Summer semester 2008
Department of Mathematics, UniFR
Introduction to statistics: Test Theory

Christian Mazza*

April 2008

Contents

*Department of Mathematics, University of Fribourg, Pérolles, Chemin du Musée 23, CH-1700 Fribourg

1 Introduction

We shall be mainly concerned with decision problems where we must choose one of two densities f_0 and f_1, when the observed data take values in $\mathbb{R}$ or $\mathbb{R}^n$, or between two probabilites P_0 and P_1, when the data take their values in a countable space. Later, we shall again consider this kind of problems, but without specifying precisely the form of the densities, and develop non-parametric tests based on ranks and permutations. Let us begin with some illustrative examples.

Example 1.1. Assume that a scientist wants to test if some kind of animal is able to learn from experiments. The scientist works with an ensemble of n animals, which must go through a corridor which divides in two parts, the left and the right. If the animal chooses the left, it receives some quantity of food. But if the animal chooses the right, the animal will instead receive a small electrical impulse. Each animal will cross the corridor several times; after these runs, the experimentalist will assign a number to each animal, which is 1 when the animal chooses more often the left, and 0 otherwise. They are basically two hypotheses, which are

$$H_0: \quad \text{the animals are unable to learn}$$

and

$$H_1: \quad \text{the data shows evidence of learning capacities.}$$

H_0 is called the **null hypothesis** and H_1 **the alternative hypothesis**. It is also common to write H for H_0 and K for H_1. Here, scientific experiments suggest that H_1 might be true: we think that animals can learn.

Mathematical formulation: The observed data is here modeled by a set of n Bernoulli random variables $\varepsilon_1, \cdots, \varepsilon_n$, which can be assumed to be i.i.d. with some probability of success $p \in [0, 1]$. The above hypotheses can be translated in a mathematical form as

$$H_0: \ p = p_0 = \frac{1}{2} \text{ and } H_1: \ p = p_1 > \frac{1}{2}.$$

Intuitive decision rule: choosing a critical domain. Clearly, common sense suggests that a natural test might consist in opting for H_1 when the number of animals choosing the left is large enough, or equivalently if the statistics $S = \varepsilon_1 + \cdots + \varepsilon_n \in \{0, 1, \cdots, n\}$ is larger than some threshold value j.
Let P_0 (resp. P_1) be the law of the sample $\{\varepsilon_1, \cdots, \varepsilon_n\}$ when $p = p_0$ (resp. $p = p_1$). The domain of rejection of H_0, also called the **critical domain**, is the subset

$$R = \{j, j+1, \cdots, n\},$$

that is, our decision rule rejects H_0 when the observed statistics S belongs to R. Otherwise H_0 is not rejected.

Errors of type I and II

Error of the first kind: the level of the test. As usual in real life experiments, any decision can lead to erroneous statements or actions. Having in mind that we think that H_1 is correct, we naturally look for the **probability of stating that H_1 is true when H_0 is true**. This will corresponds to the so-called **error of first kind** or **level** of the test. Mathematically, one looks for the probabiliy that $\{S \geq j\}$, that is $\{S \in R\}$, when the data is distributed according to P_0. Here, the statistics S is Binomial of parameter of success p_0, and therefore the level of this test is given by

$$\sum_{k=j}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}, \tag{1.1}$$

which can computed when n is not too large. For example, the threshold $j = 8$ gives a level given by

$$\frac{1}{2^{10}}\left(\binom{10}{8} + \binom{10}{9} + \binom{10}{8}\right) \approx 0.055,$$

corresponding to a level approximatively equal to 5.5 %.

Error of the second kind and power: Clearly, we can also commit a second error, **the error of second kind**, which is the **probability of not rejecting H_0 when H_1 is true**. Usually, one is more interested in the level since it is more prejudiciable to

state that H_1 is true when it is not, than the contrary. Here, the prejudice is related to a wrong scientific statement, but in most cases, such wrong affirmations can lead to strong consequences, like if H_0 means that the water of some lake is contaminated by some toxic chemical product and H_1 is the statement that there is no cause of concern, so that you can safely take a bath.
In our example, this second probability is given by

$$\tilde{\alpha} := \sum_{k=0}^{j-1} \binom{n}{k} p_1^k (1-p_1)^{n-k},$$

which is obtained by setting $p = p_1$ since we assume that H_1 is true. In fact, one looks at the **power** β of the test which is the probability of accepting H_1 when H_1 is true, with

$$\beta = 1 - \tilde{\alpha} = \sum_{k=j}^{n} \binom{n}{k} p_1^k (1-p_1)^{n-k}. \tag{1.2}$$

Minimizing both type of errors: Ideally, the level should be as small as possible while the power β should be as large as possible, that is close to 1, that is (1.1) should be small while (1.2) should be close to 1.

The free parameter in this setting is the threshold $j \in \{1, 2, \cdots, n\}$, and we will adjust it to ensure that the probability of making an error of type I is smaller that some given value α.

Consider the dependence of (1.1) and (1.2) as function of j: Clearly, these functions are decreasing functions of j, so that this ideal aim can not be attained: one can not simultaneously minimize (1.1) and maximize (1.2).

Hence, comparing the risks, a natural method is to control the error of first kind by imposing that it is smaller than a fixed value α, which seems reasonable, and is accepted by most scientists. The standard choice is $\alpha = 5\%$, but $\alpha = 1\%$ is quite often used in the literature.

The next relevant observation is that, up to now, we did not use any data: We just defined a test by choosing conveniently the

threshold j, to get a level close to 5.5%. After the experiment, the scientist obtains some number S. If S belongs to the critical domain, that is if $S \in R = \{8, 9, 10\}$, then there is strong evidence for learning capacities, at a significant level $\alpha = 5.5\%$. If however $S \notin R$, one can not reject significatively H_0.

The notion of p-value: There is also a related way of analysing the observed data S: Given S, adopting the above generic test, we see that any choice j smaller than the observed value S will provide a test rejecting H_0, while any j larger that S will lead to the acceptation of H_0. Hence S is a frontier between two different regimes, the rejection and the acceptation of H_0.
On the other hand, the probabilty given in (1.1) is decreasing as function of j, and H_0 is rejected as soon as j crosses the observed S. This leads naturally to the notion of **p-value** of the test, which is the smallest level α leading to the rejection of H_0.

In our example, the p-value is therefore given by

$$p_{val} = p_{val}(S) = \sum_{k=S}^{n} \binom{n}{k} p_0^k (1 - p_0)^{n-k}. \tag{1.3}$$

In some sense, the notion of p-value is useful since, given the observed data S, we do not need to choose j ensuring a level smaller than 5% or 1%, allowing flexibility. However, we must also remember that both the level and the power are decreasing functions of j.
Numerically, the exact computations of the probabilities given by (1.1), (1.2) and (1.3) are only feasible when n is not too large. For large values of n, one can use the central limit Theorem to get estimates. We shall discuss these approximations in the forthcoming chapters. When $n = 10$, we can compute or give precise estimates for these probabilites.

We continue with a new example similar to the previous one, but here n is large.

Example 1.2. The swiss official statistics indicate that in 1972, they were $k = 47179$ boys among the $n = 91342$ babies born during this year. The proportion of boys is then $k/n \approx 0.5165$.

The natural hypothesis to test in this context is wether it is true that the probability p to get a boy is larger than $p_0 = 1/2$. This problem is similar to that considered in Example 1.1, but here n is large, and the formulas given by (1.1), (1.2) and (1.3) can not be used efficiently. We keep the basic ideas, but transform the events in order to use the central limit Theorem. H_0 and H_1 are the same,

$$H_0 : \ p = p_0 = \frac{1}{2} \text{ and } H_1 : \ p = p_1 > p_0.$$

The relevant statistics is still S_n where

$$S_n = \sum_{i=1}^{n} \varepsilon_i,$$

where now $\varepsilon_i = 1$ when the ith baby is a boy, and $\varepsilon_i = 0$ otherwise. H_0 is rejected when S_n belongs to the domain R given by $R = \{j, \cdots, n\}$. The level is therefore given by (1.1), with $n = 91342$. The central limit Theorem yields that, when H_0 is true,

$$\frac{\frac{S_n}{n} - p_0}{\sqrt{\mathrm{Var}(\frac{S_n}{n})}} = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}},$$

is approximatively standard normal N(0,1).
The point here is that the statistics

$$U(X) = \frac{X - np_0}{\sqrt{np_0(1-p_0)}},$$

is strictly increasing as a function of X. Therefore

$$X \in R \text{ if and only if } U(X) \geq U(j).$$

This shows that we can equivalently define a domain of rejection of H_0 of the form

$$R = \{U(X) > z\},$$

for some z. We thus reject H_0 when $U(S_n)$ is larger than some threshold z. The level is then given by

$$P(Z > z), \text{ where } Z \text{ is standard normal.} \tag{1.4}$$

Given a prescribed level α, we choose the threshold $z = z_\alpha$ so that

$$P(Z > z_\alpha) = \alpha.$$

These values are obtained from statistical tables. Common values are given by $z_\alpha = 1.645$ when $\alpha = 5\%$, $z_\alpha = 1.96$ when $\alpha = 2.5\%$ and $z_\alpha = 2.326$ when $\alpha = 1\%$.
Numerically, one has $x = 47179$, so that

$$U(x) = \frac{47179 - \frac{1}{2}91324}{\sqrt{91324\frac{1}{2}\frac{1}{2}}} \approx 10.$$

Looking at the various values of z_α for $\alpha = 5\%, 2.5\%,\ 1\%$, we see that 10 is always in R, meaning that H_0 is strongly rejected, and there is a strong evidence that $p > 1/2$.

Remark 1.1. Example 1.1 is concerned with random variables of binomial type. Most statistical problems however deal with data of continuous type, taking values in $\mathbb{R}^d$. A simple way of describing such data is to use probability densities f, and the decision problem consists in choosing in a set of two densities f_0 and f_1, when precise models are known. We shall see in the next chapter a generic way of constructing efficient tests. When no relevant parametric information about the distribution of the data is available, we shall discuss non-parametric procedures.

2 Parametric tests

Consider a probability space $(\mathcal{X}, \mathcal{B}, (\mathrm{P}_\theta)_{\theta\in\Theta})$, where $\mathcal{X}$ denotes a space like $\mathbb{R}^n$, $\mathcal{B}$ denotes the Borel σ-algebra on $\mathcal{X}$, and where the family of probability measures P_θ, indexed by parameters $\theta \in \Theta$, models some stochastic phenomenon of interest. Notice that $\mathcal{X}$ denotes the set of values taken by random vector X. For readers having difficulties with measure theory, imagine that the family of probability measures P_θ is defined through probability densities so that $\mathrm{P}_\theta(dx) = f_\theta(x)dx$, for some f_θ.

