by $\mu = \mathbb{P}_{\theta_0} + \mathbb{P}_{\theta_1}$.

Contrary to the general framework, the powers of two tests are in this case always comparable. So one might consider the existence of a most powerful test.

Definition: A **Neyman-Pearson test** is of the form

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } \mathcal{L}(X,\theta_1) > k\mathcal{L}(X,\theta_0) \\ 0 & , \quad \text{if } \mathcal{L}(X,\theta_1) < k\mathcal{L}(X,\theta_0) \\ c & , \quad \text{if } \mathcal{L}(X,\theta_1) = k\mathcal{L}(X,\theta_0) \end{cases} .$$

Remark: If $\mathbb{P}_{\theta_0}(\{x, \mathcal{L}(x,\theta_1) = k\mathcal{L}(x,\theta_0)\}) = 0$, then $c = 0$ or $c = 1$, and the test is not randomized.

Remark: The Neyman-Pearson test is also called **likelihood ratio test (LRT)**. H_0 is rejected when the quotient $\frac{\mathcal{L}(X,\theta_1)}{\mathcal{L}(X,\theta_0)}$ is large, that is when θ_1 appears as being more likely than θ_0.

Proposition (existence): For all $\alpha \in [0,1]$, there exists a Neyman-Pearson test of size α.

Proof: TODO.

Neyman-Pearson Lemma: Let $\alpha \in [0,1]$.

1. A Neyman-Pearson test of size α is the most powerful test among all tests of H_0 against H_1 of size α.
2. Conversely, a most powerful test of size α is a Neyman-Pearson test of size α.

Proof:

1. Consider $\phi_{k,c}$ a Neyman-Pearson test, of size α, et consider also $\phi^\star$, a test of level α: $\mathbb{E}_{\theta_0}(\phi^\star) \leq \alpha$. One has

   - $\phi(x) = 1 \geq \phi^\star(x)$, if $\mathcal{L}(x,\theta_1) > k\mathcal{L}(x,\theta_0)$;
   - $\phi(x) = 0 \leq \phi^\star(x)$, if $\mathcal{L}(x,\theta_1) < k\mathcal{L}(x,\theta_0)$.

   Hence, for all $x \in E$,

   $$(\phi(x) - \phi^\star(x))(\mathcal{L}(x,\theta_1) - k\mathcal{L}(x,\theta_0)) \geq 0 \ \Rightarrow \int_E (\phi(x) - \phi^\star(x))(\mathcal{L}(x,\theta_1) - k\mathcal{L}(x,\theta_0))d\mu(x) \geq 0 \ .$$

   It follows that

   $$\begin{aligned} \mathbb{E}_{\theta_1}(\phi(X)) - \mathbb{E}_{\theta_1}(\phi^\star(X)) &\geq k(\mathbb{E}_{\theta_0}(\phi(X)) - \mathbb{E}_{\theta_0}(\phi^\star(X)) \\ &\geq k(\alpha - \mathbb{E}_{\theta_0}(\phi^\star(X))) \\ &\geq 0 \ . \end{aligned}$$

   Consequently, $\mathbb{E}_{\theta_1}(\phi(X)) \geq \mathbb{E}_{\theta_1}(\phi^\star(X))$, which means that ϕ is more powerful than $\phi^\star$.

2. According to the previous proposition, there exists a Neyman-Pearson test ϕ of size α. Consider also a test $\phi^\star$, of level α, most powerful among all tests of level α. According to (1), ϕ is also the most

powerful among all tests of level α. Hence, $\mathbb{E}_{\theta_1}(\phi(X)) = \mathbb{E}_{\theta_1}(\phi^\star(X))$. We also checked in (1) that $(\phi(x) - \phi^\star(x))(\mathcal{L}(x,\theta_1) - k\mathcal{L}(x,\theta_0)) \geq 0$. It follows that

$$\begin{aligned}
\textstyle\int_E (\phi(x) - \phi^\star(x))(\mathcal{L}(x,\theta_1) - k\mathcal{L}(x,\theta_0))d\mu(x) &= \mathbb{E}_{\theta_1}(\phi) - \mathbb{E}_{\theta_1}(\phi^\star) + k(\mathbb{E}_{\theta_0}(\phi^\star) - \mathbb{E}_{\theta_0}(\phi)) \\
&= k(\mathbb{E}_{\theta_0}(\phi^\star) - \mathbb{E}_{\theta_0}(\phi)) \\
&= k(\mathbb{E}_{\theta_0}(\phi^\star) - \alpha) \\
&\geq 0\ .
\end{aligned}$$

We thus get that $\mathbb{E}_{\theta_0}(\phi^\star) \geq \alpha$. Since $\phi^\star$ is of level α, $\phi^\star$ is of size α if

$$\int_E (\phi(x) - \phi^\star(x))(\mathcal{L}(x,\theta_1) - k\mathcal{L}(x,\theta_0))d\mu(x) = 0\ ,$$

which means that, μ-a.s., $(\phi(x) - \phi^\star(x))(\mathcal{L}(x,\theta_1) - k\mathcal{L}(x,\theta_0)) = 0$, or also $\phi(x) = \phi^\star(x)$ for μ-a.s. every x such that $\mathcal{L}(x,\theta_1) \neq k\mathcal{L}(x,\theta_0)$. We may then conclude that $\phi^\star$ is a Neyman-Pearson test.

Remark: If T is an exhaustive statistic for the parameter θ, according to Neyman's factorization theorem, one has that $\mathcal{L}(x,\theta_0)) = h_{\theta_0}(T(x))g(x)$, $\mathcal{L}(x,\theta_1)) = h_{\theta_1}(T(x))g(x)$, and the Neyman-Pearson test in this case can be written in terms of $T(x)$ only.

Example: Univariate Gaussian

Let us consider $(X_1, ..., X_n)$ an n-sample from a Gaussian distribution $\mathcal{N}(m,1)$, where $m \in \mathbb{R}$ is the unknown parameter and the variance is a priori known, $\sigma^2 = 1$. The hypothesis to be tested are $H_0 : m = m_0 = 0$ versus $H_1 : m = m_1 = 1$, with a level $\alpha = 0.05$. Let us build the most powerful test, according to the Neyman-Pearson lemma.

The likelihood of the n-sample for any value m of the parameter is

$$\begin{aligned}
\mathcal{L}(X_1, ..., X_n; m) &= (2\pi)^{-\frac{n}{2}} \exp\left(-\tfrac{1}{2}\textstyle\sum_{i=1}^n (X_i - m)^2\right) \\
&= (2\pi)^{-\frac{n}{2}} \exp\left(-\tfrac{1}{2}\textstyle\sum_{i=1}^n X_i^2\right) \exp\left(-\frac{nm^2}{2}\right) \exp\left(m\textstyle\sum_{i=1}^n X_i\right)
\end{aligned}$$

The likelihood ratio may then be written :

$$\begin{aligned}
V(X_1, ..., X_n) &= \frac{\mathcal{L}(X_1,...,X_n;m_0)}{\mathcal{L}(X_1,...,X_n;m_1)} = \frac{\exp\left(-\frac{nm_0^2}{2}\right)\exp\left(m_0\sum_{i=1}^n X_i\right)}{\exp\left(-\frac{nm_1^2}{2}\right)\exp\left(m_1\sum_{i=1}^n X_i\right)} \\
&= \exp\left(-\tfrac{n}{2}(m_0^2 - m_1^2)\right)\exp\left((m_0 - m_1)\textstyle\sum_{i=1}^n X_i\right)
\end{aligned}$$

The rejection region is then of the form :

$$W = \{V(X_1, ..., X_n) \leq k\} = \left\{(m_0 - m_1)\sum_{i=1}^n X_i \leq k'\right\}\ ,$$

and, since $m_0 < m_1$, W may be further written as $W = \{\sum_{i=1}^n X_i \geq k''\} = \left\{\overline{X} \geq C\right\}$.

In order to obtain the exact value of the threshold C, one should use the value fixed for the value of the test:

$$\alpha = \mathbb{P}_{H_0}(W) = \mathbb{P}_{m=m_0}(\overline{X} \geq C) = \mathbb{P}_{m=m_0}\left(\sqrt{n}(\overline{X} - m_0) \geq \sqrt{n}C\right) = \mathbb{P}(\mathcal{N}(0,1) \geq \sqrt{n}C)\ .$$

For $\alpha = 0.05$, one gets $\sqrt{n}C = 1.64$.

A practical exercice

A politician running as candidate for the elections is interested in estimating the proportion p of electorate which will vote for him. Two possible hypothesis are considered:

$$\begin{cases} H &: \ p = 0,48 \\ K &: \ p = 0,52 \end{cases}$$

The campaign staff runs an opinion poll and interviews a sample of size n. According to the results of the survey, the candidate will decide to intensify his campaign or not.

1. Which are the two possible errors?
2. The statistician of the staff selects H as null hypothesis and fixes the level of the test, $\alpha = 0.05$. Which is the meaning of this choice?

3. The statistician selects critical regions having the form $\mathcal{R}(n) = \{f_n \geq a\}$, where f_n is the proportion of voters which are favorable in the sample. Which is the decision rule if $n = 900$? Compute the power of this test.
4. How is the critical region varying when n increases? It will be explicitly computed for: $n = 900$, 2500 and 10000. Starting from which value of n the power becomes greater than 95%?
5. Which is the decision of the candidate if half of the interviewed sample is favorable when the sample size is, respectively, 900, 2 500 and 10 000? Are these results coherent?
6. Show that the critical regions $\mathcal{R}(n)$ are the optimal decision rules for testing H against K.
N.B. : The binomial distribution may be approximated here by a Gaussian distribution.

1.4 Monotone likelihood ratio

Let us go back to the Gaussian example in the previous section.

1. Let us first remark that we never used the exact value of m_1 in the construction of the test and of the rejecting region. Hence, the hypothesis could have been written as well
$$H_0 : m = m_0 = 0 \ ; \ H_1 : m > m_0$$
2. Also, let us remark that, for some $m' < m_0$,
$$\mathbb{P}_{m'}(W) = \mathbb{P}_{m'}(\bar{(X)} \geq \frac{1.64}{\sqrt{n}}) < \mathbb{P}_{m_0}(\bar{(X)} \geq \frac{1.64}{\sqrt{n}}) = \alpha \ .$$
Hence, if the hypothesis of the test are changed into
$$H_0 : m \leq m_0 = 0 \ ; \ H_1 : m > m_0 \ ,$$
the rejection region remains the same and the level of the test is still $\alpha = \sup_{m \leq m_0} \mathbb{P}_m(W)$.

This example shows that simple hypothesis testing could be easily transformed into composite hypothesis testing. Let us now check if one can always do this extension, which are the necessary conditions and if the resulting test for composite hypothesis is uniformly most powerful (UMP).

Definition: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model,$\Theta \subset \mathbb{R}$. Let us denote
$$V_{\theta_1,\theta_0}(x) = \frac{\mathcal{L}_{\theta_1}(x)}{\mathcal{L}_{\theta_0}(x)}$$
If $T(X)$ is an exhaustive statistic for the parameter of this model, one says that the model has a **monotone likelihood ratio (MLR)** in T whenever $V_{\theta_1,\theta_0}(x)$ (which can be written in terms of $T(x)$ only by exhaustivity) is an increasing function of $T(x)$, for $\theta_1 < \theta_0$.

Remark: If $\theta \mapsto g(\theta)$ is an increasing function, the exponential model $\mathcal{L}_\theta(x) = h(x) \exp(g(\theta)T(x) - B(\theta))$ has a monotone likelihood ratio in T.

Proposition: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model. Let ϕ be a test of level α of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, with $\Theta_0 \cap \Theta_1 = \emptyset$.
If there exists $\theta_0 \in \Theta_0$ such that $\mathbb{E}_{\theta_0}(\phi) = \alpha$ and if, for all $\theta_1 \in \Theta_1$, there exists $k > 0$ such that

- $\phi(x) = 1$, if $\mathcal{L}(x, \theta_1) > k\mathcal{L}(x, \theta_0)$;
- $\phi(x) = 0$, if $\mathcal{L}(x, \theta_1) < k\mathcal{L}(x, \theta_0)$.

Then, ϕ is the uniformly most powerful test of level α, UMP(α).

Proof: Let $\phi^\star$ be a test of level α of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. We consider the following reduced problem of hypothesis testing :
$$H_0' : \theta = \theta_0 \ ; \ H_1' : \theta = \theta_1 \ .$$
For this reduced problem, ϕ is a Neyman-Pearson test of size α. Hence, it is the most powerful test of level α. Furthermore,
$$\mathbb{E}_{\theta_0}(\phi^\star(X)) \leq \sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\phi^\star(X)) \leq \alpha \ ,$$
hence $\mathbb{E}_{\theta_1}(\phi(X)) \geq \mathbb{E}_{\theta_1}(\phi^\star(X))$.

1.4.1 Unilateral tests

Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model, with $\Theta \in \mathbb{R}$.

Unilateral test $H_0 : \theta = \theta_0$ **against** $H_1 : \theta > \theta_0$

Theorem: Suppose the parametric model has a monotone likelihood ratio, strictly increasing in $T(X)$, an exhaustive statistic. For any $\alpha \in]0, 1[$, there exists an UMP(α) test, having the following form:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) > k \\ 0 & , \quad \text{if } T(x) < k \\ c & , \quad \text{if } T(x) = k \end{cases} .$$

Proof: Let $\theta' > \theta_0$. We consider the Neyman-Person test, of size α,

$$H_0 : \theta = \theta_0 \ ; \ H_1' : \theta = \theta' \ .$$

We've already seen that the Neyman-Pearson test may be written as

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) > k \\ 0 & , \quad \text{if } T(x) < k \\ c & , \quad \text{if } T(x) = k \end{cases} .$$

Consider now $\theta_1 > \theta_0$ and since $\frac{\mathcal{L}(x,\theta_1)}{\mathcal{L}(x,\theta_0)} = V_{\theta_1,\theta_0}(T(x))$ with V_{θ_1,θ_0} strictly increasing in $T(x)$, it follows that there exists k_1 such that $T(x) > k$ is equivalent to $\mathcal{L}(x,\theta_1) > k_1\mathcal{L}(x,\theta_0)$ and $T(x) < k$ is equivalent to $\mathcal{L}(x,\theta_1) < k_1\mathcal{L}(x,\theta_0)$. According to the previous proposition, one gets that ϕ is the UMP(α) test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Unilateral test $H_0 : \theta = \theta_0$ **against** $H_1 : \theta < \theta_0$

In a similar manner, one gets the following theorem:

Theorem: Suppose the parametric model has a monotone likelihood ratio, strictly increasing in $T(X)$, an exhaustive statistic. For any $\alpha \in]0, 1[$, there exists an UMP(α) test, having the following form:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) < k \\ 0 & , \quad \text{if } T(x) > k \\ c & , \quad \text{if } T(x) = k \end{cases} .$$

Unilateral test $H_0 : \theta \leq \theta_0$ **against** $H_1 : \theta > \theta_0$

Lehman theorem: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model, with $\Theta \in \mathbb{R}$. Suppose that the model has a monotone likelihood ratio, strictly increasing in $T(X)$, an exhaustive statistic. For any $\alpha \in]0, 1[$, there exists an UMP(α) test, having the following form:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) > k \\ 0 & , \quad \text{if } T(x) < k \\ c & , \quad \text{if } T(x) = k \end{cases} .$$

Furthermore, the size α is reached for $\theta = \theta_0$, that is $\sup_{\theta \leq \theta_0} \mathbb{E}_\theta(\phi(X)) = \mathbb{E}_{\theta_0}(\phi(X)) = \alpha$.

