eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
FACULTY OF SCIENCE AND TECHNOLOGY
UNIVERSITY OF THE BASQUE COUNTRY

Lecture Notes of Statistical Inference

3rd Year of the Degree in Mathematics

Josu Doncel Vicente

Department of Mathematics

Faculty of Science and Technology

University of the Basque Country, UPV/EHU

Updated version on May 2025

Author contact information:

Email: josu.doncel@ehu.eus
Web: http://josudoncel.github.io/
Phone: (+34) 94 601 78 95
Address:
University of the Basque Country (UPV/EHU)
Faculty of Science and Technology
Mathematics Department
Barrio Sarriena s/n, 48940 Leioa, Spain.

Contents

1 Sampling and Estimations

1.1 Introduction to sampling

1.1.1 General notions about statistical inference

Statistical inference consists of a set of methods with which one can make inferences or generalizations of the results obtained for a sample to the entire population with a certain degree of probability. If a character of a population is studied with the elements of a sample, the result through the inference leads to an error, its complement is the degree of reliability. We will study two methods to deal with this problems: the estimation of parameters and the hypothesis tests.

In general terms, there are two types of sampling: probabilistic sampling (each element of the population has a probability of being selected) and non-probability sampling (samples casual, voluntary or expert selection).

The inference is important because in many cases it is not possible to study a character of the population taking all its components, that is, it is not always possible to carry out a census. Some examples:

- Destructive test studies: let X be the life time of a bulb, it cannot be test all production to estimate average lifetime.
- Elements that can exist conceptually: let X be the number of defective parts that a machine will produce. To estimate the average number of defective parts do not you have all the production; the process begins and at a given moment there are n.
- In very large populations, taking all the elements is unfeasible in time or cost.

Let us now study what is probability, descriptive statistics and statistical inference.

Probability

In probability, we suppose there is a population from which a perfectly described character is studied. This character is defined by a random variable.

Example 1. *In an industrial process, the diameter of a device is controlled. The buyer establishes in its specifications that the diameter must be* 3 ± 0.01 *cm and none is accepted part that deviates from this specification. It is known that the diameter follows a normal distribution* $\mathcal{N}(3, 0.005)$*. What percentage of devices will be discarded?*

Answer: Since $\mathbb{P}(3 - 0.01 < X < 3 + 0.01) = \mathbb{P}(-2 < Z < 2) = 0.9544$, the probability that a device is rejected is 0.0456.

Descriptive Statistics

In descriptive statistics, there is a population from which a character is studied of which the random variable that describes it is not known. Thus, a sample of size n is taken and if the character is quantitative (continuous or discrete) and is represented by a statistical variable. The qualitative character is represented by categorical variables or attributes and the modalities are not measurable. There is a descriptive analysis of the sample obtained from the population.

Example 2. *In an industrial process, a sample of size 9 is taken from devices whose diameters are 3.01, 2.97, 3.03, 3.04, 2.99, 2.98, 2.99, 3.01 and 3.03 cm. find the diameter average.*

Answer: The diameter average of the sample is $\bar{X} = 3.0056$.

Statistical Inference

In statistical inference, there is a population from which a character is studied of which the random variable that describes it is partially known. For example, it is known the shape of the random variable that describes it but not the value of all its parameters. Using inference methods, these values are estimated and the population is perfectly defined.

Example 3. *A machine produces cylindrical metal parts. We take a sample of pieces whose diameters are 3.01, 2.97, 3.03, 3.04, 2.99, 2.98, 2.99, 3.01 and 3.03 cm. Find the average diameter of production if a distribution is assumed to be approximately normal with deviation 0.005.*

Answer: There is a random variable X $\mathcal{N}(\mu, 0.005)$ and a sample of 9 elements. The best estimator of μ is $\bar{X}$, as it will be shown in this course. Therefore, we consider $\mu = \bar{X} = 3.0056$ and the variable that defines the character of a population is perfectly defined.

1.1.2 Types of Sampling

Within the methods of probabilistic sampling, simple random sample is studied here: each element has the same probability of being chosen. The population must be infinite or else the sampling will be carried out with replacement.