Coming back to Example 1.1, the basic probability space is $\mathcal{X} = \{0, 1, 2, \cdots, n\}$. The binomial random variable $X = S$ is such

that

$$P_\theta(S=k)=\binom{n}{k}p^k(1-p)^{n-k} \text{ with } \theta=p,\ \Theta=[0,1].$$

The test we have considered in this setting is based on critical domain R of the form $R=\{j,\cdots,n\}$. Having fixed the threshold j, the rule is to reject the null hypothesis $H_0:\ p=p_0$ and to accept the alternative H_1 when the observed statistics S belongs to R. This defines an **algorithm** $\phi(S)$ having S as an input, and $\phi=0$ and $\phi=1$ as possible outputs. More precisely,

- $\phi(S)=1$ means that we reject H_0, and
- $\phi(S)=0$ means that we reject H_1,

the function ϕ being defined by the formula

$$\phi(S)=\mathbb{I}_R(S),$$

where $\mathbb{I}$ denotes the indicator function. This is one of the many possible tests, as shown by the general Definition 2.2.

Definition 2.1 (**Hypothesis and alternative).** In the general setting, the set of parameters Θ is partitioned in two subsets: The *hypothesis* $H\subset\Theta$ and the *alternative* $K=\Theta\setminus H$. When both H and K reduce to singletons like $H=\{\theta_0\}$ and $K=\{\theta_1\}$, we recover the setting of Example 1.1 with $H=H_0$ and $K=H_1$.

Basic decision problem: Having observed $X=x$, one must decide between H and K, that is, we must check if the observed data $X=x$ is distributed according to some f_θ for $\theta\in H$ of for $\theta\in K$.

Definition 2.2. A statistical test ϕ is a measurable map $(\mathcal{X},\mathcal{B})\longrightarrow([0,1],\mathcal{B})$. Having observed $X=x$, the test rejects H with probability $\phi(x)$ and does not reject H with probability $1-\phi(x)$.

DECISION RULE

Having observed some data $X=x$, we apply this a test ϕ to get a number $\phi(x)\in[0,1]$. The next step consists in throwing

a coin having the property that the probability to get a head is $\phi(x)$. If head appears, then we choose K, otherwise one chooses H. In the extreme case where $\phi(x) = 1$, no additional stochastic experiment is necessary and one chooses K. Of course we choose H when $\phi(x) = 0$.

TWO KIND OF ERRORS

As it is usual in real life experiments, nothing is perfect, and any decision can lead to erroneous affirmations. The next step consists in quantifying the probability of errors. As stated in Example 1.1, one can consider basically two kind of errors (if we assume that at least one of the two hypotheses is true): The first one is the level of the test, and is defined below.

Definition 2.3 (**Level).** Consider a decision problem which consists in choosing between the hypotheses H and K, on the basis of some observed data $X = x$. Let ϕ be a test related to this decision problem. The **level** of the test (error of the first kind) is related to the average probability of choosing K when H is true. Given a *level of significance* $0 < \alpha < 1$, one looks for a test ϕ such that

$$\sup_{\theta \in H} \mathbb{E}_\theta(\phi(X)) \leq \alpha. \tag{2.1}$$

$\sup_{\theta \in H} \mathbb{E}_\theta(\phi(X))$ is the *size* or the *level* of the test. If the test ϕ satisfies the constraint (2.1), one says that it is of *level* smaller than α.

Definition 2.4 (**Power).** The error of second kind $\tilde{\alpha} = 1 - \beta$, where β is the *power* of the test, is similarly related to the average probability of choosing H when K is true. For a given $\theta \in K$, the power $\beta(\theta)$ of the test ϕ is then defined as

$$1 - \beta(\theta) = \mathbb{E}_\theta(1 - \phi(X)) = 1 - \mathbb{E}_\theta(\phi(X)), \quad \theta \in K.$$

Optimization problem: Clearly, given some level of significance α, one looks for tests satisfying (2.1) while maximizing the power $\beta(\theta)$, $\theta \in K$.
A natural question is to ask for the existence of a test ϕ so that both α and $1 - \beta$ are simultaneously as small as possible. One

can show that no such test exists (see e.g. Example 1.1). This is the reason why we control the level by imposing (2.1), and then looks for a test maximizing the power $\beta(\theta)$, for every fixed $\theta \in K$, among all tests having a level of significance smaller than α.

2.1 Likelihood ratio tests

This chapter develops a rigourous mathematical approach for constructing tests in a classical parametric setting when the hypothesis is $H = \{\theta_0\}$ and the alternative is $K = \{\theta_1\}$, that is we suppose that the data X is distributed according to P_{θ_0} or P_{θ_1}. The mathematical theory related to test theory uses mainly probabilistic results. We here focus on simple and relevant notions, which are very useful in every practical situations.

NOTATIONS: we suppose in the sequel that all the random vectors X possess either a density f in the continuous case or a discrete probability measure P. When the data is continuous, replace P by f. We shall write P_0 and P_1 for P_{θ_0} and P_{θ_1}. Similarly, $\mathbb{E}_0$ and $\mathbb{E}_1$ will denote the expectations under the probability measures P_0 and P_1. When dealing with P_θ, we write $\mathbb{E}_\theta$ to denote expectation with respect to P_θ.

The following Lemma is fundamental in test theory, and is vey useful in many applied settings. It is also called the **likelihood ratio test**. The idea is that, having observed $X = x$, P_1 is more likely than P_0 if the **likelihood ratio**

$$\frac{\mathrm{P}_1(x)}{\mathrm{P}_0(x)} \quad \text{is large.}$$

Theorem 2.1 (Neyman-Pearson Lemma)**.** *Let $\alpha \in (0,1)$ be given. Consider the test problem*

$$H: \ \theta = \theta_0 \ \textit{against the alternative} \ K: \ \theta = \theta_1.$$

1. *We can find constants k, γ such that the following test ϕ^* has the required level $\alpha = \mathbb{E}_0(\phi^*(X))$. This test has the form*

$$\phi^*(x) = \begin{cases} 1 & \text{when } \mathrm{P}_1(x) > k\mathrm{P}_0(x), \\ \gamma & \text{when } \mathrm{P}_1(x) = k\mathrm{P}_0(x), \\ 0 & \text{when } \mathrm{P}_1(x) < k\mathrm{P}_0(x). \end{cases}$$

2. *ϕ^* is the most powerful among all tests of level smaller that α.*

Example 2.1.

Consider again Example 1.1. The statistical model consists in a sequence of $n = 10$ i.i.d. Bernoulli random variables of probability of success $\theta = p \in [0, 1]$. The null hypothesis is $H_0 : \ p = p_0 = 1/2$ and the alternative is $H_1 : \ p = p_1 > p_0$. Let $X = \sum_{i=1}^{n} \varepsilon_i$ be the related Binomial random variable of parameter (n,p). Given the observation $X = x$, The Neyman-Pearson Lemma deals with the likelihood ratio

$$\begin{aligned} \frac{P_{\theta_1}(x)}{P_{\theta_0}(x)} &= \frac{\theta_1^x(1-\theta_1)^{n-x}}{\theta_0^x(1-\theta_0)^{n-x}} \\ &= \Big(\frac{\theta_1}{\theta_0}\Big)^x \Big(\frac{1-\theta_0}{1-\theta_1}\Big)^x \Big(\frac{1-\theta_1}{1-\theta_0}\Big)^n, \end{aligned}$$

where the last factor is independent of the data x.

The NP test rejects with probability one H_0 when the likelihood ratio $P_{\theta_1}(x)/P_{\theta_0}(x)$ is larger than some threshold value.

By assumption,

$$\theta_0 < \theta_1,$$

so that both

$$\theta_1/\theta_0 \text{ and } (1-\theta_0)/(1-\theta_1) \text{ are larger than } 1$$

It then follows that the likelihood ratio

$$\begin{aligned}\frac{P_{\theta_1}(x)}{P_{\theta_0}(x)} &= \frac{\theta_1^x(1-\theta_1)^{n-x}}{\theta_0^x(1-\theta_0)^{n-x}} \\ &= \Big(\frac{\theta_1}{\theta_0}\Big)^x\Big(\frac{1-\theta_0}{1-\theta_1}\Big)^x\Big(\frac{1-\theta_1}{1-\theta_0}\Big)^n,\end{aligned}$$

is increasing as a function of x, so that

$$\frac{P_{\theta_1}(x)}{P_{\theta_0}(x)} > k \text{ if and only if } x > j.$$

The domain of rejection of the NP Test is therefore of the form

$$R = \{x > j\},$$

for some constant j, in accordance with what we have done intuitively in Example 1.1. We reject with probability one H_0 when $x > j$, and reject H_1 with probabiliy one when $x < j$. At the border $x = j$, we choose H_1 with probability γ and choose H_0 with probability $1 - \gamma$. This case was however not included in the test provided in Example 1.1. The constants j and γ must be chosen correctly in order to catch the right level α.

PROOF OF THE NP LEMMA:

ASSERTION I

The first assertion of the NP Lemma states that we can find constants k, γ such that the following test ϕ^* has the required level $\alpha = \mathbb{E}_0(\phi^*(X))$. This test has the form

$$\phi^*(x) = \begin{cases} 1 & \text{when } \mathrm{P}_1(x) > k\mathrm{P}_0(x), \\ \gamma & \text{when } \mathrm{P}_1(x) = k\mathrm{P}_0(x), \\ 0 & \text{when } \mathrm{P}_1(x) < k\mathrm{P}_0(x). \end{cases}$$

Let

$$B_c = \{x | \mathrm{P}_1(x) > c\mathrm{P}_0(x)\},$$

and consider the function

$$\alpha(c) = \mathrm{P}_0(B_c).$$

Cleary α is decreasing in c.

$$\text{If } c < 0 \text{ then } \alpha(c) = 1 \text{ and } \lim_{c\to\infty} \alpha(c) = 0$$

Moreover α is right continuous since,

under P_0, $1 - \alpha(c)$ corresponds to the distribution function of the random variable $\mathrm{P}_1(X)/\mathrm{P}_0(X)$, which is well defined P_0 almost surely.

Let $\alpha(c - 0)$ be the left limit evaluated at c.