Proof: We already know that there exists a test ϕ, such that $= \mathbb{E}_{\theta_0}(\phi(X)) = \alpha$ and having the form

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) > k \\ 0 & , \quad \text{if } T(x) < k \\ c & , \quad \text{if } T(x) = k \end{cases} .$$

Consider $\theta' < \theta''$. Since $\frac{\mathcal{L}(x,\theta'')}{\mathcal{L}(x,\theta')} = V_{\theta'',\theta'}(T(x))$ with $V_{\theta'',\theta'}$ strictly increasing in $T(x)$, ϕ is a Neyman-Pearson test for the hypothesis

$$H_0' : \theta = \theta' \ ; \ H_1'' : \theta = \theta'' \ .$$

According to the Neyman-Pearson lemma, for all test $\phi^\star$ such that $\mathbb{E}_{\theta'}(\phi^\star(X)) \leq \mathbb{E}_{\theta'}(\phi(X))$, it follows that $\mathbb{E}_{\theta''}(\phi(X)) \geq \mathbb{E}_{\theta''}(\phi^\star(X))$.

Let $\theta' = \theta_0$, $\theta'' = \theta > \theta_0$, and $\phi^\star$ a test of level α for the initial test problem $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Then,

$$\mathbb{E}_{\theta_0}(\phi^\star(X)) \leq \sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\phi^\star(X)) \leq \alpha = \mathbb{E}_{\theta_0}(\phi(X)) \ ,$$

hence $\mathbb{E}_\theta(\phi^\star(X)) \leq \mathbb{E}_\theta(\phi(X))$ and ϕ is more powerful than $\phi^\star$.

It remains to prove that ϕ is of size α for the initial test problem, $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

Consider now $\theta' < \theta_0$, $\theta'' = \theta_0$ and $\phi^\star$ the constant test equal to $\mathbb{E}_{\theta'}(\phi(X))$. Then

$$\mathbb{E}_{\theta'}(\phi^\star(X)) = \mathbb{E}_{\theta'}(\phi(X)) \ \Rightarrow \ \mathbb{E}_{\theta_0}(\phi(X)) \geq \mathbb{E}_{\theta_0}(\phi^\star(X)) \ ,$$

that is $\mathbb{E}_{\theta_0}(\phi(X)) \geq \mathbb{E}_{\theta'}(\phi(X))$. It follows than that, $\forall \theta' < \theta_0$, $\mathbb{E}_{\theta'}(\phi(X)) \leq \mathbb{E}_{\theta_0}(\phi(X)) = \alpha$, and $\sup_{\theta \leq \theta_0} \mathbb{E}_\theta(\phi(X)) = \mathbb{E}_{\theta_0}(\phi(X)) = \alpha$.

Remark: In the previous proof, we showed that ϕ is a Neyman-Pearson test for all hypothesis testing situations of $H_0' : \theta = \theta'$ against $H_1'' : \theta = \theta''$, with $\theta' < \theta''$, hence it is unbiased and $\mathbb{E}_{\theta'}(\phi(X)) < \mathbb{E}_{\theta''}(\phi(X))$ unless $\mathbb{P}_{\theta'} = \mathbb{P}_{\theta''}$ a.e.

Indeed, if $\mathbb{E}_{\theta'}(\phi(X)) = \mathbb{E}_{\theta''}(\phi(X))$ and if $\phi^\star$ is a constant test of size $\mathbb{E}_{\theta'}(\phi(X))$, then $\phi^\star$ has the same power function as ϕ, hence it is UMP($\mathbb{E}_{\theta'}(\phi(X))$). It follows that it is of Neyman-Pearson type and $\mathcal{L}(x, \theta') = k\mathcal{L}(x, \theta'')$, μ-a.e. Since $\mathcal{L}$ is a likelihood, $k = 1$ and $\mathbb{P}_{\theta'} = \mathbb{P}_{\theta''}$.

It also follows that the model is identifiable, the mapping $\theta \mapsto \mathbb{E}_\theta(\phi(X))$ is strictly increasing.

Example:
Consider the observed values $(x_1, ..., x_n)$ of a iid n-sample $(X_1, ..., X_n)$ of a Gaussian distribution, $\mathcal{N}\left(0, \frac{1}{\theta}\right)$, $\theta > 0$. The size of the sample is $n = 15$.

1. Build a UMP(0.05) test of $H_0 : \theta = 1$ against $H_1 : \theta > 1$.
2. Compute the power function of this test.
3. Which is the decision if the observed statistic is $\sum_{i=1}^{15} x_i^2 = 6.8$? For which level of the test α one would have reached to the opposite decision? Which is the p-value associated to this test?

1.4.2 Bilateral tests

Within the framework of exponential parametric models with strictly monotone ratios, one may build "optimal" (in a sense to be defined) bilateral tests.

Theorem : Neyman-Pearson generalized lemma
Consider $\mathbb{P}_1, ..., \mathbb{P}_{m+1}$ probability measures defined on $(E, \mathcal{E})$. Suppose that there exists σ-finite probability measure μ such that $d\mathbb{P}_i = f_i d\mu$, $\forall i = 1, ..., m+1$ (for example, $\mu = \sum_{i=1}^{m+1} \mathbb{P}_i$. Let us denote $\mathbb{E}_i(\phi(X)) = \int_E \phi(x) f_i(x) d\mu(x)$ and let $\mathcal{C}_m$ the set of statistical tests ϕ verifying the constraints:

$$\mathbb{E}_1(\phi(X)) = c_1, \mathbb{E}_2(\phi(X)) = c_2, ..., \mathbb{E}_m(\phi(X)) = c_m,$$

for $c_1, ..., c_m \in \mathbb{R}$, fixed. Then

1. There exists a test $\phi \in \mathcal{C}_m$ which maximizes $\mathbb{E}_{m+1}(\phi(X))$.
2. Every $\phi \in \mathcal{C}_m$ having the expression

   $$\phi(X) = \left\{ \begin{array}{lll} 1 & , & \text{if } f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \\ 0 & , & \text{if } f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \end{array} \right. .$$

   is maximizing $\mathbb{E}_{m+1}(\phi(X))$.
3. Every $\phi \in \mathcal{C}_m$ having the expression

   $$\phi(X) = \left\{ \begin{array}{lll} 1 & , & \text{if } f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \\ 0 & , & \text{if } f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \end{array} \right. ,$$

   with $k_1 \geq 0, ..., k_m \geq 0$ is maximizing $\mathbb{E}_{m+1}(\phi(X))$, among all tests such that

   $$\mathbb{E}_1(\phi(X)) \leq c_1, \mathbb{E}_2(\phi(X)) \leq c_2, ..., \mathbb{E}_m(\phi(X)) \leq c_m \ .$$

4. The set
$$C_m = \{(\mathbb{E}_1(\phi(X)), ..., \mathbb{E}_m(\phi(X))) \;\;,\; \phi : (E, \mathcal{E}) \to ([0,1], \mathcal{B}([0,1]))\}$$
is closed and convex. If $(c_1, ..., c_m) \in \overset{\circ}{C}_m$, there exists a generalized Neyman-Pearson test in $\mathcal{C}_m$, and every test in $\mathcal{C}_m$ maximizing $\mathbb{E}_{m+1}(\phi(X))$ is a generalized Neyman-Pearson test.

Bilateral test $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ against $H_1 : \theta \in]\theta_1, \theta_2[$

Theorem: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model with respect to a probability measure μ, and $\Theta \in \mathbb{R}$. The likelihood is given by

$$\mathcal{L}(x, \theta) = C(\theta) h(x) \exp(\eta(\theta) T(x)) \;.$$

Suppose that $\eta(\theta)$ is a strictly increasing function, hence the model has a monotone likelihood ratio, strictly increasing in $T(x)$. For any $\alpha \in]0, 1[$, there exists a UMP(α) test having the following expression:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } k_1 < T(x) < k_2 \\ 0 & , \quad \text{if } T(x) < k_1 \text{ or } T(x) > k_2 \\ c_1 & , \quad \text{if } T(x) = k_1 \\ c_2 & , \quad \text{if } T(x) = k_2 \end{cases} \;.$$

Furthermore, the size α of ϕ is reached for $\theta = \theta_1$ and $\theta = \theta_2$, that is $\sup_{\theta \leq \theta_1} \mathbb{E}_\theta(\phi(X)) = \mathbb{E}_{\theta_1}(\phi(X)) = \alpha$ and $\sup_{\theta \geq \theta_2} \mathbb{E}_\theta(\phi(X)) = \mathbb{E}_{\theta_2}(\phi(X)) = \alpha$.

Proof: TODO.

Remark: In practice, the issues with applying this test lie mainly in the choice of thresholds k_1 and k_2, such that $\mathbb{E}_{\theta_1}(\phi(X)) = \mathbb{E}_{\theta_2}(\phi(X)) = \alpha$.

Bilateral test $H_0 : \theta \neq \theta_0$ against $H_1 : \theta = \theta_0$

This problem is in some sense similar to the previous one. One may show, in a similar manner, that, in the framework of general exponential models, there exists a UMP(α) test.

Theorem: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model with respect to a probability measure μ, and $\Theta \in \mathbb{R}$. The likelihood is given by

$$\mathcal{L}(x, \theta) = C(\theta) h(x) \exp(\eta(\theta) T(x)) \;.$$

Suppose that $\eta(\theta)$ is a strictly increasing function, hence the model has a monotone likelihood ratio, strictly increasing in $T(x)$. For any $\alpha \in]0, 1[$, there exists a UMP(α) test having the following expression:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } k_1 < T(x) < k_2 \\ 0 & , \quad \text{if } T(x) < k_1 \text{ or } T(x) > k_2 \\ c_1 & , \quad \text{if } T(x) = k_1 \\ c_2 & , \quad \text{if } T(x) = k_2 \end{cases} \;.$$

Furthermore, the size α of ϕ is reached for $\theta = \theta_0$ and the thresholds k_1 and k_2 are computed using the following equations

$$\begin{cases} \mathbb{E}_{\theta_0}(\phi(X)) = \alpha \\ \mathbb{E}_{\theta_0}(\phi(X) T(X)) = \alpha \mathbb{E}_{\theta_0}(T(X)) \;. \end{cases}$$

Remark: The computation of k_1 and k_2 is simplified if the probability distribution of $T(X)$ is symmetric whenever $X \sim \mathbb{P}_{\theta_0}$. If one selects a test $\phi = h(T)$, with h symmetric with respect to a (i.e. $(k_1 + k_2)/2 = a$ and $c_1 = c_2 = c$), and such that $\mathbb{E}_{\theta_0}(\phi(X)) = \alpha$, then

$$\mathbb{E}_{\theta_0}(\phi(X) T(X)) = \mathbb{E}_{\theta_0}((T(X) - a) h(T(X))) + a \mathbb{E}_{\theta_0}(\phi(X)) = a\alpha = \alpha \mathbb{E}_{\theta_0}(T(X)) \;,$$

and the second equation is verified.

Bilateral test $H_0 : \theta \in [\theta_1, \theta_2]$ **against** $H_1 : \theta < \theta_1$ **or** $\theta > \theta_2$

Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model with respect to a probability measure μ, and $\Theta \in \mathbb{R}$. Let us also suppose that the model has a monotone likelihood ratio, strictly increasing in $T(x)$. We shall now show that there does not exist a UMP(α) test.

Indeed, if such a test ϕ existed, then, for any other test $\phi^\star$ such that $\sup_{\theta \in [\theta_1, \theta_2]} \mathbb{E}_\theta(\phi^\star(X)) \leq \alpha$,

$$\mathbb{E}_\theta(\phi(X)) \geq \mathbb{E}_\theta(\phi^\star(X)) \text{ , for all } \theta > \theta_2 \text{ or } \theta < \theta_1 \text{ .}$$

Then ϕ would also be UMP(α) for the test problem $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_1' : \theta < \theta_1$ or against $H_1' : \theta > \theta_2$. Following the remark given after the Lehmann theorem, one would deduce that the mapping $\theta \mapsto \mathbb{E}_\theta(\phi(X))$ is strictly decreasing on $\Theta \cap] - \infty, \theta_2]$, and strictly increasing on $\Theta \cap [\theta_1, +\infty[$, which is impossible.

Then, we shall look for "optimal" tests in a more restricted class than that of tests with a fixed level α.

Definition: A statistical test ϕ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is said to be **unbiased of level** α if, for all $\theta_0 \in \Theta_0$, $\mathbb{E}_{\theta_0}(\phi(X)) \leq \alpha$ and, for all $\theta_1 \in \Theta_1$, $\mathbb{E}_{\theta_1}(\phi(X)) \geq \alpha$.

Definition: A statistical test ϕ is said to be **uniformly most powerful among all unbiased tests of level** α**, UMPU(**α**)** if it is unbiased of level α and if it is uniformly most powerful than any other unbiased test of level α.