Simple random sampling is used when the elements of the population are homogeneous regarding the character to be studied. Other sampling methods are more convenient when the population is finite and the sampling is without replacement it may be convenient to perform: Stratified Random Sampling, Sampling by Conglomerates and Systematic Sampling are examples of these alternative methods.

Definition 1.1.1. *Let X be a continuous random variable with density function $f(x)$. We say that $(X_1, \ldots, X_n)$ is a simple random sample if for all $(x_1, \ldots, x_n)$ its joint density function g() satisfies that*

$$g(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n).$$

Let X be a discrete random variable with probability-law function $p(x)$. We say that $(X_1, \ldots, X_n)$ is a simple random sample if for all $(x_1, \ldots, x_n)$ its joint probability function g() satisfies that

$$g(x_1, \ldots, x_n) = p(x_1) \ldots p(x_n).$$

Definition 1.1.2. *Let $(X_1, \ldots, X_n)$ be a simple random sample of size n. We define*

- *the average:* $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- *the variance:* $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- *the quasi-variance:* $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- *the sample proportion : assuming that the image of X_i is $\{0, 1\}$,* $p^* = \frac{1}{n} \sum_{i=1}^{n} X_i$

Definition 1.1.3. *Let X be a random variable. We define the k-th moment around zero as $\alpha_k = \mathbb{E}[X^k]$ and the k-th moment around the mean as $\mu_k = \mathbb{E}[(X - \mathbb{E}[X])^k]$.*

Throughout this course, we assume that X is a random variable that defines the character under study of the considered population and we use the following notation: $\mathbb{E}[X] = \mu$ and $Var[X] = \sigma^2$.

Definition 1.1.4. *Let $(X_1, \ldots, X_n)$ be a simple random sample of size n. We define the k-th sample moment around zero as $a_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$ and the k-th sample moment around the mean as $m_k = \frac{1}{n} \sum_{i=1}^{n} (X_i^k - \bar{X})^k$.*

Proposition 1.1.1. *Let X be a random variable with expected value μ and variance σ^2. Then,*

1. $\mathbb{E}[\bar{X}] = \mathbb{E}[X]$

2. $\mathbb{E}[a_k] = \alpha_k$

3. $Var[\bar{X}] = \frac{\sigma^2}{n}$

4. $\mathbb{E}[S^2] = \sigma^2$

Proof. We prove the above properties as follows:

1. $\mathbb{E}[\bar{X}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \mu.$
2. $\mathbb{E}[a_k] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n X_i^k] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^k] = \frac{1}{n}\sum_{i=1}^n \alpha_k = \alpha_k$
3. $Var[\bar{X}] = Var[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n^2}Var[\sum_{i=1}^n X_i] \overset{indep}{=} \frac{1}{n^2}\sum_{i=1}^n Var[X_i] = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$
4. We show that $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$, which implies that the desired result follows.

   $$\mathbb{E}[S_n^2] = \mathbb{E}[a_2 - a_1^2] = \mathbb{E}[a_2] - \mathbb{E}[a_1^2] = \alpha_2 - \mathbb{E}[a_1^2]$$

   We now show that $\mathbb{E}[a_1^2] = \frac{\alpha_2 + (n-1)\alpha_1^2}{n}$.

   $$\begin{aligned}\mathbb{E}[a_1^2] = \mathbb{E}[\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] \\ &= \frac{1}{n^2}\left(\mathbb{E}\left[\sum_{i=1}^n X_i^2\right] + \mathbb{E}\left[\sum_{i\neq j} X_i X_j\right]\right) \overset{indep}{=} \frac{1}{n^2}\left(\mathbb{E}\left[\sum_{i=1}^n X_i^2\right] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^n \alpha_2 + \sum_{i\neq j}\alpha_1^2\right) = \frac{1}{n^2}\left(n\alpha_2 + n(n-1)\alpha_1^2\right) = \frac{\alpha_2 + (n-1)\alpha_1^2}{n}\end{aligned}$$