We next compute the probability

$$\mathrm{P}_0(\mathrm{P}_1(X) = c\mathrm{P}_0(X)).$$

Consider the probability

$$\mathrm{P}_0(c - 1/n < \mathrm{P}_1(X)/\mathrm{P}_0(X) \leq c)$$

$$= \alpha(c - 1/n) - \alpha(c) \longrightarrow \alpha(c - 0) - \alpha(c),$$

showing that

$$\mathrm{P}_0(\mathrm{P}_1(X) = c\mathrm{P}_0(X)) = \alpha(c - 0) - \alpha(c).$$

Given a fixed level α, let c_0 be such that $\alpha(c_0) \leq \alpha \leq \alpha(c_0 - 0)$. We next consider the test given in 1., with $k = c_0$.

Taking expectation under $\mathbb{E}_0$, we can compute the level of the test to get

$$\begin{aligned}\mathbb{E}_0(\phi^*(X)) &= \mathrm{P}_0(\{\mathrm{P}_1(X) > k\mathrm{P}_0(X)\}) + \gamma \mathrm{P}_0(\{\mathrm{P}_1(X) = k\mathrm{P}_0(X)\}) \\ &= \alpha(k) + \gamma(\alpha(k-0) - \alpha(k)),\end{aligned}$$

and we see that we obtain the required level α when

$$\gamma = \frac{\alpha - \alpha(k)}{\alpha(k-0) - \alpha(k)}. \tag{2.2}$$

ASSERION II: ϕ^* IS THE MOST POWERFUL

Let ϕ be a test of level smaller or equal to α, and let ϕ^* be the NP test as given in 1. We consider the function

$$(\phi^*(x) - \phi(x))(\mathrm{P}_1(x) - k\mathrm{P}_0(x)),$$

which vanishes on the set $\{x;\ \mathrm{P}_1(x) = k\mathrm{P}_0(x)\}$. We consider separatly the sets

$$S_+ = \{x;\ \phi^*(x) - \phi(x) > 0\},$$

and

$$S_- = \{x;\ \phi^*(x) - \phi(x) < 0\}.$$

When $x \in S_+$, $\phi^*(x) > \phi(x)$, so that $\phi^*(x) > 0$, and $\mathrm{P}_1(x) \geq k\mathrm{P}_0(x)$. In the same way, $\mathrm{P}_1(x) \leq k\mathrm{P}_0(x)$ for all $x \in S_-$.
It follows that

$$\begin{aligned}&\int (\phi^*(x) - \phi(x))(\mathrm{P}_1(x) - k\mathrm{P}_0(x))\mathrm{d}x \\ &= \int_{S_- \cup S_+} (\phi^*(x) - \phi(x))(\mathrm{P}_1(x) - k\mathrm{P}_0(x))\mathrm{d}x \geq 0.\end{aligned}$$

Hence

$$\mathbb{E}_1(\phi^*(X)) - \mathbb{E}_1(\phi(X)) \geq k(\mathbb{E}_0(\phi^*(X)) - \mathbb{E}_0(\phi(X))) \geq 0,$$

where the last inequality follows from the fact that $\mathbb{E}_0(\phi^*(X)) = \alpha$ and $\mathbb{E}_0(\phi(X)) \leq \alpha$.

3 Distributions with monotone likelihood ratio

Example 2.1 provides a setting where the likelihood ratio $\mathrm{P}_1(x)/\mathrm{P}_0(x)$ is increasing as a function of the observed x. It turns out that the monotony of the likelihood ratio holds for important families of probability distributions, which occur often in applications.

Definition 3.1. Assume that $\Theta \subset \mathbb{R}$. The real-parameter family of distributions P_θ is said to have *monotone likelihood ratio* if there exists a real-valued statistics $T(X)$ such that, for any $\theta < \theta'$, the distribution P_θ and $\mathrm{P}_{\theta'}$ are distinct, and the ratio $\mathrm{P}_{\theta'}(x)/\mathrm{P}_\theta(x)$ is a nondecreasing function of $T(x)$.

Theorem 3.1 (Monotone likelihood ratio test)**.** *Let $\Theta \subset \mathbb{R}$, and let the random variable X have a probability distribution $\mathrm{P}_\theta(x)$ with monotone likelihood ratio in $T(x)$.*

1. *For testing $H : \ \theta = \theta_0$ against $K : \ \theta > \theta_0$, the NP test is given by*

$$\phi^*(x) = \begin{cases} 1 & \text{when } T(x) > C, \\ \gamma & \text{when } T(x) = C, \\ 0 & \text{when } T(x) < C, \end{cases} \tag{3.1}$$

where C and γ are determined by the boundary consition

$$\mathbb{E}_{\theta_0}(\phi^*(X)) = \alpha. \tag{3.2}$$

Example 3.1. THE GAUSSIAN CASE

Consider the special case of normal random vectors composed of i.i.d. components X_i, $i = 1, \cdots, n$. The statistical model might be of the form

$$X_i = \theta + \varepsilon_i, \quad i = 1, \cdots, n,$$

where the random variables ε_i are i.i.d. Normal $N(0, \sigma^2)$. We assume that σ is known. Here $\theta = \mathbb{E}(X)$, and X is normal $\mathrm{N}(\theta, \sigma^2)$. The joint density is given by

$$f_\theta(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{\theta}{\sigma^2}\sum_{i=1}^n x_i - \frac{n\theta^2}{2\sigma^2}).$$

Suppose that the problem consists in testing

$$H: \ \theta = \theta_0 \text{ against the alternative } \theta = \theta_1 > \theta_0.$$

The likelihood ratio is given by

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \exp(\frac{(\theta_1 - \theta_0)}{\sigma^2}\sum_i x_i + \frac{n(\theta_0^2 - \theta_1^2)}{2\sigma^2}).$$

Here, $\theta_1 > \theta_0$, so that the likelihood ratio is increasing as a function of

$$T(x) = \sum_{i=1}^n x_i.$$

The critical or rejection domain of the test is then of the form

$$R = \{x: \ T(x) > c\},$$

where the constant c must be adjusted to get the required level of significance. The statistics

$$U_n(x) = \frac{\bar{x}_n - \theta_0}{\sigma/\sqrt{n}}, \tag{3.3}$$

is standard normal N(0,1), so that the test rejects the hypothesis when $U_n(x)$ is larger than some threshold value.

Example 3.2 (Test based on the central limit Theorem)**.** We consider here a setting similar to that given in the preceeding example, but where no normality assumption is made. We assume a statistical model where the observed data is such that

- $X_i = \theta + \varepsilon_i$, are i.i.d., with
- $\mathbb{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$, where
- σ is unknown, and
- the sample size is large, that is $n >> 1$ or $n \geq 30$!

We consider the statistics (3.3) where σ^2 is replaced by the **empirical variance** s_n^2, where

$$s_n^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x}_n)^2,$$

to get the statistics

$$Z_n(x) = \frac{\bar{x}_n - \theta_0}{s_n/\sqrt{n}},$$

which is approximatively gaussian $\mathrm{N}(0,1)$ when $n >> 1$. The critical domain associated to the test problem

$$H: \ \theta = \theta_0 \text{ against the alternative } \theta = \theta_1 > \theta_0,$$

is given by

$$R = \{Z_n(x) > z_\alpha\},$$

where z_α is obtained from the table associated with the standard normal distribution, and is such that

$$\frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{\infty} \exp(-u^2/2) \mathrm{d}u = \alpha.$$

Example 3.3 (Exponential densities)**.** In this example, we assume that the i.i.d. sample is distributed according to probability measures having densities of the exponential family, of the generic form

$$f_\theta(x) = \exp(c(\theta)T(x) + d(\theta) + S(x)), \ x \in \mathbb{R}^n,$$

where we assume that $\Theta \subset \mathbb{R}$, $c(\theta)$ is such that $dc/d\theta \neq 0$, so that the model is well defined (bijection). Notice that this family contains gaussian densities. The likelihood ratio satisfies

$$\frac{f_{\theta'}(x)}{f_\theta(x)} = \exp(T(x)(c(\theta') - c(\theta))) \exp(d(\theta') - d(\theta)).$$

Consider the test problem $H: \ \theta = \theta_0$ against the alternative $K: \ \theta > \theta_0$. By assumption, c can be increasing or decreasing ($c'(\theta) \neq 0$).

- If c is increasing, the NP test rejects H when $T(x) > C$.
- If c is decreasing, the test rejects H when $T(x) < C$.

Notice that since one has densities, the probability of the event $\{T(X) = C\}$ vanishes.

Example 3.4 (Poisson random variables)**.** Recall that a Poisson random variable X takes its values in $\mathbb{N}$, and is defined through a positive parameter $\lambda > 0$. The model is thus given by the family of probabilities P_λ, where

$$\mathrm{P}_\lambda(j) = \frac{\lambda^j}{j!} e^{-\lambda}, \quad j \in \mathbb{N}.$$

Given an i.i.d. sample x_i, $i = 1, \cdots, n$, the joint distribution of the random vector $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ is given by

$$\mathrm{P}_\lambda(x_1, \cdots, x_n) = \frac{\exp(T(x)\ln(\lambda))}{x_1! \cdots x_n!} \exp(-n\lambda),$$

where the statistics $T(x)$ is just $\sum_{i=1}^n x_i$. Notice that this is an example of an element of the exponential family, in the discrete case.
Consider the test problem $H: \ \lambda = \lambda_0$ again the alternative $K: \lambda > \lambda_0$. Considering the likelihood ratio $\mathrm{P}_{\lambda'}(x)/\mathrm{P}_\lambda(x)$ with $\lambda' > \lambda$, one checks that it is increasing as a function of the statistics $T(x)$, so that the NP test rejects H with probability one when $T(x) > C$. Under P_{λ_0}, $T(X)$ is Poisson of parameter $n\lambda_0$, and we can fix C using standard statistical tables.

Example 3.5 (Hypergeometric random variables)**.** From a lot containing N items of a manufactured product, a sample of size n is selected at random. Let D be the number of defective items. This will be the unknown parameter $\theta = D \in \{1, 2, \cdots, N\}$. The related family of distributions is then hypergeometric, with

$$\mathrm{P}_\theta(d) = \frac{\binom{\theta}{d}\binom{N-\theta}{n-d}}{\binom{N}{n}}.$$

Consider the testing problem $H : \ D = D_0$ against $K : \ D > D_0$. We again check if the related family of probabilites has the monotone likelihood ratio property. To this purpose, one first consider a ratio of the form $\mathrm{P}_{D_1}(d)/\mathrm{P}_{D_0}(d)$, where $D_1 = D_0 + 1$. The likelihood ratios are such that

$$\frac{\mathrm{P}_{D+1}(d)}{\mathrm{P}_D(d)} = \frac{(D+1)(N-D-n+d)}{(N-D)(D+1-d)},$$

which is monotone increasing as a function of d. For arbitraries $D_1 > D_0$, one then use the factorization

$$\frac{\mathrm{P}_{D_1}(d)}{\mathrm{P}_{D_0}(d)} = \frac{\mathrm{P}_{D_1}(d)}{\mathrm{P}_{D_1-1}(d)} \frac{\mathrm{P}_{D_1-1}(d)}{\mathrm{P}_{D_1-2}(d)} \cdots \frac{\mathrm{P}_{D_0+1}(d)}{\mathrm{P}_{D_0}(d)},$$

to obtain that

$$\frac{\mathrm{P}_{D_1}(d)}{\mathrm{P}_{D_0}(d)},$$

is monotone increasing as a function of d. The likelihood ratio test rejects thus H when the observed number of defective items d is larger than some threshold value.