Theorem: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model with respect to a probability measure μ, and $\Theta \in \mathbb{R}$. The likelihood is given by

$$\mathcal{L}(x, \theta) = C(\theta) h(x) \exp(\eta(\theta) T(x)) \text{ .}$$

Suppose that $\eta(\theta)$ is a strictly increasing function, hence the model has a monotone likelihood ratio, strictly increasing in $T(x)$. For any $\alpha \in]0, 1[$, there exists a UMPU(α) test having the following expression:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) < k_1 \text{ or } T(x) > k_2 \\ 0 & , \quad \text{if } k_1 < T(x) < k_2 \\ c_1 & , \quad \text{if } T(x) = k_1 \\ c_2 & , \quad \text{if } T(x) = k_2 \end{cases} \text{ .}$$

Furthermore, the size α of ϕ is reached for $\theta = \theta_1$ and $\theta = \theta_2$, that is

$$\sup_{\theta \in [\theta_1, \theta_2]} \mathbb{E}_\theta(\phi(X)) = \mathbb{E}_{\theta_1}(\phi(X)) = \mathbb{E}_{\theta_2}(\phi(X)) = \alpha \text{ .}$$

Proof: TODO.

Bilateral test $H_0 : \theta = \theta_0$ **against** $H_1 : \theta \neq \theta_0$

This problem is close to the previous one. Hence, we may show that in this case there does not exist a UMP(α) test. However, we state the existence of a UMPU(α) test.

Theorem: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model with respect to a probability measure μ, and $\Theta \in \mathbb{R}$. The likelihood is given by

$$\mathcal{L}(x, \theta) = C(\theta) h(x) \exp(\eta(\theta) T(x)) \text{ .}$$

Suppose that $\eta(\theta)$ is a strictly increasing function, hence the model has a monotone likelihood ratio, strictly increasing in $T(x)$. For any $\alpha \in]0, 1[$, there exists a UMPU(α) test having the following expression:

$$\phi_{k,c}(X) = \begin{cases} 1 & , \quad \text{if } T(x) < k_1 \text{ or } T(x) > k_2 \\ 0 & , \quad \text{if } k_1 < T(x) < k_2 \\ c_1 & , \quad \text{if } T(x) = k_1 \\ c_2 & , \quad \text{if } T(x) = k_2 \end{cases} \text{ .}$$

Furthermore, the size α of ϕ is reached for $\theta = \theta_0$ and the thresholds k_1 and k_2 are computed using the following equations

$$\begin{cases} \mathbb{E}_{\theta_0}(\phi(X)) = \alpha \\ \mathbb{E}_{\theta_0}(\phi(X) T(X)) = \alpha \mathbb{E}_{\theta_0}(T(X)) \text{ .} \end{cases}$$

Remark: Here also, the practical computation of the thresholds k_1 and k_2 is greatly simplified whenever the probability distribution of $T(X)$ is symmetric when $X \sim \mathbb{P}_{\theta_0}$.

Example: Consider two positive numbers θ_0 and θ_1 ($\theta_0 < \theta_1$) and an n-sample of a probability distribution characterized by the following density:

$$f(x) = \begin{cases} \theta^2 x e^{-\theta x} & , \quad x > 0 \\ 0 & , \quad x \leq 0 \end{cases}$$

Show that, for all $\alpha \in [0, 1]$, there exists a most powerful test of level α for testing "$\theta = \theta_0$" against "$\theta = \theta_1$". Is this test unbiased? Propose a statistical test for the hypothesis "$\theta = \theta_0$" against "$\theta > \theta_0$", or against "$\theta \neq \theta_0$"? Which are the properties of these tests?

1.4.3 Exercices

Exercice 1: Consider $X = (X_1, ..., X_n)$ an n-sample of iid random variables, issued from a uniform distribution, $\mathcal{U}[0, \theta]$ and let $\alpha \in]0, 1[$.

1. Consider ϕ a statistical test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Prove that ϕ is uniformly most powerful among all tests of level α if and only if $\mathbb{E}_{\theta_0}(\phi(X)) = \alpha$ and $\phi(X) = 1$ whenever $\max(X_1, ..., X_n) > \theta_0$.
2. Prove that there exists a uniformly most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, of level α and given by

$$\phi(X) = \begin{cases} 1 & , \quad \max(X_1, ..., X_n) > \theta_0 \text{ or } \min(X_1, ..., X_n) \leq \theta_0 \alpha^{1/n} \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Exercice 2: Consider $X = (X_1, ..., X_n)$ an n-sample of iid random variables having as density with respect to the Lebesgue measure:

$$f_{(a,b)}(x) = a \exp(-a(x - b)) \mathbb{1}_{(x \geq b)} \ .$$

1. Suppose that a is known, fixed.
   (a) Compute the probability distribution of $Y_i = \exp(-aX_i)$.
   (b) Show that there exists a unique uniformly most powerful test of $H_0 : b = b_0$ against $H_1 : b \neq b_0$. Describe the details of this test.
2. Show that there exists a uniformly most powerful test of $H_0 : a = a_0, b = b_0$, against $H_1 : a > a_0, b < b_0$. Describe the details of this test.

1.5 Likelihood ratio tests (LRT) and asymptotic LRT

As illustrated in the previous sections, the likelihood is a valuable tool in hypothesis testing. Another way of building a test is to maximize the likelihood ratio.

Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric dominated model with respect to a probability measure μ. Consider also the test problem $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\Theta_0, \Theta_1 \subset \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

The first way of defining the test is to consider the ratio

$$T(x) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(x, \theta)}{\sup_{\theta \in \Theta_1} \mathcal{L}(x, \theta)} \ ,$$

and define a rejection region as $W = \{T(x) < k\}$.

One may show that this test is also equivalent to the maximum likelihood ratio test (LRT), defined by the test statistic

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(x, \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(x, \theta)} \ ,$$

The LRT decision rule may be written as $\phi(x) = \mathbb{1}_{\{\Lambda(x) \leq k\}}$.

Remark: The LRT statistic $\Lambda(x)$ is related to the maximum likelihood estimate. Indeed, if $\hat{\theta}$ is the ML estimate of θ and $\hat{\theta}_0$ is the ML estimate on the restricted parameter set Θ_0, then $\Lambda(x) = \frac{\mathcal{L}(x,\hat{\theta}_0)}{\mathcal{L}(x,\hat{\theta})}$.

Remark: Even if in the simple cases (simple hypothesis), the LRT is equivalent to the optimal tests introduced in the previous sections, the LRT is, in general, not necessarily optimal!

The effective computation of $\Lambda(x)$ is greatly simplified in practice if there is an exhaustive statistic, available for the considered model.

Theorem: Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$ a parametric model, dominated by some probability measure μ, with likelihood $\mathcal{L}(x,\theta)$. If T is an exhaustive statistic for θ, then, for any $\Theta_0 \subset \Theta$, the LRT statistic $\Lambda(x)$ is factorizing across T, that is there exists a real function $\tilde{\lambda}$ such that, for all x, $\Lambda(x) = \tilde{\lambda}(T(x))$.

In the case where the probability distribution of $\Lambda(X)$ is not easily tractable, one may use the following asymptotic result:

Theorem (Wilks): Consider $(E, \mathcal{E}, \mathbb{P}_\theta, \theta \in \Theta)$, $\Theta \in \mathbb{R}^p$, a parametric **regular** model, dominated by some probability measure μ, with likelihood $\mathcal{L}(x,\theta)$. Consider also the test problem $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Then, with the notation

$$\Lambda_n = \frac{\mathcal{L}(X_1, ..., X_n, \theta_0)}{\sup_{\theta \in \Theta} \mathcal{L}(X_1, ..., X_n, \theta)} \ ,$$

one has that

$$-2 \ln \Lambda_n \xrightarrow[n \to \infty]{(l)} \chi^2(p) \ .$$

The rejection region of this asymptotic test is then $W = \{-\ln \Lambda_n > q_{1-\alpha}\}$, where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the $\chi^2(p)$ distribution. Furthermore, the sequence of tests given by Λ_n has a power asymptotically equal to 1, for a fixed level α.

Example:
Let $(X_1, ..., X_n)$ be an n-sample of $\mathcal{N}(m, 1)$, where $m \in \mathbb{R}$ is the unknown parameter. Compute the LRT test of level $\alpha = 0.05$ in the following cases

1. $H_0 : m = 0$ against $H_1 : m = 2$.
2. $H_0 : m = 0$ against $H_1 : m \neq 0$.

1.5.1 Exercices

Exercice 1: Consider $\tau_1, \tau_2, ..., \tau_m, ...$ independent random variables distributed according to $\mathcal{E}(\theta)$, $\theta > 0$ an unknown parameter. For all $n \in \mathbb{N}^\star$, let us denote $T_n = \tau_1 + ... + \tau_n$. Design a uniformly most powerful test of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, of level $\alpha = 0.05$, using the following :

1. the available statistic is $N_t = \sum_{n \geq 1} \mathbb{1}_{(T_n \leq t)}$, for a fixed $t > 0$ (hint: show that N_t is distributed according to a Poisson $\mathcal{P}(\theta t)$).
2. the available statistics are $(T_1, ..., T_m)$, where $m \in \mathbb{N}^\star$ is fixed.

Exercice 2: The annual income of the individuals in a given population is a random variable, distributed according to a Pareto law, with density

$$f \ : \ x \mapsto \frac{a k^a}{x^{a+1}} \mathbb{1}_{[k, +\infty[}(x) \ , \ a, k > 0.$$

1. Estimate the parameters of the considered density using the maximum likelihood procedure.
2. One wishes to test $H_0 : a = 1$ against $H_1 : a \neq 1$. Compute the likelihood ratio test for a given level, $\alpha = 0.05$.

1.6 Wald test

When performing statistical hypothesis testing, the maximum likelihood estimate is a valuable tool. However, the difficulty consists in computing exactly the probability distribution of the ML estimate, $\hat{\theta}_n$, for a fixed sample size, n. If this is possible, then $\hat{\theta}_n$ may be used as test statistic.

Otherwise, more generally, the asymptotic distribution of $\hat{\theta}_n$ is known under regularity conditions. Hence, when n is sufficiently large, the probability distribution of $\hat{\theta}_n$ may be approximated by a Gaussian. However, this is quite tricky: the asymptotic covariance matrix, which is the inverse of the Fisher information matrix, depends on the parameter θ. The following test statistic $\hat{T}$ will be then preferred instead:

Definition: Consider a **regular** parametric dominated statistical model $(E^n, \mathcal{E}_n, \mathbb{P}_\theta, \theta \in \Theta)$, where $\Theta \subset \mathbb{R}^p$. The **Wald statistic** for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$ is

$$\hat{T}_n = n \cdot (\hat{\theta}_n - \theta)^t \cdot I(\theta) \cdot (\hat{\theta}_n - \theta),$$

where $\hat{\theta}_n$ is the maximum likelihood estimate and $I(\theta)$ is the Fisher information matrix computed for one random variable only.

In order to prove the relevance of this test, let us next consider the sequence $(\hat{T}_n)$ in the "big" asymptotic model.

Theorem: Given the regular dominated parametric model $\left(E^{\mathbb{N}}, \mathcal{E}_{\mathbb{N}}, (p_\theta \cdot d\mu)^{\otimes \mathbb{N}}, \theta \in \Theta\right)$, where μ is the dominating measure, and given the hypothesis testing problem $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, the Wald test statistic $\hat{T}_n$ for the projected model of size n and under the hypothesis H_0 is convergent in law:

$$\hat{T}_n \xrightarrow[n\to\infty]{(l)} \chi^2(p) \ .$$

The asymptotic rejection region will have the form $W_n = \left\{\hat{T}_n > q_{1-\alpha}\right\}$, where $q_{1-\alpha}$ is the $1-\alpha$-quantile of the $\chi^2(p)$ distribution. The power of the sequence of tests with rejection regions W_n is converging to 1, when the level α is fixed.

Proof: The asymptotic probability distribution of $\hat{\theta}_n$ induces that of $\hat{T}_n$, since $\sqrt{n} \cdot I(\theta)^{\frac{1}{2}} \cdot (\hat{\theta}_n - \theta)$ is asymptotically distributed according to $\mathcal{N}(0, I_p)$ under the hypothesis H_0 and $\hat{T}_n = \|\sqrt{n} \cdot I(\theta)^{\frac{1}{2}} \cdot (\hat{\theta}_n - \theta)\|^2$.

An exercice: Let $(X_1, ..., X_n)$ be an n-sample of $\mathcal{N}(m, 1)$, where $m \in \mathbb{R}$ is the unknown parameter. Compute the Wald test of level $\alpha = 0.05$ in the following cases

1. $H_0 : m = 0$ against $H_1 : m = 2$.
2. $H_0 : m = 0$ against $H_1 : m \neq 0$.

Chapter 2

Gaussian tests

2.1 One-sample tests

2.1.1 Testing for the mean of a Gaussian distribution

Consider $(X_1, ..., X_n)$, an n-sample from a Gaussian distribution, $\mathcal{N}(m, \sigma^2)$. For defining the following test statistics, we use the unbiased empirical estimates for the mean and the variance:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ ; \ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Hypothesis testing on m, with σ^2 known

Under H_0, the test statistic verifies

$$\frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

Hypothesis testing on m, with σ^2 unknown and n small

Under H_0, the test statistic verifies

$$\frac{\overline{X} - m_0}{\frac{S}{\sqrt{n}}} \sim St(n-1)$$

Hypothesis testing on m, with σ^2 unknown and n large

Under H_0, the test statistic verifies

$$\frac{\overline{X} - m_0}{\frac{S}{\sqrt{n}}} \xrightarrow[n\to\infty]{(l)} \mathcal{N}(0, 1)$$

Example:
A plant is manufacturing fluorescent tubes whose lifespan X, expressed in hours, is a random variable distributed according to a Gaussian $\mathcal{N}(m, \sigma^2)$. In order to have an acceptable output, m has to be equal to 450 and σ equal to 10. A random sample of tubes of size 16 is drawn and their lifespans are recorded. The resulting estimation for the empirical mean and the empirical variance are

$$\overline{x} = \frac{1}{16}\sum_{i=1}^{16} x_i = 454 \ ; \ s^2 = \frac{1}{15}\sum_{i=1}^{15}(x_i - \overline{x})^2 = 121 \ .$$

For a fixed level $\alpha = 0.1$, test the statistical hypothesis $H_0 : m = 450$ against $H_1 : m \neq 450$. Which is the statistical decision?

2.1.2 Testing for the variance of a Gaussian distribution

With the framework defined in the previous subsection:

Hypothesis testing on σ^2, with m known

Under H_0, the test statistic verifies

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - m)^2 \sim \chi^2(n).$$

Hypothesis testing on σ^2, with m unknown

Under H_0, the test statistic verifies

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sim \chi^2(n-1).$$

If n is large, the following approximations may be used:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \xrightarrow{\cong} \mathcal{N}\left(\sigma^2, \sqrt{\frac{2\sigma^4}{n}}\right)$$

and

$$\sqrt{2\chi^2(n)} \xrightarrow{\cong} \mathcal{N}\left(\sqrt{2n-1}, 1\right)$$

<u>**Example**</u>:
Using the same framework as in the previous example and for a fixed level $\alpha = 0.1$, test the statistical hypothesis $H_0 : \sigma = 10$ against $H_1 : \sigma \neq 10$. Which is the statistical decision?