   Therefore,

   $$\mathbb{E}[S^2] = \alpha_2 - \left(\frac{\alpha_2 + (n-1)\alpha_1^2}{n}\right) = \frac{n-1}{n}(\alpha_2 - \alpha_1^2) = \frac{n-1}{n}\sigma^2.$$

□

1.1.3 Distribution of $\bar{X}$ and S^2

Proposition 1.1.2. *Let $X_1, \ldots, X_n$ be a simple random sample such that $X_i \sim \mathcal{N}(\mu, \sigma)$. Then, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$*

Proof. The random variable resulting from summing normally distributed variables is normally distributed and multiplying a normal random variable by a scalar is also a normal random variable. The values of the first and the second parameters of the resulting normal distribution follow from 1. and 3. of Proposition 1.1.1. □

It is easy to see that the above result can be alternatively stated as follows:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, \sqrt{n}\sigma).$$

The above result requires that X_i is normally distributed for $i = 1, \ldots, n$. Using the Central Limit Theorem, we conclude that this normality assumption is not required when n is large.

Theorem 1.1.1 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ independent and identically distributed random variables with mean μ and variance σ^2. Then, $\bar{X}$ tends to $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ when $n \to \infty$.*

It is easy to see that the above result can be also alternatively stated as $\sum_{i=1}^{n} X_i \to \mathcal{N}(n\mu, \sqrt{n}\sigma)$ when $n \to \infty$ (in practice $n \geq 30$). The Central Limit Theorem has been seen in Probability Calculus and its proof requires tools from complex analysis applied to random variables (which are studied in advanced courses of probability); therefore, the proof is omitted here. We now present other results that have been studied in Probability Calculus and that will be used in this course.

Proposition 1.1.3. *If $X_1, \ldots, X_n$ are independent random variables such that $X_i \sim \mathcal{N}(0, 1)$, then*

$$X_1 + \cdots + X_n \sim \chi^2_n.$$

Proposition 1.1.4. *If $X_1, \ldots, X_n$ are independent random variables such that $X_i \sim \chi^2_{k_i}$, then*

$$X_1^2 + \cdots + X_n^2 \sim \chi^2_k,$$

where $k = \sum_{i=1}^{n} k_i$.

Theorem 1.1.2 (Fisher's Theorem)**.** *Let $X_1, \ldots, X_n$ independent random variables such that $X_i \sim \mathcal{N}(\mu, \sigma)$. Then,*

- *$\bar{X}$ and S^2 are independent*
- *$\frac{(n-1)S^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2}$ and follows the chi square distribution with $n-1$ degrees of freedom.*

From the above result, we conclude that $\mathbb{E}[\frac{nS_n^2}{\sigma^2}] = n - 1$ and also that $Var[\frac{nS_n^2}{\sigma^2}] = 2(n-1)$.

We now provide some results regarding the distributions of a simple random sample.

Proposition 1.1.5. *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \sim \mathcal{N}(\mu, \sigma)$, where σ is unknown. Then,*

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

Proof. By definition of student's t distribution: $t_{n-1} = \frac{Z\sqrt{n-1}}{\sqrt{\chi^2_{n-1}}}$. Since $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$, we have that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z$. Moreover, from Fisher's theorem, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Therefore,

$$\frac{Z\sqrt{n-1}}{\sqrt{\chi^2_{n-1}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sqrt{n-1}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}},$$

which simplifying gives $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ and the desired result follows. □

We know that, when $n \to \infty$, then $t_n \to Z$; in practice we consider that it is true when $n \geq 30$. As a result of this and the above result, we have thus that, when $n \geq 30$, $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

Proposition 1.1.6. *Let $X_{11}, \ldots, X_{1n}$ be a simple random sample of size n from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \ldots, X_{1m}$ be a simple random sample of size m from a population defined by $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, where X_1 and X_2 are independent random variables. Then,*

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F_{n-1,m-1},$$

where S_1^2 (resp. S_2^2) is the quasi-variance of $X_{11}, \ldots, X_{1n}$ (resp. of $X_{21}, \ldots, X_{2m}$).