4 Student t-Test

This is perhaps one of the most used test is statistics. Consider again the statistical model with additive errors

$$X_i = \mu + \varepsilon_i, \quad i = 1, \cdots, n,$$

where we assume no systematic errors, that is $\mathbb{E}(\varepsilon_i) = 0$, and that the errors have a certain variance σ^2 which is usually unknown. We have already considered this model in

- Example 3.1, where the X_i were i.i.d. normal, with σ known, and in
- Example 3.2, where the X_i were i.i.d., σ unknown, but where the sample size n was large, so that we replaced σ by the empirical standard deviation, and then used the Central Limit Theorem to fix the test.

Here, we suppose that

- the sample is i.i.d. gaussian $\mathrm{N}(\mu, \sigma^2)$,
- the sample size n is not too large,
- σ is unknown.

We will consider two basic types of test problems:

Unilateral problems:

a) $H: \ \mu = \mu_0$ against the alternative $K: \ \mu = \mu_1 > \mu_0$,

or

b) $H: \ \mu = \mu_0$ against the alternative $K: \ \mu = \mu_1 < \mu_0$,

Bilateral problem:

c) $H: \ \mu = \mu_0$ against the alternative $K: \ \mu = \mu_1 \neq \mu_0$.

Suppose we have observed an i.i.d. $\mathcal{N}(\mu, \sigma^2)$ sample $x_1, \ldots, x_n$, and consider the problem given in a):

$$H: \ \mu = \mu_0 \quad K: \ \mu = \mu_1 > \mu_0.$$

Taking inspiration by looking at Example 3.1, using the approximation of σ provided in Example 3.2, and the fact that the gaussian family has the monotone likelihood ratio property, see Example 3.1 and Example 3.3, common sense suggests the statistics

$$t(x) = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}},$$

where s_n is the sample standard deviation, and **T-test** for a) given by

$$\phi(x) = \begin{cases} 1 & \text{if } t(x) = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha, n-1} \\ 0 & \text{otherwise.} \end{cases}$$

This T-test is fixed by noting that the statistics $t(X)$ follows a **Student distribution of $\nu = n-1$ degrees of freedom**, of density

$$g_\nu(x) = \frac{\left(\frac{\nu}{\nu+x^2}\right)^{\frac{1+\nu}{2}}}{\sqrt{\nu}\,\beta(\frac{\nu}{2},\frac{1}{2})},$$

where $\beta(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$. The cases b) and c) are treated in a similar way, see below:

Summary:

Alternative K	Critical domain
$K : \mu > \mu_0$	$R :\ t(x) > t_{\alpha,n-1}$
$K : \mu < \mu_0$	$R :\ t(x) < -t_{\alpha,n-1}$
$K : \mu \neq \mu_0$	$R :\ \lvert t(x)\rvert > t_{\alpha/2,n-1}$

where $t_{\alpha,\nu}$ is given by $\int_{t_{\alpha,\nu}}^{\infty} g_\nu(w)\mathrm{d}w = \alpha$.

Example 4.1 (Safety level)**.** A city health department aims to determine if the mean bacteria count per unit volume of water at a lake beach is within the safety level of 200. A researcher collected 10 water samples of unit volume and found the bacteria counts to be

$$175, 190, 215, 198, 184, 207, 210, 193, 196, 180.$$

Do the data strongly indicate that there is no cause for concern?

Let μ denote the current mean bacteria count per unit volume of water. Then, the statement "'no cause of concern "' translates to $\mu < 200$, and the researcher is seeking strong evidence in support of this hypothesis. So, the formulation of the null and alternative hypotheses should be

$$H :\ \mu = 200 \text{ against } K :\ \mu < 200.$$

Here, one puts the risky statement under K to control the error of type I. Since the counts are spread over a wide range, an

approximation by a continuous distribution is not unrealistic for inference about the mean. Assuming further that the measurments constitute a sample from a normal population, we use the t-test with the statistics

$$t(x) = \frac{\bar{x}_n - 200}{s_n/\sqrt{10}}, \quad d.f. = n - 1 = 9.$$

The test rejects H when $t(x)$ is smaller than some threshold value $-t_{\alpha,9}$. Here,

$$\bar{x}_n = 194.8, \quad s_n = 13.14,$$

and

$$t(x) = \frac{194.8 - 200}{13.14/\sqrt{10}} = \frac{-5.2}{4.156} = -1.25.$$

Because the observed $t = -1.25$ is larger than $-t_{0.01,9} = -2.821$, the null hypothesis is not rejected at the level $\alpha = 0.01$. On the basis of the data obtained from these 10 measurments, there does not seem to be strong evidence that the true mean is within the safety level.

R code

```
> (mean(x)-200)/(sd(x)/sqrt(10))
[1] -1.251570
>  qt(c(0.01), df=9, lower.tail=TRUE)
[1] -2.821438
> t.test(x,mu=200,alternative=c("less"))

        One Sample t-test
data:  x
t = -1.2516, df = 9, p-value = 0.1211
alternative hypothesis: true mean is less than 200
95 percent confidence interval:
     -Inf 202.4162
sample estimates:
mean of x
    194.8
```

The upper part shows that T-statistics equals -1.25170 and that $\mathrm{P}(T < -2.8214) = 0.01 = \alpha$ and so the $\mu_0 = 200$ cannot be

rejected at level $\alpha = 0.01$. All this can be done automatically by invoking

```
t.test(x,mu=200,alternative=c("less"))
```

5 Tests and confidence intervals

Assume the statistical model

$$X_i = \mu + \varepsilon_i,\ i_1, \cdots, n,$$

where the ε_i are i.i.d. centered gaussian $\mathrm{N}(0, \sigma^2)$. We assume here that σ is unknown.

Definition 5.1. A confidence intervall for the mean μ of confidence level $1 - \alpha$ is any random interval $I(X_1, \cdots, X_n)$ which contains the mean μ with probability $1 - \alpha$.

Confidence intervals are not restricted to the above simple additive model. We illustrate here this notion with this simple example. The maximum likelihood estimator for the mean is the empirical average $\bar{x}_n$, and we have in the preceeding chapter that the statistics

$$t(x) = \frac{\bar{x}_n - \mu}{s_n/\sqrt{n}},$$

follows a Student distribution of $(n-1)$ d.f. When n is large, the denominator of the above expression converges to the standard deviation σ, so that $t(X)$ is close to a standard normal r.v. Z, $Z \sim \mathcal{N}(0, 1)$ when $n >> 1$. The threshold is usually fixed to $n = 30$. This means that

$$P(-t_{\alpha/2,n-1} \leq t(X) \leq t_{\alpha/2,n-1}) = 1 - \alpha,$$

where $t_{\alpha/2}$ is the quantile related to the Student distribution with $n - 1$ degrees of freedom. This provides the natural confidence interval

$$I(x_1, \cdots, x_n) = [\bar{x}_n - \frac{t_{\alpha/2,n-1} s_n}{\sqrt{n}}, \bar{x}_n + \frac{t_{\alpha/2,n-1} s_n}{\sqrt{n}}],$$

which possess the required properties. In fact, the notion of confidence interval is very useful: once we have computed the interval I, we can get ideas of the results of many testing problems.

By Definition, $P(\mu \in I(X)) = 1 - \alpha$: consider the test problem

$$H : \ \mu = \mu_0 \text{ against } K : \ \mu \neq \mu_0,$$

the natural test is the bilateral t-test

$$\phi(x) = \begin{cases} 1 & \text{if } |t(x)| = |\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}| > t_{\alpha/2,n-1} \\ 0 & \text{otherwise.} \end{cases}$$

One thus see that the hypothesis $H : \ \mu = \mu_0$ is not rejected if $\mu_0 \in I(x_1, \cdots, x_n)$, and, on the contrary is rejected when $\mu_0 \notin I(x_1, \cdots, x_n)$. We thus see directly which hypotheses will be accepted at a α-level by looking at the confidence interval.

Example 5.1 (Confidence intervals for variances)**.** Suppose that $X_1, \ldots, X_n$are i.i.d with $\mathcal{N}(\mu, \sigma^2)$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. We have that

$$\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2_{n-1},$$

i.e. follows a Chi-Square distribution of $n-1$ degrees of freedom. This leads to

$$\mathrm{P}(\chi^2_{1-\frac{\alpha}{2};n-1} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2};n-1}) = 1 - \alpha,$$

where $\mathcal{X}_{1-\alpha/2,n-1}$ denotes related quantiles, or equivalently

$$\mathrm{P}(\frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}) = 1 - \alpha, \tag{5.1}$$

where the probabilityl distribution function of a χ^2_n is given by

$$f(x) = \frac{2^{-n/2} e^{-x/2} x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}.$$

Notice that (5.1) provides a $1 - \alpha$ confidence interval for the variance, which permits to test bilateral hypotheses.

6 Comparing two treatements

A recurrent problem in applied settings consists in comparing two treatments, or two populations. For example, the statistical

problem might consist in comparing a standard treatement A against a new one B; clearly, one expects that treatment B is better than treatment A. If this is a new medicament, one chooses a group a patient of size n_1 which receives A and a group of size n_2 which gets B. After the experiment, one has two samples

$$X_1, X_2, \cdots, X_{n_1},$$

and

$$Y_1, Y_2, \cdots, Y_{n_2},$$

which might be weights, blood pressure... The mean effects of these treatments are estimated using the sample means

$$\bar{X}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \quad \text{and} \quad \bar{Y}_{n_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j,$$

which estimate the means

$$\mu_1 = \mathbb{E}(X) \quad \text{and} \quad \mu_2 = \mathbb{E}(Y).$$

Statistical problems:

- Is it true that B is better than A ?, for example is $\mu_2 < \mu_1$?
- Give a confidence interval for the difference $\mu_1 - \mu_2$.