2.1.3 Testing for a theoretical proportion

Consider $(X_1, ..., X_n)$, an n-sample from a Bernoulli distribution, $\mathcal{B}(p)$, $0 < p < 1$. For defining the following test statistic, we use the empirical frequency, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Under $H_0 : p = p_0$,

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \xrightarrow[n\to\infty]{(l)} \mathcal{N}(0,1).$$

<u>**Example**</u>:
Several studies on the motor development of children showed that 50% of the babies are walking at the age of 12 months. We wish to investigate the possible delay of preterm babies in acquiring the capacity of walking. The hypothesis to be tested is that preterm babies walk later than normal babies. A sample of 80 preterm babies is drawn at random in the general population. Among them, 35 are walking at 12 months. For a fixed level of the test, $\alpha = 0.05$, may we accept the former hypothesis? Which is the p-value associated to this test?

2.2 Two-samples tests

Consider two samples $(X_1, ..., X_{n_1})$ and $(Y_1, ..., Y_{n_2})$ issued from two Gaussian distributions, $\mathcal{N}\left(m_1, \sigma_1^2\right)$ and $\mathcal{N}\left(m_2, \sigma_2^2\right)$. Define the unbiased empirical estimates for the means and the variances in the two samples :

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \; ; \; S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2 \; ;$$

$$\overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j \; ; \; S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_j - \overline{Y})^2 \; .$$

2.2.1 Testing for the equality of the variances

Testing for $\sigma_1 = \sigma_2$, with m_1 and m_2 known

Under H_0, the test statistic verifies

$$\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - m_1)^2}{\frac{1}{n_2}\sum_{j=1}^{n_2}(Y_j - m_2)^2} \sim \mathcal{F}(n_1, n_2)$$

Testing for $\sigma_1 = \sigma_2$, with m_1 and m_2 unknown

Under H_0, the test statistic verifies

$$\frac{\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_i-\overline{X}\right)^2}{\frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(Y_j-\overline{Y}\right)^2} \sim \mathcal{F}\left(n_1-1, n_2-1\right)$$

<u>**Example**</u>:
We are interested in the lifespan of two types of devices, A and B. The lifespan is denoted by X and is considered to be a random variable, distributed according to a Gaussian distribution, $\mathcal{N}(m_A, \sigma_A^2)$ for devices A and $\mathcal{N}(m_B, \sigma_B^2)$ for devices B.
In order to make a statistical test, two samples are randomly chosen, of sizes $n_A = 30$ and $n_B = 60$, and we compute the following estimations (measured in hours):

$\overline{x}_A = 2\ 000$	the estimated empirical mean in sample A;
$\overline{x}_B = 2\ 200$	the estimated empirical mean in sample B;
$s_A = 300$	the estimated empirical standard deviation in sample A;
$s_B = 360$	the estimated empirical standard deviation in sample B.

Test the equality of the variances in the two populations, for a fixed level $\alpha = 0.05$.

2.2.2 Testing for the equality of the means

Testing for $m_1 = m_2$, with σ_1^2 and σ_2^2 known

Under H_0, the test statistic verifies

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

Testing for $m_1 = m_2$, with σ_1^2 and σ_2^2 unknown, n_1 and n_2 large

Under H_0, the test statistic verifies

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \xrightarrow[n\to\infty]{(l)} \mathcal{N}(0,1)$$

Testing for $m_1 = m_2$, with σ_1^2 and σ_2^2 unknown, but equal

Under H_0, the test statistic verifies

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{\sum_{i=1}^{n_1}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{n_2}\left(Y_j-\overline{Y}\right)^2}{n_1+n_2-2}}} \sim St\left(n_1+n_2-2\right)$$

<u>**Example**</u>:
Considering the test from the previous example states the equality of variances, may one conclude that the mean lifespan of devices B is significantly larger than the mean lifespan of devices A, for a fixed level $\alpha = 0.05$?

2.2.3 Testing for the correlation coefficient (Pearson test)

In this section, let us consider (X_1, Y_1),...,(X_n, Y_n), a n-sample of iid random vectors, distributed according to a Gaussian $\mathcal{N}\left(\begin{pmatrix} m_X \\ m_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$. The parameter of this model is $\theta = \left(m_X, m_Y, \sigma_X^2, \sigma_Y^2, \rho\right) \in \Theta \subset \mathbb{R}^5$.

In this framework, testing $\rho = 0$ against $\rho \neq 0$ is equivalent to testing the independence of X and Y.

When writing the log-likelihood of the n-sample, one gets

$$\begin{aligned}\ln \mathcal{L}(X_1^n, Y_1^n, \theta) &= -n\ln(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}) \\ &\quad -\frac{1}{2(1-\rho^2)}\left(\frac{1}{\sigma_X^2}\sum_{i=1}^n(X_i-m_X)^2 - \frac{2\rho}{\sigma_X\sigma_Y}\sum_{i=1}^n(X_i-m_X)(Y_i-m_Y) + \frac{1}{\sigma_Y^2}\sum_{i=1}^n(Y_i-m_Y)^2\right)\end{aligned}$$

After maximizing the log-likelihood, one gets the ML estimates for θ:

$$\hat{m}_X = \overline{X} \ ; \ \hat{m}_Y = \overline{Y} \ ; \ \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ ; \ \hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

and $\hat{\rho} = \frac{1}{\hat{\sigma}_X \hat{\sigma}_Y}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$.

Under $H_0 : \Theta_0 = \{\theta \in \Theta \mid \rho = 0\}$, the ML estimate is $\hat{\theta}_0 = (\overline{X}, \overline{Y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, 0)$.

The log-likelihood ratio may then be written as follows

$$\ln \lambda(X_1^n, Y_1^n) = \ln \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(X_1^n, Y_1^n, \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(X_1^n, Y_1^n, \theta)} = \ln \frac{\mathcal{L}(X_1^n, Y_1^n, \hat{\theta}_0)}{\mathcal{L}(X_1^n, Y_1^n, \hat{\theta})} = \frac{n}{2}\ln(1 - \hat{\rho}^2) \ .$$

One may remark that the likelihood ratio is a decreasing function of $|\hat{\rho}|$, but also of $|T_n|$, where $T_n = \frac{\sqrt{n-2}\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$. If $\rho = 0$, one may show that $T_n \sim St(n-2)$ and derive the corresponding rejection region.

<u>**Example**</u>: In the following table, the cylinder capacity and the power are reported for 28 car brands. The estimated correlation coefficient between the two is equal to 0.9475. Is this statistically significant, for a given level $\alpha = 0.05$?

	A	B	C	D	E	F	G
1	Numero	Modele	Cylindree	Puissance	XY	X²	Y²
2	1	Daihatsu Cuore	846	32	27072	715716	1024
3	2	Suzuki Swift 1.0 GLS	993	39	38727	986049	1521
4	3	Fiat Panda Mambo L	899	29	26071	808201	841
5	4	VW Polo 1.4 60	1390	44	61160	1932100	1936
6	5	Opel Corsa 1.2i Eco	1195	33	39435	1428025	1089
7	6	Subaru Vivio 4WD	658	32	21056	432964	1024
8	7	Toyota Corolla	1331	55	73205	1771561	3025
9	8	Opel Astra 1.6i 16V	1597	74	118178	2550409	5476
10	9	Peugeot 306 XS 108	1761	74	130314	3101121	5476
11	10	Renault Safrane 2.2. V	2165	101	218665	4687225	10201
12	11	Seat Ibiza 2.0 GTI	1983	85	168555	3932289	7225
13	12	VW Golf 2.0 GTI	1984	85	168640	3936256	7225
14	13	Citroen ZX Volcane	1998	89	177822	3992004	7921
15	14	Fiat Tempra 1.6 Liberty	1580	65	102700	2496400	4225
16	15	Ford Escort 1.4i PT	1390	54	75060	1932100	2916
17	16	Honda Civic Joker 1.4	1396	66	92136	1948816	4356
18	17	Volvo 850 2.5	2435	106	258110	5929225	11236
19	18	Ford Fiesta 1.2 Zetec	1242	55	68310	1542564	3025
20	19	Hyundai Sonata 3000	2972	107	318004	8832784	11449
21	20	Lancia K 3.0 LS	2958	150	443700	8749764	22500
22	21	Mazda Hachtback V	2497	122	304634	6235009	14884
23	22	Mitsubishi Galant	1998	66	131868	3992004	4356
24	23	Opel Omega 2.5i V6	2496	125	312000	6230016	15625
25	24	Peugeot 806 2.0	1998	89	177822	3992004	7921
26	25	Nissan Primera 2.0	1997	92	183724	3988009	8464
27	26	Seat Alhambra 2.0	1984	85	168640	3936256	7225
28	27	Toyota Previa salon	2438	97	236486	5943844	9409
29	28	Volvo 960 Kombi aut	2473	125	309125	6115729	15625
30	n		Moyenne			Somme	
31	28		1809.07	77.71	4451219	102138444	197200
32							
33			Numérateur	614679.571			
34			Dénominateur	643168.291			
35			**Corrélation**	**0.9475**			
36							
37			Coef Corr. Excel	0.9475			

Figure 2.1: Relation between cylinder capacity and power (`eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf`)

2.2.4 Testing for the equality of two proportions

Consider two samples $(X_1, ..., X_{n_1})$ and $(Y_1, ..., Y_{n_2})$ issued from two Bernoulli distributions, $\mathcal{B}(p_1)$ and $\mathcal{B}(p_2)$. Define the empirical frequencies in the two samples

$$\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \ ; \ \hat{p}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j \ ,$$

as well as the overall frequency $\hat{f} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} Y_j\right) = \frac{1}{n_1+n_2}(n_1\hat{p}_1 + n_2\hat{p}_2)$.

Under $H_0 : p_1 = p_2$,

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{f}(1-\hat{f})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \xrightarrow[n\to\infty]{(l)} \mathcal{N}(0,1).$$

Example:
A teaching assistant wishes to test whether the fact of students attending the class **improves** their results on the exam. For this, he considers a random sample of 100 students with the following characteristics :

	Exam passed	Exam missed	Overall
Attend the class	28	12	40
Don't attend the class	33	27	60
Overall	61	39	100

Formalize the wording of the exercice; which is the decision rule for a fixed level $\alpha = 0.1$? Which is the decision of the teaching assistant given the observed values in the sample?

2.3 The linear model

2.3.1 Analysis of variance

We consider here k independent samples, $(X_{ij})_{i=1,...,k;j=1,...,n_i}$ of respective sizes n_i for $i = 1, ..., k$, issued respectively from k Gaussian distributions, $X_{ij} \sim \mathcal{N}(m_j, \sigma^2)$. The variances are supposed to be equal (**homoscedasticity**).

The issue here is to test the equality of the means within the k groups :
$H_0 : m_1 = ... = m_k = m$ against $H_1 : \exists i \neq i'$ such that $m_i \neq m_{i'}$.

In order to do so, we need to introduce the empirical means within the k groups and on all groups

$$\overline{X}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \ , \ \forall i = 1, ..., k \ ; \ \overline{X}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} \ ,$$

where $n = n_1 + ... + n_k$. With the above assumptions, one may remark that, under H_0, $\overline{X}_{i\bullet} \sim \mathcal{N}(m, \frac{\sigma^2}{n_i})$, $\forall i = 1, ..., k$ and $\overline{X}_{\bullet\bullet} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$.

Let us remark that we may alternatively write the model in a linear form

$$X_{ij} = m_i + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \ , \ \forall j = 1, ..., n_i \ , \ i = 1, ..., k \ .$$

Furthermore, with the notation $m_i = \mu + \alpha_i$, where μ is the main, general effect and α_i is the specific effect of factor i, and with the constraint $\sum_{i=1}^{k} n_i\alpha_i = 0$. With these notations, the previous hypothesis may be written
$H_0 : \alpha_1 = ... = \alpha_k = 0$ against $H_1 : \exists i \in \{1, ..., k\}$ such that $\alpha_i \neq 0$.

Under H_0, $X_{ij} = \mu + \varepsilon_{ij}$ and, using least squares (or ML) estimates, one gets that $\hat{\mu} = \overline{X}_{\bullet\bullet}$ and the forecast $\hat{X}_{ij} = \hat{\mu}$. Then, the residual variance or the residual sum of squares is

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{\bullet\bullet}\right)^2 ,$$

which may be further decomposed into the within sum of squares (SSW) and the between sum of squares (SSB):

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{i\bullet}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet}\right)^2 = SSW + SSB \ .$$

By remarking that, under H_0, $SSW \sim \sigma^2\chi^2(n-k)$ and $SSB \sim \sigma^2\chi^2(k-1)$, one may easily derive the test statistic and its probability distribution

$$F^2 = \frac{SSB/(k-1)}{SSW/(n-k)} \sim_{H_0} \mathcal{F}(k-1, n-k) \ .$$

Example: A school wishes to test if there is an examiner effect on the results obtained by the candidates at the entrance examination. For each of the three examiners of the school, a small sample of results was selected, resumed in the following table :

Examiner	A	B	C
Grades	10, 11, 11 12, 13, 15	8, 11, 11, 13 14, 15, 16, 16	10, 13, 14, 14 15, 16, 16
Sample size	6	8	7

Under the regular assumption (independent samples, Gaussian distribution, homoscedasticity), make a statistical test in order to check the existence of an examiner effect.

Chapter 3

Nonparametric hypothesis testing

3.1 χ^2-tests

The χ^2 test was originally designed for checking assumptions on the goodness-of-fit of a probability distribution, but it may also be used for other purposes, such as checking whether two variables are independent or if two random samples are issued from the same distribution. The test is essentially based on an asymptotic property of the multinomial distribution.

Consider $(X_1, ..., X_n)$ an iid n-sample issued from the probability model $(E, \mathcal{E}, P)$. Let us also consider a partition of E, $\{E_1, ..., E_m\}$, such that $E = \cup_{k=1}^m E_k$ and $E_k \cap E_l = \emptyset$, $\forall k \neq l \in \{1, ..., m\}$.