Proof. By definition of Fisher distribution and, from Theorem 1.1.2, we know that $\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n-1}$ and $\frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2_{m-1}$, therefore

$$F_{n-1,m_1} = \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)} = \frac{\frac{\frac{(n-1)S_1^2}{\sigma_1^2}}{n-1}}{\frac{\frac{(m-1)S_2^2}{\sigma_2^2}}{m-1}},$$

which simplifying gives $\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$ and the desired result follows. □

Remark 1.1.1. *Recall that for the Fischer distribution, we have that $F_{1-\alpha,n,m} = \frac{1}{F_{\alpha,m,n}}$. The proof is here:*

$$\alpha = \mathbb{P}(F_{n,m} \geq F_{\alpha,n,m}) = \mathbb{P}\left(\frac{\frac{\chi^2_n}{n}}{\frac{\chi^2_m}{m}} \geq F_{\alpha,n,m}\right) = \mathbb{P}\left(\frac{\frac{\chi^2_m}{m}}{\frac{\chi^2_n}{n}} \leq \frac{1}{F_{\alpha,n,m}}\right) = \mathbb{P}\left(F_{m,n} \leq \frac{1}{F_{\alpha,n,m}}\right)$$

Therefore, $1-\alpha = \mathbb{P}\left(F_{m,n} \geq \frac{1}{F_{\alpha,n,m}}\right)$ and as a result, the desired result follows.

Proposition 1.1.7. *Let $X_{11}, \ldots, X_{1n}$ be a simple random sample of size n from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \ldots, X_{1m}$ be a simple random sample of*

size m from a population defined by $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$*, where* X_1 *and* X_2 *are independent random variables. Assume that* σ_1 *and* σ_2 *are known. Then,*

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1).$$

Proof. Since $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, then $\bar{X}_1 \sim \mathcal{N}(\mu_1, \frac{\sigma_1}{n_1})$ (see Proposition 1.1.2). Likewise, $\bar{X}_2 \sim \mathcal{N}(\mu_2, \frac{\sigma_2}{n_2})$. Let $Y = \bar{X}_1 - \bar{X}_2$. We know that Y is normally distributed and, to end the proof, we need to show that: (i) $\mathbb{E}[Y] = \mu_1 - \mu_2$ and (ii) $Var[Y] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

We show (i).

$$\mathbb{E}[\bar{X}_1 - \bar{X}_2] = \mathbb{E}[\bar{X}_1] - \mathbb{E}[\bar{X}_2] = \mu_1 - \mu_2.$$

We now show (ii). We use that X_1 and X_2 are independent and, therefore, $\bar{X}_1$ and $\bar{X}_2$ are independent

$$Var[\bar{X}_1 - \bar{X}_2] = Var[\bar{X}_1] + Var[\bar{X}_2] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

□

Proposition 1.1.8. *Let* $X_{11}, \ldots, X_{1n}$ *be a simple random sample of size n from a population defined by* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ *and* $X_{21}, \ldots, X_{1m}$ *be a simple random sample of size m from a population defined by* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$*, where* X_1 *and* X_2 *are independent random variables and* σ_1 *and* σ_2 *are unknown but equal, i.e,* $\sigma_1 = \sigma_2 = \sigma$*. Then,*

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2},$$

where $S_p = \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}$.

Proof. From Proposition 1.1.7, we know that, if σ_1 and σ_2 are known,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1).$$