We must again consider various situations: gaussian samples, n_i, $i = 1, 2$ not too large; arbitrary i.i.d. samples with both n_1 and n_2 large etc.

We assume that the two samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are i.i.d., of the form

$$X_i = \mu_1 + \varepsilon_i, \quad i = 1, \cdots, n_1,$$

$$Y_j = \mu_2 + \tilde{\varepsilon}_j, \quad j = 1, \cdots, n_2,$$

where we assume no systematic errors,

$$\mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{E}(\tilde{\varepsilon}_j) = 0,$$

and the existence of variances

$$\mathrm{Var}(\varepsilon_i) = \sigma_1^2, \quad \mathrm{Var}(\tilde{\varepsilon}_j) = \sigma_2^2.$$

The next elements of interest are the sample variances

$$s_{n_1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x}_{n_1})^2,$$

and

$$s_{n_2}^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \bar{y}_{n_2})^2,$$

6.1 Two independent (large) samples

We now turn to statistical questions related to treatment effects. As stated previously, one might be interested in the difference $\mu_1 - \mu_2$. If the medicament is supposed to act against high blood pressure, two natural hypotheses to be tested are

- No treatment effect, that is $H: \quad \mu_2 = \mu_1$;
- Treatment effect, $K: \quad \mu_2 < \mu_1$.

This suggests the introduction of the difference $\delta = \mu_1 - \mu_2$. Then, the two basic hypotheses to be tested are

$$H: \ \delta = \delta_0 = 0 \text{ and } \quad K: \ \delta > 0.$$

The sample difference is such that

$$\mathbb{E}(\bar{X}_{n_1} - \bar{Y}_{n_2}) = \mathbb{E}(\mu_1) - \mathbb{E}(\mu_2) = \mu_1 - \mu_2,$$

and the related variances are given by, using independence,

$$\mathrm{Var}(\bar{X}_{n_1} - \bar{Y}_{n_2}) = \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

When both n_1 and n_2 are large, say larger than 30, we can check that the following statistics is approximatively normal:

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \approx \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

The statistics of interest is thus obtained by *studentizing* the difference to get the statistics

$$Z = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{s_{n_1}^2/n_1 + s_{n_2}^2/n_2}} \tag{6.1}$$

which is apprximatively $\mathcal{N}(0,1)$.
One can then consider a confidence interval

$$I(X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2}),$$

for the mean difference $\mu_1 - \mu_2$, of the form

$$[\bar{x}_{n_1} - \bar{y}_{n_2} - z_{\alpha/2}\sqrt{s_{n_1}^2/n_1 + s_{n_2}^2/n_2}, \bar{x}_{n_1} - \bar{y}_{n_2} + z_{\alpha/2}\sqrt{s_{n_1}^2/n_1 + s_{n_2}^2/n_2}],$$

with

$$P(\mu_1 - \mu_2 \in I(x_1, \cdots, x_{n_1}, y_1, \cdots, y_{n_2})) = 1 - \alpha.$$

We can now proceed as in Section 4, using gaussian approximations, and obtain the tests:

Let $\delta = \mu_1 - \mu_2$. The relevant statistics is given by

$$Z = \frac{\bar{X}_{n_1} - \bar{Y}_{n_1} - \delta_0}{\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2}}},$$

is approximately $\mathcal{N}(0,1)$ and leads to the following table:

Alternative K	Critical domain
$K : \delta > \delta_0$	$R : \ Z > z_\alpha$
$K : \delta < \delta_0$	$R : \ Z < -z_\alpha$
$H_1 : \mu_1 - \mu_2 \neq \rho_0$	$R : \ \lvert Z\rvert > z_{\alpha/2}$

where z_α is given by $\int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t = 1 - \alpha$.

We provide an example where the problem consists in comparing two populations. The reader will check without problem that the above notions can be applied in this new setting:

Example 6.1. Suppose we have the following sufficient summary of ages of the first marriage of women in two tribes A and B where $n_1 = n_2 = 100$. A sufficient summary is provided by the following 2x2 table:

	A	B
Mean	20.7	18.5
Std-Dev	6.3	5.8

One finds that

$$\sqrt{\frac{s_{n_1}^2}{n_1}+\frac{s_{n_2}^2}{n_2}}=0.8563,\ z_{\alpha/2}=1.96,$$

for a confidence interval at level of confidence of 95 %. The confidence interval is given by

$$I=[2.2-1.68,2.2+1.68],$$

which does not contain 0. Using the relation presented in Section 5 between test problems and confidence intervals, we can infer that there is strong evidence that the women of tribe B marry sooner than the woment of tribe A. If one perfom a test for $H:\ \mu_1=\mu_2$ against the alternative $K:\ \mu_1\neq\mu_2$, the test rejects H a the level $\alpha=0.025$ since $0\notin I$.

```
> num
[1] 2.2
> den
[1] 0.8563294
> qlow<-qnorm(c(0.025), mean=0, sd=1,lower.tail=TRUE)
> qlow*den+num
[1] 0.5216253
> -qlow*den+num
[1] 3.878375
```

Example 6.2. Every winter, the roads surrounding a lake are treated with salt, which contains chlorine. The government wishes to get information on the level of chlorine present in the lake. No salt is used during two years. One gets in this way a sample of size $n_1=25$ in 1993, and one of size $n_2=110$ in 1995. The mean and standard deviation are summarized in the following table (concentration of chlorine)

	1993	1995
Mean	18.8	17.8
Std-Dev	1.2	1.8

The basic hypothesis we would like to study is $\mu_2<\mu_1$. We therefore set

$$H:\ \mu_1=\mu_2 \text{ against } K:\ \mu_2<\mu_1.$$

The statistics is again given by

$$Z = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{s^2_{n_2}}{n_2}}},$$

where we denote it as z since we make a normal approximation (both n_1 and n_2 are large). We reject H and thus accept the fact that the treatment is efficient if the observed statstics z is larger than a threshold value. The observed statistics is given by $z = 2.32$, and the p-value is $P(Z \geq 2.32) = 0.0102$. The data indicates thus strong evidence for treatment effect.

In what follows we give methods for dealing with small samples.

6.2 Comparison of two small samples

When the sample sizes are small, one can not rely on gaussian approximations. We therefore assume here that:

1. The populations are both normal and independent,
2. they have the same variance σ^2.

The statistical translation of these statements is that the two samples

$$(X_i)_{1 \leq i \leq n_1} \text{ and } (Y_i)_{1 \leq i \leq n_2},$$

are i.i.d. gaussian $\mathrm{N}(\mu_1, \sigma^2)$ and $\mathrm{N}(\mu_2, \sigma^2)$, with, using independence,

$$\mathrm{Var}(\bar{X}_{n_1} - \bar{Y}_{n_2}) = \sigma^2(1/n_1 + 1/n_2).$$

The *pooled* variance

$$s^2_{\text{pooled}} = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_{n_1})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y}_{n_2})^2}{n_1 + n_2 - 2}$$

is an unbiased estimator of σ^2, and the above statistics is distributed as a Chi-square random variable $\mathcal{X}^2_{n_1+n_2-2}$ of $n_1 + n_2 - 2$ d.f. Then, the statistics

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\mu_1 - \mu_2)}{s_{\text{pooled}}\sqrt{1/n_1 + 1/n_2}}$$

follows a Student distribution of $n_1 + n_2 - 2$ d.f. For more information on these various distributions, the reader is encouraged to solve the exercices. The related confidence intervals and tests are obtained as in Section 6.1 by replacing the quantiles of the standard normal $\mathcal{N}(0,1)$ distribution by the quantiles t_{α,n_1+n_2-2} of the related Student distribution.

7 Comparison of two treatments: the Mann-Whitney Wilcoxon test

Suppose we must compare two treatments A and B, when no parametric information is available. The group of patients is divided in two subgroups of sizes n_1 and n_2; the patients of the first group receive treatment A while those of the second group receive treatment B. This leads to two i.i.d. samples

$$(X_i)_{1\leq i\leq n_1} \text{ and } (Y_i)_{1\leq i\leq n_2}.$$

Statistical model: We assume here that the random variable X has some continuous density $f(x)$, and that Y has a density $g(y)$, which is a shifted version of f, that is

$$g(y) = f(y - \triangle),$$

for some parameter $\triangle$. When $\triangle > 0$, Y *is stochastically larger than* X, which means that

$$P(X < t) \geq P(Y < t), \quad \forall t,$$

since

$$\begin{aligned} P(Y < t) &= \int_{-\infty}^{t} g(y)\mathrm{d}y \\ &= \int_{-\infty}^{t} f(y - \triangle)\mathrm{d}y \\ &= \int_{-\infty}^{t-\triangle} f(x)\mathrm{d}x \\ &= P(X < t - \triangle) \\ &\leq P(X < t). \end{aligned}$$

The hypotheses of interest are the null hypothesis of no difference between A and B,

$$H : \triangle = 0,$$

and the alternative stating that Y is stochastically larger than X with $\triangle > 0$,

$$K : \triangle > 0.$$

Rank statistics: We merge the two samples, and order this global sample to get

$$Z_{(1)} < Z_{(2)} < \cdots Z_{(n_1+n_2)}.$$

Some of the $Z_{(i)}$ correspond to the Y random variables. For each data of the intial samples, let R_j denote the rank of the jst data.

Statistical procedure: The Wilcoxon test rejects he null hypothesis $H : \triangle = 0$ when the sum of the ranks associated to the data of group A is smaller than some threshold value, that is when

$$W_A := \sum_{j \in A} R_j < c.$$

Let $N = n_1 + n_2$ be the size of the global sample which includes data from A and B. As

$$\sum_{j \in A} R_j + \sum_{j \in B} R_j = \frac{N(N+1)}{2},$$

one sees that it is equivalent to reject the hypothesis of no difference when

$$W_B := \sum_{j \in B} R_j > \tilde{c},$$

for some threshold value $\tilde{c}$, to be fixed in order to get a acceptable level of significance.

When the null hypothesis is true, that is when $\triangle = 0$, all the possible subset $\{R_j;\ j \in A\}$ are equiprobable. This permits to fix the constant c in order to get the level α, see the example below.