Next, let us define, for all $k \in \{1, ..., m\}$,

$$N_k(n) = \sum_{i=1}^n \mathbb{1}_{X_i \in E_k} \ .$$

If $p_k = \mathbb{P}(X_i \in E_k)$, $\forall k \in \{1, ..., m\}$, then the random vector $(N_1(n), ..., N_m(n))$ is distributed according to a multinomial distribution, $\mathcal{M}(n; p_1, ..., p_m)$:

$$\mathbb{P}(N_1(n) = n_1, ..., N_m(n) = n_m) = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m} \ .$$

Let us also denote $\mathbb{P}_n$ the empirical distribution allowing to approximate $\mathbb{P}$ based on the sample $(X_1, ..., X_n)$: $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$.

Definition: The **(pseudo) χ^2-distance** between $\mathbb{P}_n$ and $\mathbb{P}$ is defined by

$$D(\mathbb{P}_n, \mathbb{P}) = \sum_{k=1}^m \frac{(N_k(n) - np_k)^2}{np_k} \ .$$

Theorem: The following asymptotic result holds for $D(\mathbb{P}_n, \mathbb{P})$,

$$D(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n \to \infty]{(l)} \chi^2(m-1) \ .$$

Proof: In order to prove this theorem, we shall use a central limit theorem for random vectors. Let us first introduce the random vectors $Y_i = (\mathbb{1}_{X_i \in E_1}, ..., \mathbb{1}_{X_i \in E_m})$, for all $i = 1, ..., n$. Then, $N(n) = \sum_{i=1}^n Y_i$ and

$$Cov(Y_{i,k}, Y_{i,l}) = \mathbb{E}\left(\mathbb{1}_{X_i \in E_k} \mathbb{1}_{X_i \in E_l}\right) - p_k p_l = \delta_k^l p_k - p_k p_l \ .$$

The covariance matrix of Y_i is then given by $\Sigma = \Delta_\pi - \pi\pi'$, where $\pi' = (p_1, ..., p_m)$ and Δ_π is a diagonal matrix, whose diagonal is equal to π elements. The CLT for vectors states that

$$\frac{N(n) - n\pi}{\sqrt{n}} \xrightarrow[n \to \infty]{(l)} \mathcal{N}_m(0, \Sigma) \ .$$

If we consider now $f : \mathbb{R}^m \to \mathbb{R}_+$ such that $f(t) = \sum_{k=1}^m \frac{t_k^2}{p_k}$, then $D(\mathbb{P}_n, \mathbb{P}) = f\left(\frac{N(n) - n\pi}{\sqrt{n}}\right)$, hence

$$D(\mathbb{P}_n, \mathbb{P}) \xrightarrow[n \to \infty]{(l)} f(Z) \ , \ Z = (Z_1, ..., Z_m) \sim \mathcal{N}_m(0, \Sigma) \ .$$

It remains to derive the probability distribution of $f(Z)$?
First, let us remark that $f(Z) = \|AZ\|^2$, with

$$A = \begin{pmatrix} \frac{1}{\sqrt{p_1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{p_m}} \end{pmatrix} .$$

For any orthogonal transformation $U : \mathbb{R}^m \to \mathbb{R}^m$, $f(Z) = \|AZ\|^2 = \|UAZ\|^2$. But, AZ is distributed according to a centered Gaussian distribution with covariance matrix $A\Sigma A' = I_m - \sqrt{\pi}\sqrt{\pi}'$, hence UAZ is distributed according to a centered Gaussian distribution with covariance matrix $U(I_m - \sqrt{\pi}\sqrt{\pi}')U' = I_m - (U\sqrt{\pi})(U\sqrt{\pi})'$. By considering U such that $U\sqrt{\pi} = (0, ..., 0, 1)'$ (this is possible since $\|\sqrt{pi}\| = 1$), then

$$U(I_m - \sqrt{\pi}\sqrt{\pi}')U' = \begin{pmatrix} 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} .$$

Eventually, $f(Z) = \|\tilde{Z}\|^2$, where $\tilde{Z} \sim \mathcal{N}\left(0, \begin{pmatrix} I_{m-1} & 0 \\ 0 & 1 \end{pmatrix}\right)$. Using the results on Gaussian vectors, one deduces that $\tilde{Z}_k$ are independent, that $\tilde{Z}_k \sim \mathcal{N}(0,1)$, $\forall k = 1, ..., m-1$ and $\tilde{Z}_m = 0$ a.s. The probability distribution of $f(Z)$ is $\chi^2(m-1)$.

Remark: The above proof is very similar to that of the Cochran theorem, which may also be used here directly.

3.1.1 Goodness-of-fit χ^2-test

Consider here $X_1, X_2, ..., X_n$ an n-sample of iid random variables with probability distribution $\mathbb{P}$, supposed to be unknown, and defined on $(E, \mathcal{E})$. We would like to test the hypothesis that $\mathbb{P}$ is equal to a particular distribution $\mathbb{P}_0$, i.e. to decide between

$$H_0 : \mathbb{P} = \mathbb{P}_0 \quad \text{against} \quad H_1 : \mathbb{P} \neq \mathbb{P}_0 .$$

Furthermore, consider also a partition of E, $\{E_1, ..., E_m\}$, such that $E = \cup_{k=1}^m E_k$ and $E_k \cap E_l = \emptyset$, $\forall k \neq l \in \{1, ..., m\}$. In most of the cases, the sample is resumed by the random vector $N(n) = (N_1(n), ..., N_m(n))$, where $N_k(n) = \sum_{i=1}^n \mathbb{1}_{X_i \in E_k}$, $\forall k = 1, ..., m$.

The test probability distribution $\mathbb{P}_0$ is resumed by $p_0^k = \mathbb{P}_0(E_k)$, $\forall k = 1, ..., m$.

Intuitively, if the X_i's are distributed according to $\mathbb{P}_0$, the (psedo) χ^2-distance $D(\mathbb{P}_n, \mathbb{P}_0)$ should be small. Furthermore, under H_0, $D(\mathbb{P}_n, \mathbb{P}_0) \xrightarrow[n\to\infty]{(l)} \chi^2(m-1)$, and if there exists k such that $p_k \neq p_0^k$, the law of large number states that

$$\frac{N_k(n)}{n} \xrightarrow[n\to\infty]{a.s.} p \neq p_0^k \ \Rightarrow \ D(\mathbb{P}_n, \mathbb{P}_0) \xrightarrow[n\to\infty]{a.s.} 0 .$$

With the above remarks, we may conclude that $D(\mathbb{P}_n, \mathbb{P}_0)$ may be a suitable test statistic:

$$T_n(X_1^n) = D(\mathbb{P}_n, \mathbb{P}_0) = \sum_{k=1}^m \frac{(N_k(n) - np_0^k)^2}{np_0^k} .$$

The rejection region may then be written as $W = \{T_n(X_1^n) > q_{1-\alpha}\}$, where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the $\chi^2(m-1)$ probability distribution. Also, the decision rule of the test may be written as $\phi(x) = \mathbb{1}_W(x) = \mathbb{1}_{\{T_n(x_1^n) > q_{1-\alpha}\}}$.

Remark: The χ^2 test is an asymptotic test. In practice, we consider that the results are reliable provided that $np_0^k \geq 5$, $\forall k = 1, ..., m$. If this is not the case, the partition of E may be modified by concatenating some of its elements.

Remark: If one wishes to test whether the sample is issued from a family of parametric probability distributions, indexed by $\theta \in \mathbb{R}^d$, then θ will be estimated from the sample $(X_1, ..., X_n)$ - using, for example, the maximum likelihood or any other consistent and asymptotically Gaussian estimate -, and the test is identical, except for the number of degrees of freedom of the χ^2-distribution, which is now equal to $m-d-1$.

Example:
We would like to test whether the number of births in a maternity ward is uniformly distributed across the year. A sample of 88 births is available and may be resumed as follows:

April - June	July - August	September - October	November - March
27	20	8	33

The hypothesis to be tested are:
H_0 : "the births are uniformly distributed across the year";
H_1 : "the births are not uniformly distributed across the year".

Under H_0, we are in an equiprobability framework. Under this hypothesis, the expected values np_0^k previously defined may be resumed in the following table :

April - June	July - August	September - October	November - March
$\frac{3}{12}88 = 22$	$\frac{2}{12}88 = 15$	$\frac{2}{12}88 = 15$	$\frac{5}{12}88 = 37$

The test statistic is then equal to :

$$D^2 = \frac{(27-22)^2}{22} + \frac{(20-15)^2}{15} + \frac{(8-15)^2}{15} + \frac{(33-37)^2}{37} = 6.51 \ .$$

Moreover, under H_0, $D^2 \sim \chi^2(4-1) = \chi^2(3)$.

If the level of the test is $\alpha = 0.05$, one gets from the tables that $q_{0.95} = 7.815$, the 0.95-quantile of the $\chi^2(3)$ distribution. The rejection region of the test is then $W = \{D^2 \geq 7.815\}$.

Since $D^2 = 6.51 < 7.815 = q_{0.95}$, the null H_0 hypothesis cannot be rejected given the observed data, and we conclude that the births in the maternity ward are uniformly distributed.

We may also compute the p-value associated to this test, p-value $= \mathbb{P}\left(\chi^2(3) \geq 6.51\right) = 0.09$. With a test level $\alpha = 0.05$, H_0 cannot be rejected, but with $\alpha = 0.10$, H_0 is rejected. The test is significant at 10%, but not at 5%.

Let us now consider the statistics for the entire population of the country, for a whole year. This may be summarized by the following table:

April - June	July - August	September - October	November - March
27385	19978	8106	33804

In this case, the expected values np_0^k are:

April - June	July - August	September - October	November - March
$\frac{3}{12}89273 = 22318$	$\frac{2}{12}89273 = 14879$	$\frac{2}{12}89273 = 14879$	$\frac{5}{12}89273 = 37197$

and the test statistic is

$$D^2 = \frac{(27385-22318)^2}{22318} + \frac{(19978-14879)^2}{14879} + \frac{(8106-14879)^2}{14879} + \frac{(33804-37197)^2}{37197} = 6290.39 \ .$$

For a fixed level, $\alpha = 0.05$, the critical value of the χ^2 does not change. Hence, since

$$D^2 = 6290.39 > 7.815 = q_{0.95} \ ,$$

we may conclude that, as $D^2 \in W$, H_0 is rejected and the probability distribution of the births is not uniform across the year.

We may also compute the p-value of the test, p-value $= \mathbb{P}\left(\chi^2(3) \geq 6290.39\right) = 0$, hence the level of the test α may be fixed as small as wished.

Remark: Among the two tests, the second is preferred since the size of the sample is much larger and the results are thus more reliable.

3.1.2 χ^2-test for independence

The χ^2-test may also be used for checking the independence of two random variables. Consider (X_1, Y_1),..., (X_n, Y_n) an n-sample of iid random vectors, where the X_i's are valued in $\{a_1, ..., a_{m_1}\}$ and the Y_i's are valued in $\{b_1, ..., b_{m_2}\}$.

The hypothesis to be tested are:

$$\{H_0 : X \text{ and } Y \text{ are independent}\} \text{ ; against } \{H_1 : X \text{ and } Y \text{ are not independent}\} .$$

In this case, the hypothesis H_0 is equivalent to

$$\mathbb{P}\{(X, Y) = (a_i, b_j)\} = p_{i\bullet}p_{\bullet j} = \mathbb{P}(X = a_i)\mathbb{P}(Y = b_j), \ \forall i = 1, ..., m_1; j = 1, ..., m_2.$$

The hypothesis H_0 maybe also written in terms of membership to a parametric family of probability distributions. Indeed, the unknown parameter would be

$$\theta = (p_{1\bullet}, ..., p_{m_1-1\bullet}, p_{\bullet 1}, ..., p_{\bullet m_2-1}) \in]0, 1[^{m_1+m_2-2}.$$

The maximum likelihood estimate for θ is then

$$\hat{\theta} = \left(\frac{N_{1\bullet}}{n}, ..., \frac{N_{m_1-1\bullet}}{n}, \frac{N_{\bullet 1}}{n}, ..., \frac{N_{\bullet m_2-1}}{n}\right) ,$$

where

- $N_{i\bullet} = \sum_{l=1}^{n} \mathbb{1}_{\{X_l=a_i\}}$, $\forall i = 1, ..., m_1$;
- $N_{\bullet j} = \sum_{l=1}^{n} \mathbb{1}_{\{Y_l=b_j\}}$, $\forall j = 1, ..., m_2$;
- and also, let us denote $N_{i,j} = \sum_{l=1}^{n} \mathbb{1}_{\{X_l=a_i, Y_l=b_j\}}$, $\forall i = 1, ..., m_1$ and $\forall j = 1, ..., m_2$.

Using the asymptotic Gaussian distribution of the maximum likelihood estimate, one may easily prove that:

$$T_n(X_1^n, Y_1^n) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{\left(\frac{N_{i\bullet}N_{\bullet j}}{n} - N_{i,j}\right)^2}{\frac{N_{i\bullet}N_{\bullet j}}{n}} \xrightarrow[n\to\infty]{(l)} \chi^2((m_1 - 1)(m_2 - 1)) .$$

Remark: The rejection region may then be written as $W = \{T_n(X_1^n, Y_1^n) > q_{1-\alpha}\}$, where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the $\chi^2((m_1 - 1)(m_2 - 1))$ probability distribution. Also, the decision rule of the test may be written as $\phi(x) = \mathbb{1}_W(x) = \mathbb{1}_{\{T_n(x_1^n, y_1^n) > q_{1-\alpha}\}}$.

Remark: Since the χ^2 test is an asymptotic test, in practice we consider that the results are reliable if $N_{i,j} \geq 5$, $\forall i = 1, ..., m_1$ and $\forall j = 1, ..., m_2$. If this is not the case, some values on the support of X and/or Y may be put together.

Remark: If the variables for which the test is performed are not categorical or discrete with a finite support, finite partitions of the supports of X and Y may be considered.

Example:
The following table resumes the results of a study conducted on 120 young adults in Switzerland and concerns the use of mobile phones. Two variables are available, X, the mean duration of the phone calls emitted or received, and Y, the education level.

Duration/Education level	High-school	University
0 to 2 minutes	7	10
2 to 4 minutes	21	15
4 to 6 minutes	42	25

Can one conclude, based on this data, at the existence of a relation between the education level and the time spent on the phone by young adults in Switzerland?