Since $\frac{(n-1)S_1^2}{\sigma_1^1} \sim \chi^2_{n-1}$ and $\frac{(m-1)S_2^2}{\sigma_2^1} \sim \chi^2_{m-1}$, we have that $\frac{(n-1)S_1^2}{\sigma_1^1} + \frac{(m-1)S_2^2}{\sigma_2^1} \sim \chi^2_{n+m-2}$. From the definition of student's t distribution and because $\sigma_1 = \sigma_2 = \sigma$,

$$t_{n+m-2} \sim \frac{Z\sqrt{n+m-2}}{\sqrt{\chi^2_{n+m-2}}} = \frac{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}\sqrt{n+m-2}}{\sqrt{\frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2}}},$$

which simplifying gives the desired result. □

Remark 1.1.2. *Under the above conditions, we have that* $\mathbb{E}[S_p^2] = \sigma^2$. *To see this,*

$$\begin{aligned}\mathbb{E}[S_p^2] &= \frac{1}{n+m-2}\mathbb{E}[nS_1^2 + mS_2^2] = \frac{1}{n+m-2}\left(n\mathbb{E}[S_1^2] + m\mathbb{E}[S_2^2]\right) \\ &= \frac{1}{n+m-2}((n-1)\sigma + (m-1)\sigma) = \sigma.\end{aligned}$$

Proposition 1.1.9. *Let* $X_{11}, \dots, X_{1n}$ *be a simple random sample of size* n *from a population defined by* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ *and* $X_{21}, \dots, X_{1m}$ *be a simple random sample of size* m *from a population defined by* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, *where* X_1 *and* X_2 *are independent random variables and* σ_1 *and* σ_2 *are unknown and unequal. Then,*

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \sim t_{df},$$

where $df = \dfrac{\left(\frac{S_1^2}{n-1} + \frac{S_2^2}{m-1}\right)^2}{\frac{\left(\frac{S_1^2}{n-1}\right)^2}{n-1} + \frac{\left(\frac{S_2^2}{m-1}\right)^2}{n-1}}$.

1.1.4 Distribution of p^*

The following result is an application of the Central Limit Theorem and Proposition 1.1.2. that uses that, for a Bernoulli distributed random variable with parameter p, the expected value is equal to p and the variance is pq.

Proposition 1.1.10. *Let* $X_1, \dots, X_n$ *be independent random variables such that* $X_i \sim Ber(p) = Bin(1, p)$. *Then, when* n *tends to infinity, we have that*

$$p^* \sim \mathcal{N}(p, \frac{\sqrt{pq}}{\sqrt{n}}).$$

In the following result, we focus on the proportion of two independent populations.

Proposition 1.1.11. *Let* $X_{11}, \dots, X_{1n}$ *be a simple random sample of size* n *from a population defined by* $X_1 \sim Ber(p_1)$ *and* $X_{21}, \dots, X_{1m}$ *be a simple random sample of size* m *from a population defined by* $X_2 \sim Ber(p_2)$, *where* X_1 *and* X_2 *are independent random variables. Then,*

$$p_1^* - p_2^* \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n} + \frac{p_2 q_2}{m}}\right).$$

Proof. The result follows because the subtraction of two normal distributions is normal and $\mathbb{E}[p_1^* - p_2^*] = p_1 - p_2$ as well as $Var[p_1^* - p_2^*] = \frac{p_1 q_1}{n} + \frac{p_2 q_2}{m}$, in which we use that X_1 and X_2 are independent random variables. □

1.2 Estimation

We consider that we have a simple random sample $X_1, \ldots, X_n$ of size n from a population defined by a cumulative density function $F(x, \theta)$, where θ is unknown (it can be a single parameter or a vector of parameters). Our goal is to estimate θ.

Definition 1.2.1. *Let $X_1, \ldots, X_n$ be a simple random sample of size n. An estimator of θ is denoted as θ^* and it is a function of the sample $X_1, \ldots, X_n$, i.e.,*

$$\begin{aligned} \theta^* :&(X_1, \ldots, X_n) \to \mathbb{R} \\ &(x_1, \ldots, x_n) \ \to \theta^*(x_1, \ldots, x_n). \end{aligned}$$

It is important to note that an estimator θ^* is a random variable since it is a function of a random vector. Its distribution depends on the parameter θ.