Example 7.1. Suppose we have data from two groups: $B = 31.8, 39.1$ and $A = 35.5, 27.6, 21.3$. Then on gets

Data	21.3	27.6	31.8	35.5	39.1
Rank	1	2	3	4	5
Group	A	A	B	A	B

Here, $W_B = \sum_{j \in B} R_j = 3 + 5 = 8$. We then consider the possible values that W_B can take:

Ranks	Sum of Ranks
1,2	3
1,3	4
1,4	5
1,5	6
2,3	5
2,4	6
2,5	7
3,4	7
3,5	8
4,5	9

which yields, under H, the law of W_B, as given by the following table

Sum of Ranks	Probability
3	1/10
4	1/10
5	1/5
6	1/5
7	1/5
8	1/10
9	1/10

Hence for example, $\mathrm{P}_{H_0}(W_B \geq 8) = 1 - \mathrm{P}(W_B < 8) = 1 - \mathrm{P}(W_B \leq 7) = 1 - 0.8 = 0.2$.

Mathematica code

```
In[1]:=
<<DiscreteMath‘Combinatorica‘
In[7]:=
l=KSubsets[Table[i,{i,1,5}],2];
```

```
Length[l]
Out[8]=
56
In[9]:=
myfunc[l_]:=ToString[l[[1]]]<>","<>ToString[l[[2]]]
In[10]:=
pivot[l_]:=Module[{sums,temp,cumsum},
    sums=Map[Apply[Plus,#]&,l];
    temp=Map[
        {#,
            Count[sums,#]/Length[sums]
            }&,Union[sums]];
    cumsum=Rest[N[FoldList[Plus,0,Transpose[temp][[2]]]]];

    Transpose[{
       Prepend[Transpose[temp][[1]],"Rank Sum"],
   Prepend[Transpose[temp][[2]],"Proba="],
   Prepend[cumsum,"proba<="]}]//
      TableForm
    ]
In[11]:=
pivot[l]
```

Which gives the following output

Rank Sum	Proba=	Proba $\leq$
3	$\frac{1}{10}$	0.1
4	$\frac{1}{10}$	0.2
5	$\frac{1}{5}$	0.4
6	$\frac{1}{5}$	0.6
7	$\frac{1}{5}$	0.8
8	$\frac{1}{10}$	0.9
9	$\frac{1}{10}$	1.

The above examples are based on artificial data. The following example is more likely to be realistic:

Example 7.2. Geologists study two different formations A and B, and look for the concentration in minerals. The data is given

by

$$A: \ 7.6, \ 11.1, \ 6.8, \ 9.8, \ 4.9, \ 6.1, \ 15.1,$$

and

$$B: \ 4.7, \ 6.4, \ 4.1, \ 3.7, \ 3.9.$$

Here they suspect that formation A is richer in minerals. The model thus assume that the density related to group A is obtained from B by a shift to the right. The critical domain is thus of the form

$$\{W_B < c\}.$$

We choose the statistics W_B insteed of W_A since group B as the smallest number of elements, which makes computations more easy. The observed sum of ranks is $W_B = 17$. On the other hand, direct computation shows that

$$c = 22: \ P_H(W_B \leq c) = 0.053,$$

$$c = 21: \ P_H(W_B \leq c) = 0.037.$$

The test rejects the null hypothesis of no difference at level $\alpha = 3.7\%$; there is thus a strong evidence that formation A contains more minerals than formation B.

Asymptotic normality

In the same setting, assume that the sizes n_1 and n_2 of groups A and B are large. It can be shown that

$$Z = \frac{W - n_1 \frac{n_1+n_2+1}{2}}{\sqrt{n_1 n_2 \frac{n_1+n_2+1}{2}}}$$

is approximatively standard normal, i.e. $Z \sim \mathcal{N}(0,1)$. We can the use this normal approximation to fix the test.

8 Sign test for small dependent samples

Sign tests are designed for comparing two treatments when no gaussian approximation is available, and when the two samples are dependent. We focus on an example involving *diastolic* P_d

and *systolic* P_s blood pressures. The average pressure is defined by $\frac{2}{3}P_d + \frac{1}{3}P_s$.
In the experiment, a sample of n=18 students is selected; one then measure for each student the average pressure when lying down and when standing up. The data consists in a collection of $n = 18$ pairs $z_i = (x_i, y_i)$, $i = 1, \cdots, n$, each being the realization of a random vector (X, Y). The aim of the study is to see if Y is *stochastically larger* than X, that if

$$P(X < t) \geq P(Y < t), \quad \forall t.$$

We thus use the differences

$$Z_i := Y_i - X_i, \ i = 1, \cdots, n,$$

which should be significatively positive. *The null hypothesis of no difference* is here modeled as

$$H: \ P(Z_i > 0) = P(Z_i < 0) = \frac{1}{2}.$$

The alternative of interest is

$$K: \ P(Z_i > 0) > \frac{1}{2}.$$

The relevant statistics is the number of positive differences

$$V := \sum_{i=1}^{n} \mathbb{I}(Z_i > 0),$$

where $\mathbb{I}$ is the indicator function.

Decision rule: The (unilateral) test considered here rejects the hypothesis H of no difference when the number of positive differences V is too large.

$$\phi(x) = \begin{cases} 1 & \text{if } V \geq c \\ 0 & \text{otherwise.} \end{cases}$$

Given a level α, the test is fixed by choosing c_α so that $P_H(V \geq c_\alpha) \leq \alpha$.

Law of V under H: To fix the test, one needs the law of V under H. The random variables $\varepsilon_i = \mathbb{I}(Z_i > 0)$, are of Bernoulli type with probability of success (under H) given by

$$P(\varepsilon = 1) = P(\varepsilon = 0) = \frac{1}{2}.$$

so that V is Binomial Bi(n,1/2). Hence

$$P(V \geq c) = \sum_{k=c}^{n} \binom{n}{k} \frac{1}{2^n},$$

which can be obtained from standard tables of the binomial distribution when n is not too large, or by using the Central Limit Theorem when n is large (exercice).

Table (observed differences)

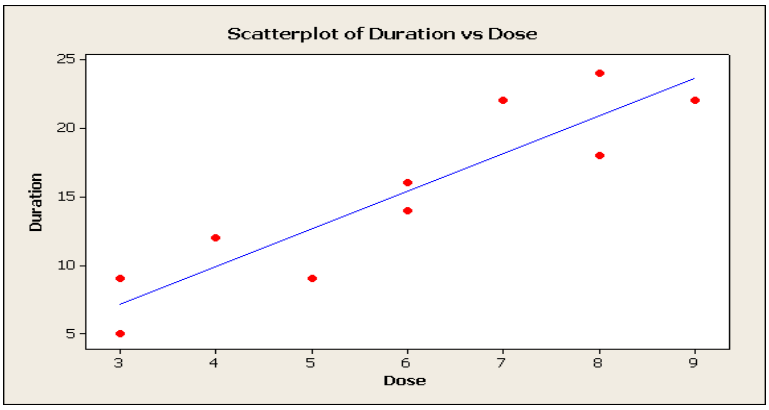
Student	1	2	3	4	5	6	7	8	9
Differences	$1\frac{2}{3}$	$1\frac{2}{3}$	$2\frac{2}{3}$	$4\frac{1}{3}$	$-2\frac{2}{3}$	$-3\frac{1}{3}$	$-8\frac{1}{3}$	$-1\frac{2}{3}$	$5\frac{1}{3}$
Student	10	11	12	13	14	15	16	17	18
Differences	5	-5	$1\frac{2}{3}$	$1\frac{2}{3}$	$3\frac{1}{3}$	5	$\frac{1}{3}$	$1\frac{2}{3}$	-5

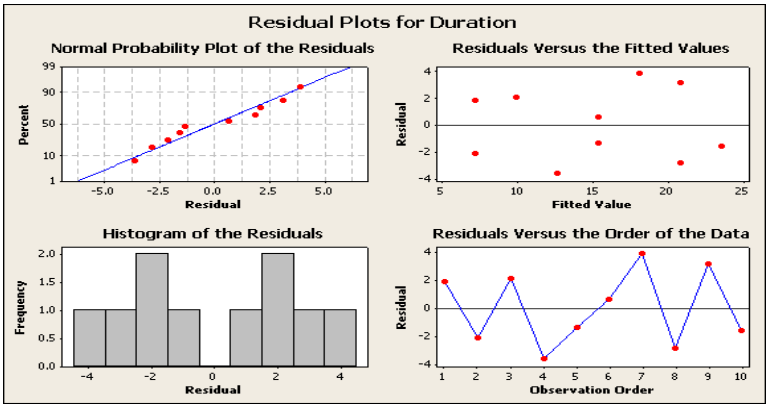
Assume that $\alpha = 5\,\%$. Using a table for $Bi(n,p)$, one gets the threshold $c_{0,05} = 13$ ($P_H(V \geq c_{0,0.5}) \leq 0,05$). Our experiment yields $V = 12$, so that H is not rejected at the level $\alpha = 5\%$.

9 Regression

We have already considered regression in the first part of this course. Regression analysis allows to predict one variable from the value of another variable.

Example 9.1. In one stage of the development of a new drug for an allergy, an experiment is conducted to study how different dosages of the drug affect the duration of relief from the allergic symptoms. Ten patients are included in the experiment. Each





patient receives a specified dosage of the drug and is asked to report back as soon as the protection of the drug seems to wear off. The observation are recorded in the following table, which shows the dosage x and duration of relief y for the $n = 10$ patients.

Dosage x [mg]	Duration of relief [days]
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

After inspection of the data, we look for a straight line of the

form

$$y = \beta_0 + \beta_1 x,$$

which should explain the data. To proceed corectly, we must set a valid statistical model, which is given here by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, n, \tag{9.1}$$

where we assume that the measurment errors are i.i.d. such that

- $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$,
- and the parameters β_0 and β_1 are unknown.

We use here the least squares method, which looks for β_k minimizing the sum of squares

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2.$$

We have already considered such kind of models when dealing with estimation, and the general model was given by

$$Y = A\theta + \varepsilon,$$

for some design matrix $A \in \mathbb{R}^{nxp}$, where $\theta \in \mathbb{R}^p$ was the unknown parameter. Under weak conditions on the rank of A, we have seen that the least squares estimate was given by

$$\hat{\theta} = (A^* A)^{-1} A^* Y,$$

which is linear in the observation Y. We can thus obtain directly a formula for (9.1) in the straight line case, and find after some computations that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n,$$

where we set

$$S_{xx} = \sum_i (X_i - \bar{X}_n)^2, \quad S_{xy} = \sum_i (X_i - \bar{X}_n)(Y_i - \bar{Y}_n),$$

and where S_{yy} is similarly defined. Proceeding in this way, we can thus find the optimal straight line. The statistical problem

is however not closed. The next step consists in checking **the residuals**

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i,\ i = 1, \cdots, n, \tag{9.2}$$

which should look like a typical sample drawn from a normal population. The next quantity of interest is the **residual sum of squares SSE** given by

$$\text{SSE} = \sum_{i=1}^{n} \bar{\varepsilon}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \tag{9.3}$$

Notice that we also already encountered this kind of expressions in the first semester when dealing with best least squares approximations of a random variable Y by an affine transformation of a second random variable X.
This provides an unbiaised estimator for the variance σ^2, given by

$$s_n^2 = \frac{\text{SSE}}{n-2}, \tag{9.4}$$

The natural question here is why we divide SSE by $(n-2)$? Intuitively, this is because two degree of freedom are lost from estimating the two parameters β_0 and β_1. Mathematically, the residual sum of squares SSE involves terms of the form $(Y_i\hat{\beta}_0 - \beta_1 X_i =^2$, where the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of the random variables Y_i, and are therefore correlated. We will come to this point later when dealing with ANOVA.