First, let us compute the marginal numbers, $N_{i\bullet}$ and $N_{\bullet j}$:

Duration/Education level	High-school	University	Overall
0 to 2 minutes	7	10	17
2 to 4 minutes	21	15	36
4 to 6 minutes	42	25	67
Overall	70	50	120

Next, let us compute the expected statistics in case on independence (under H_0), $N_{i\bullet}N_{\bullet j}$:

Duration/Education level	High-school	University	Overall
0 to 2 minutes	9.92	7.08	17
2 to 4 minutes	21	15	36
4 to 6 minutes	39.08	27.92	67
Overall	70	50	120

The test statistic is:

$$T = \frac{(7-9.92)^2}{9.92} + \frac{(10-7.08)^2}{7.08} + \frac{(21-21)^2}{21} + \frac{(15-15)^2}{15} + \frac{(42-39.08)^2}{39.08} + \frac{(25-27.92)^2}{27.92} = 2.59$$

Moreover, under H_0, $T \sim \chi^2((2-1)(3-1)) = \chi^2(2)$. If the level of the test is $\alpha = 0.05$, the critical value is the 0.95-quantile of a $\chi^2(2)$ distribution, $q_{0.95} = 5.991$.

Since

$$T = 2.59 < 5.991 = q_{0.95} \Rightarrow T \notin W \ ,$$

we decide, based on the available data, that we cannot state the existence of a relation between the education level and the duration of the phone calls.

One may also compute the p-value of this test, p-value $= \mathbb{P}\left(\chi^2(2) \geq 2.59\right) = 0.28$.

3.1.3 A few more exercices

Exercice 1
After long years of clinical trials, the survival rates (without medical treatment) for bronchial-cancer patients were established:

survival without medical treatment	≤ 6 months	7-12 months	13-24 months	≥ 24 months
rate	0.45	0.35	0.15	0.05

At the same time, on a sample of sixty patients having received medical treatment combining chemotherapy and radiotherapy, the following statistics were recorded:

survival with medical treatment	≤ 6 months	7-12 months	13-24 months	≥ 24 months
number of patients	6	24	12	18

For fixed level $\alpha = 0.05$, may one conclude that the medical treatment is efficient? Which is the p-value associated to this test?

Exercice 2
We consider the following data, issued from *Bortkiewicz, 1898*, and summarizing the number of soldiers in the Prussian army killed accidentally by horse kicks, during 200 years.

No of deaths per year	0	1	2	3	4
Frequency	109	65	22	3	1

1. Compute the empirical mean of the random variable "number of accidental deaths per year", $\overline{x}$.
2. For a fixed level, $\alpha = 0.05$, test whether the data is issued from a Poisson distribution with parameter $\lambda = \overline{x}$. Which is the p-value of this test?

Exercice 3
In a study issued in 2000, two Spanish researchers investigated hundreds of facial expressions of men and women on their wedding photo. In all, 389 couples were studied and, for each face, it was recorded whether it was smiling or not. The results may be summarized as follows:

Facial expression/Gender	Men	Women	Overall
Smile	285	366	651
No smile	104	23	127
Overall	389	389	778

May one conclude at the existence of a relation between the gender and the facial expression on the wedding photo?

3.2 Kolmogorov-Smirnov tests

3.2.1 Some generalities

We consider here $X_1, X_2, \ldots$ a sequence of iid random variables with c.d.f. $F(x) = \mathbb{P}(X_1 \leq x)$, supposed to be unknown. We would like to test the hypothesis that F is equal to a particular distribution F_0, i.e. to decide between

and we define

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$$

the **empirical cumulative distribution function** (which, like F, is increasing, left-continuous and has limit at the right in each point).

Glivenko-Cantelli lemma (add proof!)

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \to +\infty]{as} 0.$$

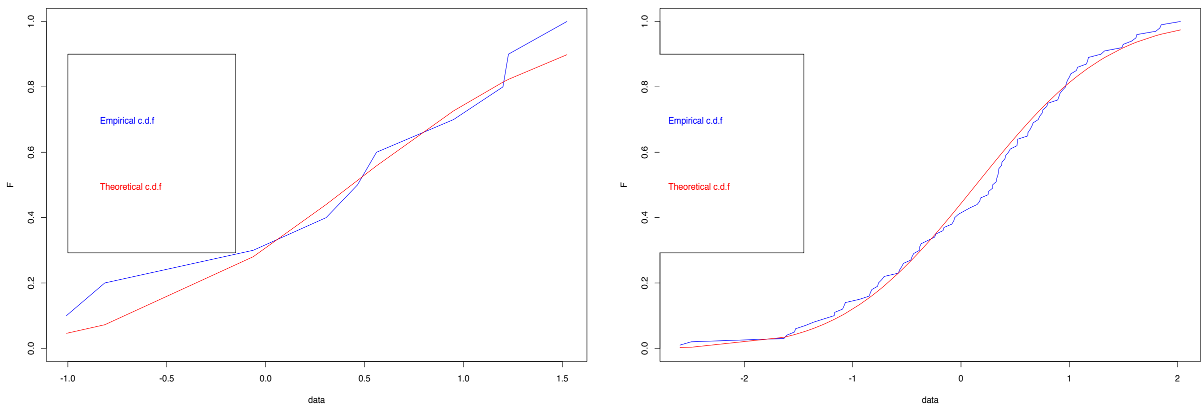

Figure 3.1: Empirical distribution (10 and 100 observations) and theoretical distribution, $\mathcal{N}(0,1)$

Remark: For simplicity reasons, we shall suppose hereafter that F is strictly increasing and continuous everywhere. Then, the inverse F^{-1} of F has its usual meaning, $F(X)$ is uniformly distributed on $[0,1]$ and $F^{-1}(U)$ has the same probability distribution as X_1, where $U \sim \mathcal{U}[0,1]$.
Without this hypothesis, it remains exact that $F^{-1}(U)$ has the same probability distribution as X_1. However, considering, for instance, $X_1 \sim \mathcal{B}\left(\frac{1}{2}\right)$ proves that $F(X_1)$, which takes three values only, cannot be uniform.

The Glivenko-Cantelli lemma justifies the idea of considering a test statistic of the form $\|F_n - F_0\|_\infty$ in order to test $H_0 : F = F_0$ against $H_1 : F \neq F_0$. If $\varepsilon \searrow 0$, the previous lemma proves that the sequence of tests having

$$W_n = \{\|F_n - F_0\|_\infty \geq \varepsilon\}$$

as rejection region is consistent. In order to consider the level of such a test, one must know the approximate quantiles of the probability distribution of $\|F_n - F_0\|_\infty$. The following result proves that this distribution depends on n only; hence, it may be tabulated after having simulated uniform distributions.

Theorem Kolmogorov - Smirnov (1): Suppose $F = F_0$. The test statistics

$$\begin{array}{rcll} D_n & = & \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| & ; \\ D_n^+ & = & \sqrt{n}\sup_{x\in\mathbb{R}} (F_n(x) - F_0(x)) & ; \\ D_n^- & = & \sqrt{n}\sup_{x\in\mathbb{R}} (F_0(x) - F_n(x)) & , \end{array}$$

have probability distributions independent of F_0. Moreover, $D_n^+ \overset{(l)}{=} D_n^-$.

Proof: Using the remark above, one may see that

$$F_n(x) - F(x) \overset{(l)}{=} U_n\left(F(x)\right) - F(x) \ ,$$

where $U_n(t)$ is the empirical c.d.f. of an n-sample issued from a uniform distribution. Then,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{(l)}{=} \sup_{t\in[0,1]} |U_n(t) - t| \ .$$

Since this statistic is independent of F, it is also called a **free statistic**.

Let us also remark that the random variables D_n^+ and D_n^- have the same probability distribution. If one disposes of an n-sample of $\mathcal{U}[0,1]$, let us eventually remark that the expressions D_n, D_n^+ and D_n^- are maximums of at most $2n$ values, since it is enough to consider the values for the observations in the sample, as well as the left limits at these points for computing the supremum in $\mathbb{R}$.
Following this, one may easily tabulate these probability distributions using the law of large numbers in the case where a sufficiently large number of independent uniform samples are available

Next, we admit the following (difficult) theorem :
Theorem Kolmogorov - Smirnov (2):

$$\lim_{n\to\infty} \mathbb{P}\left(D_n^+ > \lambda\right) = \exp\left(-2\lambda^2\right) \text{ and}$$

$$\lim_{n\to\infty} \mathbb{P}\left(D_n > \lambda\right) = 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp\left(-2\lambda^2 k^2\right) \ .$$

The asymptotic result is generally admitted whenever $n > 50$.

For a complete proof of the previous result, refer to P. Doukhan, *Empirical processes*. Let us however give some hints for understanding at least the factor $\sqrt{n}$. The following lemma, easy to prove, is left as exercice.

Lemma: Let $B_n(x) = \sqrt{n}\left(F_n(x) - F(x)\right)$. Then, for any $-\infty < x_1 \leq x_2 \leq ... \leq x_k < \infty$, one has

$$\left(B_n\left(x_1\right), ..., B_n\left(x_k\right)\right) \xrightarrow[n\to+\infty]{(l)} \left(B_1, ..., B_k\right) \ ,$$

$$\left(B_1, ..., B_k\right) \sim \mathcal{N}(0, \Sigma) \ ,$$

$$\Sigma = \left(\sigma_{ij}\right)_{i,j=1,\ldots,k} \ , \ \sigma_{ij} = F(x_i) \wedge F(x_j) - F(x_i)F(x_j) \ .$$

This lemma allows to figure that a functional central limit theorem is guiding the previous theorem. Then, if one admits that $\sqrt{n}(F_n - F) \xrightarrow[n\to+\infty]{(l)} B \circ F$, where B is a **Brownian bridge**: $(B(t))_{t\in\mathbb{R}}$ such that $\sum_{i=1}^{I} a_i B(t_i)$ is Gaussian, $\forall I$, $\forall a_i \in \mathbb{R}$, $\forall t_i \in [0,1]$, $i = 1, ..., I$ and such that $B(s) \sim \mathcal{N}(0, s(1-s))$ and $Cov(B(t), B(s)) = t \wedge s - ts$, $s, t \in [0,1]$.

The probability distributions in Theorem 2 are those of $\|B\|_\infty$.

3.2.2 Testing $F = F_0$

For testing the hypothesis $F = F_0$, $F \leq F_0$, $F \geq F_0$, one should use the $(1-\alpha)$-quantiles $d_{n,1-\alpha}$, $d^+_{n,1-\alpha}$ of the probability distributions of D_n and, respectively, $D_n^{\pm}$ and reject the null hypothesis whenever the test statistic exceeds the corresponding threshold.

- For testing $F = F_0$ against $F \neq F_0$, the rejection region is $W = \{D_n > d_{n,1-\alpha}\}$;
- For testing $F \leq F_0$ against $F > F_0$, the rejection region is $W = \left\{D_n^+ > d^+_{n,1-\alpha}\right\}$;
- For testing $F \geq F_0$ against $F < F_0$, the rejection region is $W = \left\{D_n^- < d^+_{n,\alpha}\right\}$.

All of the above tests are of level α and consistent.

In order to prove the last assertion, let us remark, for example,

$$F < F_0 \Rightarrow \limsup_n \sup_x \left(F_n(x) - F(x)\right) \leq 0 \ ,$$

hence $\mathbb{P}\left(\sup_x \left(F_n(x) - F(x)\right) < d\right) \xrightarrow[n\to+\infty]{} 1, \forall d > 0$.

The asymptotic behavior of the sequence $d_{n,1-\alpha}$ is obtained using Theorem (2). These expressions are computed using the rank representations which will be detailed in the next sections.

3.2.3 The case of two samples

Let us now consider two independent samples of iid random variables $X_1, ..., X_n \sim F$ and $Y_1, ..., Y_m \sim G$, where F and G are the corresponding cumulative distribution functions. Then, similarly to the previous reasoning, one may prove the following result.

Theorem: Let $c_{n,m} = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}}$. The test statistics

$$\begin{array}{rcll} D_{n,m} & = & c_{n,m} \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)| & ; \\ D^+_{n,m} & = & c_{n,m} \sup_{x\in\mathbb{R}} \left(F_n(x) - G_m(x)\right) & ; \\ D^-_{n,m} & = & c_{n,m} \sup_{x\in\mathbb{R}} \left(G_m(x) - F_n(x)\right) & , \end{array}$$

have probability distributions independent of F and G, if these c.d.f. are continuous and strictly increasing.

The goal is to test one of the following couples of hypothesis :

- For testing $F = G$ against $F \neq G$, the rejection region is $W = \{D_{n,m} > d_{n,m,1-\alpha}\}$;
- For testing $F \leq G$ against $F > G$, the rejection region is $W = \left\{D^+_{n,m} > d^+_{n,m,1-\alpha}\right\}$;
- For testing $F \geq G$ against $F < G$, the rejection region is $W = \left\{D^-_{n,m} < d^-_{n,m,\alpha}\right\}$.

Under these conditions, the sequences $U_i = F(X_i)$ and $U_j = G(Y_j)$ are iid and distributed according to a uniform distribution in $[0, 1]$. The quantiles of these distributions may be simulated and tabulated.

3.2.4 Rank-based expressions of the test statistics

Since we supposed the distributions to be continuous, the probability of ex-aequo in the following list is null. Although we shall return to this in a more detailed manner in the following section, in order to obtain a simplified writing of these tests, it is of interest to introduce the **rank** of X_i is a list of variables $(X_1, ..., X_n)$ (with no ex-aequo):

$$R_X(i) = \sum_{j\neq i} \mathbb{1}_{(X_j \leq X_i)} \ .$$

This is also the rank occupied by X_i when the list of variables is ordered in ascending fashion, $X_{(1)} < X_{(2)} < ... < X_{(n)}$, called **order statistic**.

Then, one may directly derive the test statistics for the two-samples problem as follows :

$$\begin{aligned} D_{n,m} &= c_{n,m} \max\left\{ |\tfrac{i}{n} - \tfrac{j}{m}| \mid U_{(i)} < V_{(j)} < U_{(i+1)} \right\} \quad ; \\ D^+_{n,m} &= c_{n,m} \max\left\{ \tfrac{i}{n} - \tfrac{j}{m} \mid U_{(i)} < V_{(j)} < U_{(i+1)} \right\} \quad ; \\ D^-_{n,m} &= c_{n,m} \max\left\{ \tfrac{j}{m} - \tfrac{i}{n} \mid U_{(i)} < V_{(j)} < U_{(i+1)} \right\} \quad ; \end{aligned}$$

and, for the one-sample problem,

$$\begin{aligned} D_n &= \sqrt{n} \max\left\{ |\tfrac{i}{n} - u| \mid U_{(i)} < u < U_{(i+1)} \right\} \quad ; \\ D^+_n &= \sqrt{n} \max\left\{ \tfrac{i}{n} - u \mid U_{(i)} < u < U_{(i+1)} \right\} \quad ; \\ D^-_n &= \sqrt{n} \max\left\{ u - \tfrac{i}{n} \mid U_{(i)} < u < U_{(i+1)} \right\} \quad . \end{aligned}$$

3.3 Some practical examples

Example 1: lifespan of an electrical device.
A sample of 5 devices were tested. The observed lifespans, measured in hours, were $\{133; 169; 8, 122; 58\}$. One would like to test if the lifespan of these devices is distributed according to an exponential law.