Example 4. *Let $X \sim \mathcal{N}(\mu, 2)$, where μ is unknown. We take a simple random sample of size n from this random variable and we consider $\bar{X}$ as the estimator of μ. According to Proposition 1.1.2, this estimator follows a normal distribution where the first parameter is μ and the second $2/\sqrt{n}$.*

1.2.1 Methods to Obtain Estimators

Method of Moments

If we want to estimate k parameters using the method of moments, we equalize the first k-th moments about zero of the random variable (see Definition 1.1.3.) with the first k-th sample moments about zero (see Definition 1.1.4.), i.e., $a_i = \alpha_i$, for $i = 1, \ldots, k$. This can be also done with the moments about the mean instead of about the zero.

Example 5. *If $X \sim \mathcal{N}(\mu, 1)$, where μ is unknown. Thus, $a_1 = \bar{X}$ and $\alpha_1 = \mu$. Therefore, $\mu^* = \bar{X}$.*

Example 6. *If $X \sim Poi(\lambda)$, where λ is unknown. Thus, $a_1 = \bar{X}$ and $\alpha_1 = \lambda$. Therefore, $\mu^* = \bar{X}$.*

Maximum Likelihood Estimator

Let $X_1, \ldots, X_n$ be a simple random sample of size n from a population defined by the density function $f(x; \theta)$, where θ is unknown. We define the likelihood function as

$$L(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \ldots f(x_n; \theta).$$

The maximum likelihood estimator is the value of θ maximizing the above expression, i.e.,

$$L(x_1, \ldots, x_n; \theta^*) = \max_{\theta} f(x_1; \theta) \ldots f(x_n; \theta).$$

However, we will be sometimes interested in maximizing the logarithm of the likelihood function (since it becomes easier).

Example 7. *If $X \sim \mathcal{N}(\mu, 1)$, where μ is unknown, then*

$$L(x_1, \dots, x_n; \theta) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(x_1-\mu)^2} \dots \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(x_n-\mu)^2} = \frac{1}{(\sqrt{2\pi})^2} e^{\frac{-1}{2}\sum_{i=1}^n (x_i-\mu)^2}.$$

We take the logarithm and we get

$$-\frac{n}{2} Ln(2\pi) - \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2.$$

We derivate with respect to μ and we equal to zero, which gives

$$\frac{-1}{2}(-1)\sum_{i=1}^n 2(x_i - \mu) = 0 \iff \mu^* = \bar{x}.$$

We can easily show that the second derivative of the logarithm of the likelihood with respect to μ is negative. Therefore, we have shown the maximum likelihood estimator of μ is $\bar{X}$ for this case.

Example 8. *If $X \sim Poi(\lambda)$, where λ is known.*

$$\begin{aligned} L(x_1, \dots, x_n; \lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \dots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! \dots x_n!} \\ LnL(x_1, \dots, x_n; \lambda) &= -n\lambda + \left(\sum_{i=1}^n x_i\right) Ln\lambda - Ln(x_1! x_2! \dots x_n!) \\ \frac{\partial LnL(x_1, \dots, x_n; \lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0 \Rightarrow \lambda = \bar{x}. \end{aligned}$$

We can easily check that it is a maximum since the second derivative of the logarithm of the likelihood with respect to μ is negative. Therefore, we have shown the maximum likelihood estimator of θ is $\bar{X}$ for this case.

When we have more than one parameter to be estimated, we compute the derivative of the likelihood function (or the logarithm of the likelihood function) with respect to each of the variables and we solve the resulting system of equations.

1.2.2 Properties of the Estimators

Definition 1.2.2 (Bias)**.** *An estimator θ^* of θ is unbiased if $\mathbb{E}[\theta^*] = \theta$. If $\mathbb{E}[\theta^*] \neq \theta$, then θ^* is a biased estimator of θ. The bias of a estimator θ^* of θ is defined as $b(\theta^*) = \mathbb{E}[\theta^*] - \theta$.*

From this definition and Proposition 1.1.1, we conclude that the following results are true.