Example 9.2. Coming back to our basic Example 9.1, we find that

$$\bar{X}_n = 5.9,\ \bar{Y}_n = 15.1,$$

$$S_{xx} = 40.9,\ S_{yy} = 370.9,\ S_{xy} = 112.1,\ \text{SSE} = 63.6528,$$

yielding the estimates

$$\hat{\beta}_0 = -1.07 \text{ and } \hat{\beta}_1 = 2.74.$$

9.1 Inference problems

Having performed the numerical computations, hence having the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, one next consider related test problems.

These estimates are random, and have variances and standard deviations or standard errors (S.E.), given by

$$\text{S.E.}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{xx}}}, \tag{9.5}$$

and

$$\text{S.E.}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}, \tag{9.6}$$

where the unknown standard deviation σ is estimated by

$$s_n^2 = \sqrt{\frac{\text{SSE}}{n-2}}.$$

For test problems related to the slope β_1, one uses the t statistics

$$T = \frac{\hat{\beta}_1 - \beta_1}{s_n/\sqrt{S_{xx}}}, \quad d.f. = n-2,$$

which has a Student distribution of $\nu = n-2$ degrees of freedom. Concerning the intercept, one uses similarly the statistics

$$T = \frac{\hat{\beta}_0 - \beta_0}{s_n\sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{xx}}}}, \quad d.f. = n-2,$$

which has a Student distribution of $\nu = n-2$ degrees of freedom.

9.1.1 Inference about the intercept β_1

A basic problem in this setting is to check whether the random variable Y does or does not vary withe magnitude of the input variable X. According to the model,

$$\text{Expected response} = \beta_0 + \beta_1 x,$$

and the response is independent of the input if and only if $\beta_1 = 0$. We can thus consider the test problem

$$H_0 : \beta_1 = 0,$$

against some alternative hypothesis, to be chosen according to the setting. Under H_0, the relevant test statistics is then given by

$$T = \frac{\hat{\beta}_1}{s_n/\sqrt{S_{xx}}}, \quad d.f. = n-2.$$

Example 9.3. We consider again Example 9.1: *Do the data constitute strong evidence that the mean duration of relief increases with higher dosages of the drug ?* The data suggests that it is in fact the case, and we put this statement in the alternative hypothesis K. Mathematically, admitting a straight line model, the line is incrasing when $\beta_1 > 0$, so that we set here

$$K: \ \beta_1 > 0.$$

We look for the quantile of the Student distribution of $\nu = n - 2$ degrees of freedom to get $t_{0.05,n-2} = 1.860$, and compute the t-statistics to obtain that $t = 6.213$, so that the null hypothesis H_0 is strongly rejected. The P-value is much smaller that 0.005 !

More generally, assume that we can test whether or not β_1 is equal to some specified value $\tilde{\beta}_1$. The relevant statistics is similarly given by

$$H_0 : \beta_1 = \tilde{\beta}_1, \quad T = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{s_n/\sqrt{S_{xx}}}, \ d.f. = n - 2.$$

In addition to testing hypotheses, we can also provide a confidence interval for the slope using the t distribution: A 100(1-α) **confidence interval** for the slope is given by

$$(\hat{\beta}_1 - t_{\alpha/2,n-2}\frac{s_n}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\frac{s_n}{\sqrt{S_{xx}}}),$$

where $t_{\alpha/2,n-2}$ is the upper $\alpha/2$ point of the t distribution with d.f.=n-2.

Example 9.4. We can construct a 95 % confidence interval for the slope β_1 of the regression line considered in Examples **??** and 9.3. The numerical values yield the confidence interval

$$(1.72, 3.76).$$

We are 95% confident that by adding one extra milligram to the dosage, the mean duration of relief would increase somewhere between 1.72 and 3.76 days.

9.1.2 Inference about the intercept β_0

The procedures are again based on the t distribution with $\nu = n-2$ d.f. A 100(1-α) **confidence interval for the interecpt** β_0 is given by

$$(\hat{\beta}_0 - t_{\alpha/2,\nu} s_n \sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{xx}}}, \hat{\beta}_0 + t_{\alpha/2,\nu} s_n \sqrt{\frac{1}{n} + \frac{\bar{X}_n^2}{S_{xx}}}).$$

9.2 The strength of a linear regression

We examine how much of the variation in the response is explained by the fitted regression line. To this end, we decompose the observed values as follows

$$Y_i = (\hat{\beta}_0 + \hat{\beta}_1 X_i) + (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i), \tag{9.7}$$

The residual sum of squares

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

is considered as a measure of the discrepancy or departure from linearity. The total variability of the y values is reflected in the sum of squares

$$S_{yy} = \sum_i (Y_i - \bar{Y})^2.$$

The Decomposition of Variability is here defined by

$$\begin{aligned} &\textbf{Total variability of Y} \;\; S_{yy} \\ &= \textbf{Variability explained by the linear model} \;\; \frac{S_{xy}^2}{S_{xx}} \\ &\quad + \textbf{Residual or unexplained variability SSE}. \end{aligned}$$

The proportion of the y-variability explained by the linear regression is then given by

$$\frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2,$$

where r is the **sample correlation coefficient**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

with $0 \leq r^2 \leq 1$.

Example 9.5. Consider again the drug problem. We had

$$S_{xx} = 40.9,\ S_{yy} = 370.9,\ S_{xy} = 112.1,$$

and the fitted regression line was given by

$$\hat{y} = -1.07 + 2.74x.$$

How much of the variability in y is explained by the linear regression model ?
We can compute the related correlation coefficient

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = 0.83,$$

which means that 83 % of the variability in y is explained by the linear regression, and the model seems satisfactory in this respect.

10 Testing for a trend

10.1 Spearman rank correlation

Given a sample $(X_i)_{1\leq i\leq n}$, we are here interested in testing

H : the sample is i.i.d. with a continuous distribution function,

against the alternative

K : the data (i, X_i) suggests a monotonic trend.

Notice that under K, the r.v. X_i might be dependent. Alternatives like K occur often in applications; the data seems to indicate a monotonic trend, but no explicit model can be given. A typical setting where such phenomenon occurs is in linear regression, where the statistical model is given by

$$X_i = ai + b + \varepsilon_i, \quad i = 1, \cdots, n,$$

where the ε_i are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$. The standard way of estimating the parameter $\theta = (a, b)$ is to use the least squares method. The *coefficient of correlation* provides a measure of departure from linearity, and is defined by

$$\rho = \frac{\frac{1}{n}\sum_i (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X}_n)^2}\sqrt{\frac{1}{n}\sum_i (Y_i - \bar{Y}_n)^2}}.$$

The idea of the Spearman's correlation test is to use the statitics given by the correlation coefficient applied to the data (i, R_i) where R_i denotes the rank of the ist data. When for example (i, X_i) is increasing, that is when $X_{i+1} > X_i$, $R_i \equiv i$, which is linear, so that the correlation coefficient is equal to 1. When it is decreasing, $\rho = -1$.

Let R_i denote the rank of the ist observation, $R = (R_1, \ldots, R_n)$ and $I = (1, \ldots, n)$. Define the rank correlation $\rho_{\mathcal{S}} = \mathrm{Cor}(R, I)$, i.e.

$$\rho_{\mathcal{S}} = \frac{\sum_{i=1}^n (R_i - \bar{R}_n)(i - \bar{i})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R}_n)}\sqrt{\sum_{i=1}^n (i - \bar{i})}}.$$

Lemma 10.1.

$$\bar{R} = \frac{1}{n}\sum_{i=1}^n R_i = \bar{i} = \frac{1}{n}\sum_{i=1}^n i = \frac{(n+1)}{2},$$

$$\begin{aligned}\sum_{i=1}^n (R_i - \bar{R})^2 &= \sum_{i=1}^n (i - \bar{i})^2 = \sum_{i=1}^n i^2 - (\frac{n+1}{2})^2 \\ &= \frac{n(n^2-1)}{12},\end{aligned}$$

$$\sum_{i=1}^n (i - R_i)^2 = 2\sum_{i=1}^n (i^2 - iR_i),$$

$$\rho_{\mathcal{S}} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^n (R_i - i)^2 \leq 1.$$

The test is next based on the three usual possibilites

- Positive dependence, that is the data is increasing, $\rho_{\mathcal{S}}$ is close to 1,
- arbitrary monotonic dependence, either increasing or decreasing, $|\rho_{\mathcal{S}}|$ close to 1,
- negative dependence, that is the data is decreasing, $\rho_{\mathcal{S}}$ close to -1.