1. Give an estimation for the parameter of an exponential distribution which could be fit to the observed data;
2. State the hypothesis to test (null and alternative).
3. Compare the empirical distribution to the theoretical distribution using a Kolmogorov-Smirnov test. Which is the conclusion of the test?

Solution of example 1.

1. The parameter of an exponential distribution, $\mathcal{E}(\lambda)$, represents the inverse of the expected value, $\mathbb{E}(X) = \frac{1}{\lambda}$. Let us then compute the observed empirical mean :

$$\overline{x} = \frac{1}{5}(133 + 169 + 8 + 122 + 58) = 98 \ .$$

   Then, an estimation of λ is $\hat{\lambda} = \frac{1}{\overline{x}} = \frac{1}{98}$.

2. The hypothesis to test are H_0:“the lifespan of the considered devices is distributed according to $\mathcal{E}(\frac{1}{98})$” versus H_1:“the lifespan of the considered devices is not distributed according to $\mathcal{E}(\frac{1}{98})$”.
3. The information from the sample may be resumed as in the following table Then, the test statistic

Ordered observed sample $x_{(i)}$	8	58	122	133	169
Cumulative sizes n_i	1	2	3	4	5
Cumulative frequencies f_i	0.2	0.4	0.6	0.8	1
Theoretical c.d.f. $F_0(x_{(i)})$	0.078	0.0447	0.712	0.743	0.822
$\lvert F_0(x_{(i)}) - f_i \rvert$	0.122	0.047	0.112	0.057	0.178

   is $D_5 = \sqrt{5} \times 0.178 = 0.398$. For $n = 5$ and $\alpha = 0.05$, the Kolmogorov-Smirnov table provides the quantile $d_{5,\ 0.95} = 0.5633$. Since $D_5 < d_{5,\ 0.95}$, then, according to the available data, the hypothesis H_0 cannot be rejected and the exponential distribution $\mathcal{E}(\frac{1}{98})$ will be accepted as probability distribution for the sample.

Example 2: tree heights in a forest.
Two samples of trees were selected in two forests. We are interested to know if the distributions of the heights are the same within the two forests. The registered values for the two samples are :

$$\{x_1, ..., x_{12}\} = \{23.4, 24.4, 24.6, 24.9, 25, 26.2, 26.3, 26.8, 26.9, 27, 27.6, 27.7\}$$

$$\{y_1, ..., y_{14}\} = \{22.5, 22.9, 23.7, 24.0, 24.4, 24.5, 25.3, 26, 26.2, 26.4, 26.7, 26.9, 27.4, 28.5\}$$

State the hypothesis to test (null and alternative). Compare the two empirical distributions using a Kolmogorov-Smirnov test. Which is the conclusion of the test?

3.4 Exercices

Exercice 1: At the start of a horse race, there are usually eight lanes and the first lane is the closest to the fence. One supposes that a horse is more likely to win a race when it starts from a level with a low number, that is it is closed to the inner fence. The data from 144 races is available and may be summarized as follows: State the hypothesis to test (null and alternative). Compare the empirical distribution to the

Lane number	1	2	3	4	5	6	7	8
Winners	29	19	18	25	17	10	15	11

theoretical distribution using a Kolmogorov-Smirnov test. Which is the conclusion of the test?

Exercice 2: A meticulous antique dealer records in a notebook the time spent between the arrival of customers during a continuous day of work, from 10am to 6pm. The arrival times for one given day are the following :

$$\{10h15; 10h40; 11h15; 11h27; 12h57; 13h03; 14h18; 15h23; 16h20; 16h24; 17h09\}.$$

These arrival times allow to compute, for example, the waiting time, in minutes, between two costumers:

$$\{25; 35; 12; 90; 6; 75; 65; 57; 4; 45\}.$$

The antique dealer knows the costumers arrive independently and, but he would like to further know if the number of costumers arriving in a time interval T depends on T only, or also on the instant in time with respect to which the time interval T is observed.

1. Show that, if the number of costumers arriving independently in a time interval T depends on T only, the waiting time between two costumers is distributed according to an exponential law, and that the number of customers arriving during the interval T is distributed according to a Poisson law.
2. Give an estimation for the mean waiting time between two arrivals of customers.
3. State the hypothesis to test (null and alternative). Compare the empirical distribution to the theoretical distribution using a Kolmogorov-Smirnov test. Which is the conclusion of the test?

3.5 Rank tests

3.5.1 Some considerations on rank statistics

3.5.2 Wilcoxon test

Let us describe here a nonparametric test useful for deciding whether two samples are issued from the same probability distribution or not. Consider two independent samples $X_1, ..., X_n \sim F$ and $Y_1, ..., Y_m \sim G$, where F and G are the corresponding cumulative distribution functions. F and G are supposed to be continuous and strictly increasing.

The hypothesis to test are $H_0 : F = G$ against $H_1 : F \neq G$.

Let us denote $N = n + m$ and $(Z_1, ..., Z_N) = (X_1, ..., X_n, Y_1, ..., Y_m)$. Next, we consider the ranks and the order statistics related to the concatenated samples :

$$Z_{(1)} < Z_{(2)} < ... < Z_{(N-1)} < Z_{(N)}\ ,\ R_Z(i) = 1 + \sum_{j \neq i} \mathbb{1}_{Z_j < Z_i}\ ,\ 1 \leq i \leq N.$$

Then, R_Z is the permutation of $\{1, ..., N\}$ such that $Z_{R_Z(i)} = Z_{(i)}$, $\forall i = 1, ..., N$. This random variable has a uniform distribution on $\mathcal{S}_N$, the set of permutations of $\{1, ..., N\}$ (of cardinal $N!$).

Definition: The sum of ranks of the X_i , $W_n = \sum_{i=1}^{n} R_Z(i)$, is called the **Wilcoxon** statistic.

The probability distribution of W_n (which depends both on n and m) is tabulated. Let us remark that n and m can be switched, provided that W_n is replaced by a sum from $n + 1$ to N, hence the quantile tables contain the case $n \leq m$ only. Obviously, this probability distribution does not depend on F if $F = G$.

When testing $H_0 : F = G$ against $H_1 : F > G$, the rejection region has the form $W = \{W_n > w_\alpha\}$, where $\alpha > 0$ is the test level. Here w_α is the $(1-\alpha)$-quantile of the probability distribution of W_n, which can be tabulated since it is the same (under H_0) probability distribution for $W_U = \sum_{i=1}^n R_U(i)$, for a random iid sample $U = (U_1, ..., U_N)$ of uniform variables, $\mathcal{U}[0, 1]$.

Theorem: Under H_0,

$$\frac{W_n - \mathbb{E}W_n}{\sqrt{\mathbb{V}W_n}} \xrightarrow[n\to\infty]{(l)} \mathcal{N}(0,1) ,$$

where $\mathbb{E}W_n = n\mathbb{E}R_Z(1) = n\sum_{j=1}^N \frac{j}{N} = \frac{n(N+1)}{2}$ (since $\mathbb{P}(R_Z(i) = j) = \frac{1}{N}$) and $\mathbb{V}W_n = \frac{n(N+1)(N-n)}{12}$.

Example:
The concentration of a particular chemical in a river is subject to daily measurements. The values registered at two spots $P1$ and $P2$ are the following :

Point 1 (X)	5.32	5. 00	5.14	5.00	5.35	5.17	5.11	526
Point 2 (Y)	5.33	5.13	5.16	5.09	5.49	5.32	5.24	5.23

1. Is there a statistically significant difference in the concentrations on the pollutant between the two spots?
2. The second spot is situated downstream the first spot. One wants to tests whether the discharges of a plant situated between the two spots lead to an increase in the concentration of the pollutant at spot 2. The table above contains the registered values of these concentration during eight days. How is the test to be modified in this case? What is the conclusion of the test?

3.5.3 Wilcoxon signed-rank test

Consider $(X_1, ..., X_n)$ an iid n-random sample, issued from the probability distribution $\mathbb{P}$, continuous with respect to the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. P is supposed to be unknown and the test checks whether P is symmetric around zero. Let us denote $\left(X^a_{(1)}, ..., X^a_{(n)}\right)$ the ordered statistic associated to $|X_i|$.

Consider $(x_1, ..., x_n)$ an observed value of the sample.

The probability distribution $\mathbb{P}$ being continuous, one has that $\mathbb{P}(|X_i| = |X_j|) = 0$ for any $i \neq j$, hence we might consider that there are no ties in the sample.

Eventually, let $R_{|X|}$ be the vector of ranks:

$$R_{|X|}(i) = k \;\Leftrightarrow |X_i| = X^a_{(k)} \;, \forall i, k = 1, ..., n \;.$$

Symmetry test

The hypothesis to test here are

$$H_0 : \mathbb{P} \text{ is symmetric around zero} \;;\; H_1 : \mathbb{P} \text{ is not symmetric around zero.}$$

The test statistic is

$$W_n^+(X) = \sum_{i=1}^n R_{|X|}(i)\mathbb{1}_{X_i>0} = \sum_{k=1}^n kB_k^+ \;,$$

where $B_k^+ = \mathbb{1}_{\left\{X_i,\ (|X_i|=X^a_{(k)})>0\right\}}$.

The decision function associated to this test is

$$\phi(x) = \mathbb{1}_{\left\{W_n^+(x)\leq s_1\right\}} + \mathbb{1}_{\left\{W_n^+(x)\geq s_2\right\}} \;.$$

Finally, how do we choose the threshold values s_1 and s_2? Under the null hypothesis, $B_k^+ \sim \mathcal{B}(0.5)$. Moreover, the B_k^+ variables are independent. Hence, the probability distribution of the test statistic W_n^+ is free of P under H_0. Furthermore, it is symmetric with respect to its expected value, $\frac{n(n+1)}{4}$. One may thus either read s_1 and s_2 in the corresponding table for small values of n, or use an asymptotic approximation by $\mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$ for large values of n.

Test on the center of symmetry

Here, we suppose that the probability distribution P **is symmetric** and the hypothesis testing is about the value of the center of symmetry m, only:

$$H_0 : m = m_0 \ ; \text{ against } H_1 : m \neq m_0, \text{ or } m > m_0, \text{ or } m < m_0 \ .$$

If one denotes $Y_i = X_i - m_0$, $y_i = x_i - m_0$, $\forall i = 1, ..., n$, the test statistic is $W_n^+(Y)$. According to the alternative hypothesis, the decision function is

$$\phi(x) = \mathbb{1}_{\left\{W_n^+(y) \leq s_1\right\}} + \mathbb{1}_{\left\{W_n^+(y) \geq s_2\right\}}, \text{ or } \phi(x) = \mathbb{1}_{\left\{W_n^+(y) \leq s\right\}}, \text{ or } \phi(x) = \mathbb{1}_{\left\{W_n^+(y) \geq s\right\}} \ .$$

Example:

3.5.4 Spearman test

Now, consider (X_1, Y_1),..., (X_n, Y_n) an iid sequence of random vectors, with unknown probability distribution. We would like to test if the random variables X and Y are independent. For this, we shall use the Spearman statistic

$$S_n = \sum_{i=1}^{n} R_X(i) R_Y(i)$$

Under the null hypothesis H_0 : X and $Y$are independent, ont gets that

$$\mathbb{E}(S_n) = \frac{1}{4} n(n+1)^2 \ ; \ \mathbb{V}(S_n) = \frac{1}{144}(n-1)n^2(n+1)^2 \ .$$

Let us remark that the two extreme situations, $R_X = R_Y$ and $R_X = n + 1 - R_Y$, are naturally leading to the following bounds

$$\sum_{i=1}^{n} i(n+1-i) = \frac{n(n+1)(n+2)}{6} \leq S_n \leq \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \ .$$

When $n \to \infty$, the probability distribution of S_n is asymptotically Gaussian; hence a critical region of the Spearman test has the form $W = \{S_n < \underline{s}\} \cup \{S_n > \overline{s}\}$, for tabulated $\underline{s}$ and $\overline{s}$ which allow to reach any level α.

Eventually, let us recall the Spearman correlation coefficient or the empirical correlation between the random vectors R_X and R_Y:

$$\rho_{S_n} = \frac{Cov(R_X, R_Y)}{S_{R_X} S_{R_Y}} \ ,$$

where

$$Cov(R_X, R_Y) = \frac{1}{n} \sum_{i=1}^{n} R_X(i) R_Y(i) - \frac{1}{n^2} \sum_{i=1}^{n} R_X(i) \sum_{i=1}^{n} R_Y(i) \ ,$$

and

$$\begin{aligned} S_{R_X}^2 &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} R_X^2(i) - \left(\tfrac{1}{n} \sum_{i=1}^{n} R_X(i)\right)^2 = \tfrac{1}{n} \sum_{i=1}^{n} i^2 - \left(\tfrac{1}{n} \sum_{i=1}^{n} i\right)^2 = S_{R_Y}^2 \\ &= \tfrac{(n+1)(2n+1)}{6} - \left(\tfrac{n+1}{2}\right)^2 = \tfrac{n^2-1}{12} \ . \end{aligned}$$

Then, one may easily remark that

$$\rho_{S_n} = \frac{12 S_n - 3n(n+1)^2}{n(n^2-1)} \ ,$$

is an affine function of S_n.

Example:

3.5.5 Shapiro-Wilks test for Gaussian distribution

Consider $(X_1, ..., X_n)$ an iid n-random sample, issued from the probability distribution $\mathbb{P}$ and let $(x_1, ..., x_n)$ be an observed value of the sample. We wish to test:
$H_0 :$ X is distributed according to a Gaussian law ; $H_1 :$ X is not distributed according to a Gaussian law.

The test statistic here is

$$SW_n(X) = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} ,$$

where

- a_i are constants depending on the expected value m and on the variance matrix V of the order statistic associated to an n-sample of a standard Gaussian distribution, $(a_1, ..., a_n) = \frac{m'V^{-1}}{(m'V^{-1}V^{-1}m)^{1/2}}$ and available in tables/software;
- $\left(X_{(1)}, ..., X_{(n)}\right)$ is the order statistic associated to $(X_1, ..., X_n)$.

This test statistic may be seen as a coefficient of determination between the quantile vector issued from a Gaussian distribution and the empirical quantiles computed from the data. The scatter-plot of these two vectors is the **quantile-quantile** graph or the "**Q-Q plot**".

The associated decision rule is $\phi(x) = \mathbb{1}_{\{SW_n(x) \leq s\}}$ and the threshold value s may be read in the Shapiro-Wilks tables.

Chapter 4

Confidence intervals and confidence regions

In practice, giving an estimation for the parameter of a model is not enough, in most situations. One would also like to have a more precise idea on the "safety margin" one has in the knowledge of this parameter and its estimation.

Definition: Let us consider the parametric framework $(E^n, \mathcal{E}_n, \mathbb{P}_\theta, \theta \in \Theta)$, where $\Theta \subset \mathbb{R}^p$. Let $\alpha \in [0,1]$ an a priori fixed number. The **confidence region** on the parameter θ of level $1-\alpha$ is a random subset $R_{1-\alpha} \in \mathbb{R}^p$ and defined on $(E^n, \mathcal{E}_n)$, such that $\forall \theta \in \Theta$, $\{(x_1, ..., x_n) \in E^n, \theta \in R_{1-\alpha}(x_1, ..., x_n)\} \in \mathcal{E}_n$, and

$$\inf_{\theta \in \Theta} \{\mathbb{P}_\theta(\theta \in R_{1-\alpha})\} \geq 1 - \alpha \ .$$

If an observed sample $(X_1(\omega), ..., X_n(\omega))$ is known, then $R_{1-\alpha}(X_1(\omega), ..., X_n(\omega))$ is called observed confidence region. Furthermore, if the parameter is univariate ($p = 1$), one will speak of **confidence interval**.

Now the next question is how does one actually build a confidence region? First, it is obvious that for any $\alpha \in [0,1]$, $R_{1-\alpha} \subset \Theta$ (usually, α is chosen to be close to 0, and $\alpha = 0.05$ is mostly used). One possible approach pour building a confidence region is the following: naturally, one would fancy using a consistent estimate $\hat{T}$ of θ, except that the probability distribution of $\hat{T}$ is usually depending on θ, and this makes difficult (with few exceptions) its direct utilization. One should then prefer using a **pivot function** $\pi(\hat{T}, \theta)$, which is a measurable function depending on the estimate $\hat{T}$ and the parameter θ and which is a free statistic. Then, the inequality in the previous definition could be written under the alternative form:

$$\inf_{\theta \in \Theta} \left\{\mathbb{P}_\theta(\pi(\hat{T}, \theta) \in C_\alpha)\right\} \geq 1 - \alpha \ ,$$

where C_α is a deterministic region. This way, the confidence region can be subsequently computed using the quantiles (usually $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of the probability distribution of the pivot function.

Remark: For a given level, $1-\alpha$, the manner of defining a confidence interval is not unique. Indeed, if one writes $\alpha = \alpha_1 + \alpha_2$, $\alpha_1, \alpha_2 \geq 0$, it is enough to consider A_α and B_α such that $\mathbb{P}_\theta(\theta \leq A_\alpha) = \alpha_1$ and $\mathbb{P}_\theta(\theta \geq B_\alpha) = \alpha_2$, provided that $A_\alpha < B_\alpha$. But, in general, one chooses $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ (some hints justifying this symmetric choice will be given hereafter).
A confidence interval of the form $[A_\alpha \ ; \ B_\alpha]$ is called a bilateral confidence interval. However, in some cases, it might be more relevant to look for unilateral confidence intervals, generally when one bound is already known (0, for example).

4.1 The duality between hypothesis testing and building a confidence interval

There exists a strong connection between the confidence region and the rejection region of a parametric statistical test.

Duality theorem: Let $\mathcal{R}(\theta_0)$ be the rejection region of a non-randomized test of level α of the hypothesis Θ_0 against $\Theta_1 = \overline{\Theta}_0$, such that

$$\sup_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}(X \notin \mathcal{R}(\theta_0)) = \alpha \ .$$

Suppose the mapping $\theta \mapsto \mathcal{R}(\theta)$ defined for any $\theta \in \Theta$, then a confidence region of level $\geq 1-\alpha$ for the parameter θ is given by $R(X) = \{\theta \in \Theta \mid X \notin \mathcal{R}(\theta)\}$.
Conversely, this mapping associates the rejection region $\mathcal{R}(\theta_0)$ of a statistical test with level $\leq \alpha$ to each confidence region $R(X)$.

When $\Theta_0 = \{\theta_0\}$, the hypothesis $H_0 = \theta = \theta_0$ will be accepted, provided that θ_0 belong to the confidence interval of level $1-\alpha$.

4.2 The Gaussian case

4.2.1 Confidence intervals for the mean of a univariate Gaussian distribution

Case 1: the variance is known

Let $(X_1, ..., X_n)$ be an n-sample of $\mathcal{N}\left(m, \sigma_0^2\right)$. Suppose the variance σ_0^2 is known and the mean m is the parameter to be estimated. Then, according to Fisher theorem,

$$\frac{\overline{X}_n - m}{\frac{\sigma_0}{\sqrt{n}}} \sim \mathcal{N}(0,1) \ ,$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the empirical mean. If the level of the test is $1-\alpha$, and $q_{1-\alpha/2}$ is chosen such that $\mathbb{P}(\mathcal{N}(0,1) \geq q_{1-\alpha/2}) = \frac{\alpha}{2}$, then one has that

$$\mathbb{P}_m\left(-q_{1-\alpha/2} \leq \frac{\overline{X}_n - m}{\frac{\sigma_0}{\sqrt{n}}} \leq q_{1-\alpha/2}\right) = 1-\alpha \ ,$$

and, this leads to the confidence interval

$$IC = \left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \ ; \ \overline{X}_n + q_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] .$$

Case 2: the variance is unknown

Next, let $(X_1, ..., X_n)$ be an n-sample of $\mathcal{N}\left(m, \sigma^2\right)$. Suppose that both parameters, m and σ^2 are unknown. Then, still according to Fisher theorem,

$$\frac{\overline{X}_n - m}{\frac{S}{\sqrt{n}}} \sim St(n-1) \ ,$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ is the empirical unbiased variance. If the level of the test is $1-\alpha$, and $q_{1-\alpha/2}$ is chosen such that $\mathbb{P}(St(n-1) \geq q_{1-\alpha/2}) = \frac{\alpha}{2}$, then one has that

$$\mathbb{P}_m\left(-q_{1-\alpha/2} \leq \frac{\overline{X}_n - m}{\frac{S}{\sqrt{n}}} \leq q_{1-\alpha/2}\right) = 1-\alpha \ ,$$

and, this leads to the confidence interval

$$IC = \left[\overline{X}_n - q_{1-\alpha/2}\frac{S}{\sqrt{n}} \ ; \ \overline{X}_n + q_{1-\alpha/2}\frac{S}{\sqrt{n}}\right] .$$

Example

One of the characteristics of a machine may be considered as a random variable X distributed according to $\mathcal{N}\left(m, \sigma^2\right)$, where m and σ^2 are unknown. A sample of 16 machines is drawn at random, let $x_1, ..., x_{16}$ be the observed values of this characteristic and let :

$$\overline{x} = \frac{1}{16}\sum_{i=1}^{16} x_i \ , \ s^2 = \frac{1}{15}\sum_{i=1}^{16}(x_i - \overline{x})^2$$

1. Compute a symmetric confidence interval of level 98% for m.
2. The values of the observed sample led to the confidence interval: $[374\ ,\ 426]$. Compute the estimations $\overline{x}$ and s^2 in the sample.
3. Propose a strategy for decreasing the amplitude of the interval.

4.2.2 Confidence intervals for the variance of a univariate Gaussian distribution

Case 1: the mean is known

Let $(X_1, ..., X_n)$ be an n-sample of $\mathcal{N}\left(m_0, \sigma^2\right)$. Suppose the mean m_0 is known and the variance σ^2 is the parameter to be estimated. Then, according to Cochran theorem,

$$\frac{\sum_{i=1}^n \left(X_i - m_0\right)^2}{\sigma^2} \sim \chi^2(n).$$

Next, using the $q_{\alpha/2}$ and $q_{1-\alpha/2}$ quantiles of a $\chi^2(n)$ distribution, one may write that

$$\mathbb{P}_{\sigma^2}\left(q_{\alpha/2} \leq \frac{\sum_{i=1}^n \left(X_i - m_0\right)^2}{\sigma^2} \leq q_{1-\alpha/2}\right) = 1 - \alpha$$

and then derive the confidence interval

$$IC = \left[\frac{\sum_{i=1}^n \left(X_i - m_0\right)^2}{q_{1-\alpha/2}}\ ;\ \frac{\sum_{i=1}^n \left(X_i - m_0\right)^2}{q_{\alpha/2}}\right].$$

Case 2: the mean is unknown

Let us also consider the more general case $(X_1, ..., X_n)$ be an n-sample of $\mathcal{N}\left(m, \sigma^2\right)$. Suppose that both parameters, m and σ^2 are unknown. Then, still according to Cochran theorem,

$$\frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2} \sim \chi^2(n-1).$$

Next, using the $q_{\alpha/2}$ and $q_{1-\alpha/2}$ quantiles of a $\chi^2(n-1)$ distribution, one may write that

$$\mathbb{P}_{\sigma^2}\left(q_{\alpha/2} \leq \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2} \leq q_{1-\alpha/2}\right) = 1 - \alpha$$

and then derive the confidence interval

$$IC = \left[\frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{q_{1-\alpha/2}}\ ;\ \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{q_{\alpha/2}}\right].$$

Example

The machines of an industrial plant are tuned such that the mean diameter of the outcome production is equal to 5cm. Suppose that the diameter X is a random variable distributed according to a Gaussian with mean $m = 5$ and unknown variance σ^2.
A random sample of 30 pieces is drawn at random and their diameter is measured, $x_1, ..., x_{30}$.

1. Give an estimate for the variance.
2. The values in the observed sample allowed to compute

$$\sum_{i=1}^{30} \left(x_i - 5\right)^2 = 1,2$$

   Which is the corresponding estimation of σ^2?

3. Compute a unilateral confidence interval of level 95% for σ^2.

4.2.3 Confidence interval for a proportion

Examples

1. The statistical control of the quality found that, in a given industrial plant, among 100 pieces of the outcome, 10 are defective, since they do not satisfy the production standards. What can one infer on the entire production of defective pieces?

2. In a sample of 600 patients with bronchial cancer, 550 were found to be smokers. Build a confidence interval for the proportion of smokers among bronchial-cancer patients. Do these results imply that smoking is harmful?

4.3 Using asymptotically efficient estimates for building confidence regions

Remark: If the statistical model $(E^n, \mathcal{E}_n, \mathbb{P}_\theta, \theta \in \Theta)$ is regular, under the usual conditions for the asymptotic normality of the maximum likelihood estimate, one may show (using Slutsky lemma) that

$$\pi(\hat{\theta}_n, \theta_0) = \sqrt{n} \cdot (I_1(\hat{\theta}_n))^{1/2} \cdot \left(\hat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{(l)} \mathcal{N}_p(0, I_p) \ ,$$

where I_p is the identity matrix of size p and $(I_1(\theta))^{1/2} \cdot (I_1(\theta))^{1/2} = I_1(\theta)$, for all $\theta \in \Theta$. Hence, for large n, the probability distribution of $\pi(\hat{\theta}_n, \theta_0)$ may be approximated by the standard multidimensional Gaussian distribution. But, if $Z \sim \mathcal{N}_p(0, I_p)$, with $q_{1-\alpha/2}$ the quantile of a standard univariate Gaussian distribution of level $1 - \alpha/2$, such that $\mathbb{P}(Z \in [-q_{1-\alpha/2} \ ; \ q_{1-\alpha/2}]^d) \geq 1 - \alpha$. It follows that the polyhedron $\frac{1}{\sqrt{n}} \cdot (I_1(\hat{\theta}_n))^{-1/2} \cdot [-q_{1-\alpha/2} \ ; \ q_{1-\alpha/2}]^d$ recentered about $\hat{\theta}_n$ will represent the desired confidence region.

4.3.1 Exercices

Exercice 1: Consider $(E^n, \mathcal{E}_n, (\mathbb{P}_\theta)^{\otimes n}, \theta > 0)$ a parametric model, such that $\mathbb{P}_\theta$ has a density with respect to the Lebesgue measure

$$f_\theta(x) = \frac{1}{\ln 2} \cdot \frac{1}{x} \cdot \mathbb{1}_{]\theta;2\theta[}(x) \ .$$

1. Show that the maximum likelihood estimate for θ is not unique.

2. Let us denote $\hat{\theta}_n^1 = \frac{1}{2}\max(X_1, ..., X_n)$ and $\hat{\theta}_n^2 = \min(X_1, ..., X_n)$. For each of these estimates, compute their probability distributions and prove that they are consistent. Without entering into computation details, explain why their are biased. Compute a confidence interval for θ.

Exercice 2: Let $((X_1, Y_1), ..., (X_n, Y_n))$ be an iid random sample, with $X_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$. Suppose that the probability distribution of Y_i conditionally to $X_i = 0$ is $\mathcal{B}(p_0)$, with $p_0 \in]0, 1[$, and the probability distribution of Y_i conditionally to $X_i = 1$ is $\mathcal{B}(p_1)$, with $p_1 \in]0, 1[$. Moreover, $X_i \sim \mathcal{B}(1/2)$.

1. Show that this sample belongs to the exponential family.

2. Compute $\hat{\theta}_n$, the maximum likelihood estimate of $\theta = (p_0, p_1)$. Show that $\hat{\theta}_n$ is consistent and asymptotically efficient.

3. Give a confidence region of level 95% for the parameter θ.

Bibliography

[1] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics : basic ideas and selected topics*. Englewood Cliffs, N.J. Prentice Hall, 1977.

[2] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Springer US, Boston, MA, 1974.

[3] Paul Doukhan. *Estimation et introduction aux tests, Cours de statistiques M1, Université de Cergy Pontoise*, 2014.

[4] Magalie Fromont. *Tests stiatistiques : rejeter, ne pas rejeter... se risquer?*, 2015.

[5] A. W. van der Vaart. *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge (UK), New York (N.Y.), 1998. Autre tirage : 2000 (édition brochée), 2005, 2006, 2007.