We set accordingly the critical domain of the test as

Pos. Dep.	Arbitrary Dep.	Neg. Dep.
$R = \{\rho_{\mathcal{S}} > c\}$	$R = \{\|\rho_{\mathcal{S}}\| > c\}$	$R = \{\rho_{\mathcal{S}} < c\}$

When using such a test, one must choose the threshold c to get a satisfactory significance level. One must for example compute the probability

$$P_H(\rho_{\mathcal{S}} > c).$$

Under H, the random variables X_i are i.i.d., so that the distribution of the random variable $\rho_{\mathcal{S}}$ is *symmetric*, that is

$$\mathcal{L}(\rho_{\mathcal{S}}) = \mathcal{L}(-\rho_{\mathcal{S}}).$$

To check this, first notice that under H,

$$\mathcal{L}(X_1, \cdots, X_n) = \mathcal{L}(X_{\sigma(1)}, \cdots, X_{\sigma(n)}),$$

for any permutation σ of $[n]$. Writing $\tilde{\rho}_{\mathcal{S}}$ for the rank correlation coefficient applied to the permuted data, one gets

$$\mathcal{L}(\rho_{\mathcal{S}}) = \mathcal{L}(\tilde{\rho}_{\mathcal{S}}).$$

Choosing the special permutation σ given by $\sigma(i) = n + 1 - i$, one gets that, denoting by $\tilde{R}_i$ the rank associated with the ist permutated data,

$$\sum_i (\tilde{R}_i - \bar{\tilde{R}})^2 = \sum_i (R_i - \bar{R})^2,$$

and therefore

$$\begin{aligned}
\tilde{\rho}_{\mathcal{S}} &= \frac{\sum_i (\tilde{R}_i - \bar{R})(i - \bar{i})}{\sqrt{\sum_i (\tilde{R}_i - \bar{R})^2}\sqrt{\sum_i (i - \bar{i})^2}} \\
&= \frac{\sum_i (R_{n+1-i} - \bar{R})(i - \bar{i})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (i - \bar{i})^2}} \\
&= -\frac{\sum_i (R_{n+1-i} - \bar{R})(n+1-i - \bar{i})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (i - \bar{i})^2}} \\
&= -\rho_{\mathcal{S}},
\end{aligned}$$

where we have used the identity

$$i - \bar{i} = i - \frac{n+1}{2} = i - n - 1 + \frac{n+1}{2} = -(n+1-i-\bar{i}).$$

Having established the symmetry of the law of $\rho_{\mathcal{S}}$ under H, one gets that all odd moments of $\rho_{\mathcal{S}}$ vanish, that is

$$\begin{aligned}
\mathbb{E}(\rho_{\mathcal{S}}^{2k+1}) &= \int_{\rho_{\mathcal{S}}>0} \rho_{\mathcal{S}}^{2k+1} dP + \int_{\rho_{\mathcal{S}}<0} \rho_{\mathcal{S}}^{2k+1} dP \\
&= \int_{\rho_{\mathcal{S}}>0} \rho_{\mathcal{S}}^{2k+1} dP - \int_{\rho_{\mathcal{S}}<0} (-\rho_{\mathcal{S}})^{2k+1} dP \\
&= \int_{\rho_{\mathcal{S}}>0} \rho_{\mathcal{S}}^{2k+1} dP - \int_{-\rho_{\mathcal{S}}>0} (-\rho_{\mathcal{S}})^{2k+1} dP \\
&= 0,
\end{aligned}$$

by symmetry.

The idea here consists in finding some function $f(n)$ so that the rescaled random variable

$$f(n)(\rho_{\mathcal{S}} - \mathbb{E}(\rho_{\mathcal{S}})) \approx \mathrm{N}(0,1).$$

If such a statement is correct, we can then fix the threshold of Spearman's rank correlation test to the the required significance level.

The speed $f(n)$

We will need the following Theorem:

Theorem 10.1 (Carleman)**.** *Given a sequence of random variables X_n with*

$$\lim_{n\to\infty} \mathbb{E}(X_n^k) = \mathbb{E}(X^k) \qquad \forall k,$$

where X is some random variable satisfying Carleman's condition

$$\limsup_k \frac{(\mathbb{E}(X^k))^{\frac{1}{2k}}}{2k} = r < \infty. \tag{10.1}$$

Then X_n converges in law to the law of X, i. e. $X_n \overset{\mathcal{L}}{\Rightarrow} \mathcal{L}(X)$.

Our idea is to use the above Theorem to show that the sequence of probability measures associated to the rescaled random variables $f(n)\rho_{\mathcal{S}}$ weakly converges toward the law of a normal random variable. It remains to study the asymptotic behavior of the even moments
The first step consists in showing that

$$f(n)^{2k}\,\mathbb{E}(\rho_{\mathcal{S}}^{2k}) \longrightarrow \frac{(2k)!}{2^k k!}, \tag{10.2}$$

where the limit corresponds to the moment of order $2k$ of the standard normal random variable. If this statement is correct, for a well chosen function $f(n)$, one must have

$$f(n)^2 \mathrm{Var}(\rho_{\mathcal{S}}) \longrightarrow 1,$$

which will give us the function $f(n)$.
This is the good time for introducing computations based on ranks:

Lemma 10.2.

$$\mathrm{Var}(\rho_{\mathcal{S}}) = \frac{1}{n-1},$$

so that

$$f(n) = \sqrt{n-1}.$$

Proof: We first use some previously established identities like

$$\mathrm{Var}(\rho_{\mathcal{S}}) = \mathrm{Var}(1 - \frac{6}{n(n^2-1)} \sum_i (R_i - i)^2),$$

and

$$\mathrm{Var}(\rho_{\mathcal{S}}) = \frac{36\ 4}{n^2(n^2-1)^2}\mathrm{Var}(\sum i^2 - \sum iR_i),$$

where we use the fact that

$$\frac{1}{2}\sum_i (R_i - i)^2 = \sum i^2 - \sum iR_i,$$

so that

$$\mathrm{Var}(\rho_{\mathcal{S}}) = \Big(\frac{12}{n(n^2-1)}\Big)^2 \mathrm{Var}(\sum iR_i).$$

We will need the decomposition

$$\mathrm{Var}(\sum iR_i) = \sum_i \mathrm{Var}(iR_i) + \sum_{i\neq j} \mathrm{Cov}(iR_i, jR_j).$$

These covariances are computed by first looking for the law of the ranks, under H.

Assume that the distribution function F of the i.i.d. data is continuous. We can then write

$$R_i = \sum_j \mathbb{I}_{\mathbb{R}^+}(X_i - X_j), \tag{10.3}$$

where we assume without loss of generality that $X_i \neq X_j$ when $i \neq j$ since F is continuous. We also use the *exchangeability* of the collection of random variables X_i, that is the invariance of the law of the random vector $X = (X_1, \cdots, X_n)$ under permutations, since, under H,

$$\mathcal{L}(X_1, \cdots, X_n) = \mathcal{L}(X_{\sigma(1)}, \cdots, X_{\sigma(n)}),$$

for every permutation σ of $[n]$. Let $R = (R_1, \cdots, R_n)$ be the random permutation associated the random vectors associated with the ranks, that is the vector giving the rank R_i for each data X_i. First notice that the law of R is the uniform distribution on the permutation group, as a consequence of the exchangeability. Next, the marginal probability $P(R_i = k)$ can be obtained as follows:

$$P(R_i = k) = \sum_\sigma P(R_i = k, R = \sigma) = \frac{(n-1)!}{n!} = \frac{1}{n},$$

$$P(R_i = k, R_j = l) = \frac{1}{n(n-1)}, \quad \text{when } i \neq j,$$

so that

$$\mathbb{E}(R_i) = \sum_k kP(R_i = k) = \frac{1}{n}\sum_k k = \frac{n(n+1)}{2n} = \frac{n+1}{2},$$

and

$$\mathrm{Var}(R_i) = \mathbb{E}(R_i^2) - \mathbb{E}(R_i)^2 = \frac{1}{n}\sum_k k^2 - (\frac{n+1}{2})^2 = \frac{n^2-1}{12}.$$

Furthermore,

$$\begin{aligned}
\mathrm{Cov}(R_i, R_j) &= \mathbb{E}(R_i R_j) - (\frac{n+1}{2})^2 \\
&= \sum_{k \neq l} kl \frac{1}{n(n-1)} - (\frac{n+1}{2})^2 \\
&= \frac{1}{n(n-1)}(\sum_k k)^2 - \sum_k k^2) - (\frac{n+1}{2})^2 \\
&= -\frac{n+1}{2}.
\end{aligned}$$

Finally,

$$\mathrm{Var}(\sum_i iR_i) = \mathrm{Var}(R_1)\sum_i i^2 - \sum_{i \neq j} ij\mathrm{Cov}(R_i, R_j).$$

Using all of these identities, one finally obtain that

$$\mathrm{Var}(\rho_{\mathcal{S}}) = \frac{1}{n-1},$$

as required. □

Having found the correct speed $f(n)$, one can prove the following

Theorem 10.2. *Under H, the random variable*

$$\frac{\rho_{\mathcal{S}}}{\sqrt{\mathrm{Var}(\rho_{\mathcal{S}})}} = \sqrt{n-1}\rho_{\mathcal{S}},$$

converges in distribution toward a standard gaussian N(0,1) random variable.

The proof of this Theorem involves even moment computations, but we will accept it without further discussions. □

10.2 The sign test

We again consider the probleme of testing H: the sample is i.i.d. against the alternative of a monotonic trend. Under H, one has

$$P(X_i < X_{i+1}) = P(X_i > X_{i+1}) = 1/2.$$

Under K, the difference $X_i - X_{i+1}$ will show a preference for some sign. This suggests the statistics S given by

$$S = \sum_{i=1}^{n-1} \mathbb{I}_{\mathbb{R}^+}(X_i - X_{i+1}),$$

which gives the number of positive differences in the data $Z_i = \mathbb{I}_{\mathbb{R}^+}(X_i - X_{i+1})$. If the data is increasing, $S \simeq 0 = S_{\min}$ while when the data is decreasing, $S \simeq n-1 = S_{\max}$. Since under H, $P(Z_i = 0) = P(Z_i = 1) = 1/2$, it follows that $\mathbb{E}(S) = \frac{n-1}{2}$. The random variables Z_i are correlated. We again use the decomposition

$$\mathrm{Var}(S) = \sum_i \mathrm{Var}(Z_i) + \sum_{i \neq j} \mathrm{Cov}(Z_i, Z_j).$$

Suppose first that $i + 1 \leq j$: Then

$$\begin{aligned} \mathrm{Cov}(Z_i, Z_j) &= \mathbb{E}(Z_i Z_j) - \mathbb{E}(Z_i)\,\mathbb{E}(Z_j) \\ &= P(Z_i = Z_j = 1) - \frac{1}{4} \\ &= P(X_i > X_{i+1}, X_j > X_{j+1}) - \frac{1}{4} \end{aligned}$$

When $i + 1 < j$,

$$P(X_i > X_{i+1}, X_j > X_{j+1}) = \frac{1}{4},$$

and

$$\mathrm{Cov}(Z_i, Z_j) = \frac{1}{4} - \frac{1}{4} = 0,$$

while if $i + 1 = j$, one has

$$P(X_i > X_{i+1}, X_j > X_{j+1}) = \frac{1}{3!},$$

and therefore

$$\mathrm{Cov}(Z_i, Z_j) = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12}.$$

Finally one obtains that

$$\mathrm{Var}(S) = \frac{n-1}{12}.$$

Again, by means of Carleman's theorem and moments computations, one can prove that

$$\frac{S - \mathbb{E}(S)}{\sqrt{\mathrm{Var}(S)}} = \left(S - \frac{n-1}{2}\right)\sqrt{\frac{12}{n-1}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

This result permits to fix the threshold of the test based on the sign statistics S.