Proposition 1.2.1. *Let $X_1, \ldots, X_n$ be a simple random sample of size n from a population defined by a random variable X with mean $\mathbb{E}[X]$, variance $Var[X]$ and k-th moment about zero $\mathbb{E}[X^k]$.*

- *$\bar{X}$ is an unbiased estimator of μ*
- *a_k are unbiased estimators of $\mathbb{E}[X^k]$*
- *S_n^2 is a biased estimator of $Var[X]$, but S^2 is an unbiased estimator of $Var[X]$*

Since we have seen two methods to obtain estimators, we can get different estimators for the same unknown parameter. If this is the case, we would like to know which one we must select. For this purpose, we will use the notion of Mean Square Error.

Definition 1.2.3 (Mean Square Error)**.** *Let θ^* be an estimator of θ. The Mean Square Error (or risk) of θ^* is*

$$R(\theta^*, \theta) = \mathbb{E}[(\theta^* - \theta)^2].$$

In general, we will say that an estimator is better if its Mean Square Error is smaller.

Proposition 1.2.2. *Let θ^* be an estimator of θ.*

$$R(\theta^*, \theta) = Var[\theta^*] + b(\theta^*)^2.$$

Proof.

$$\begin{aligned} R(\theta^*, \theta) &= \mathbb{E}[(\theta^* - \theta)^2] = \mathbb{E}[\theta^*] + \mathbb{E}[\theta^2] - 2\mathbb{E}[\theta\theta^*] = \mathbb{E}[\theta^*] + \theta^2 - 2\theta\mathbb{E}[\theta^*] \\ &= \mathbb{E}[\theta^*] - \mathbb{E}[\theta^*]^2 + \mathbb{E}[\theta^*]^2 + \theta^2 - 2\theta\mathbb{E}[\theta^*] = Var[\theta^*] + (\mathbb{E}[\theta^*] - \theta)^2 \\ &= Var[\theta^*] + b(\theta^*)^2. \end{aligned}$$

□

Definition 1.2.4 (UMVUE)**.** *An estimator θ_0^* of the parameter θ is UMVUE (Uniformly Mimimum Variance and Unbiased Estimator) if any other unbiased estimator θ^* of θ satisfies that $Var[\theta_0^*] \leq Var[\theta^*]$.*

We now focus on the minimum variance that an estimator θ^* of a $\gamma(\theta)$ (i.e., a function of θ) can achieve. This is called the Cramer-Rao Lower Bound and it is defined as

$$CRLB = \frac{(\gamma'(\theta))^2}{I_n(\theta)},$$

where $I_n(\theta)$ is the Fisher information, which is defined as

$$I_n(\theta) = \mathbb{E}\left[\left(\frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial\theta}\right)^2\right].$$

or, since we have a simple random sample, as

$$I_n(\theta) = n\mathbb{E}\left[\left(\frac{\partial LnL(x;\theta)}{\partial\theta}\right)^2\right].$$

In the particular case where $\gamma(\theta) = \theta$, we have that

$$CRLB = \frac{1}{I_n(\theta)}.$$

We will show that the CRLB is a lower bound of the variance of any unbiased estimator of $\gamma(\theta)$. First, we provide the following auxiliary results.

Lemma 1.2.1. $\mathbb{E}\left[\frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right] = 0.$

Proof. We note that

$$\int\cdots\int L(x_1,\dots,x_n)dx_1\dots dx_n = 1,$$

We derivate with respect to θ both sides of the above expression and, assuming that the derivative and integral interchange, we have that

$$\int\cdots\int \frac{\partial L(x_1,\dots,x_n)}{\partial\theta}dx_1\dots dx_n = 0.$$

Taking into account that

$$\frac{\partial L(x_1,\dots,x_n)}{\partial\theta} = \frac{\partial LnL(x_1,\dots,x_n)}{\partial\theta}L(x_1,\dots,x_n), \tag{1.1}$$

we have that

$$\int\cdots\int \frac{\partial LnL(x_1,\dots,x_n)}{\partial\theta}L(x_1,\dots,x_n)dx_1\dots dx_n = 0.$$

From the definition of the expected value of a function of a random variable, we have that

$$\int\cdots\int \frac{\partial LnL(x_1,\dots,x_n)}{\partial\theta}L(x_1,\dots,x_n)dx_1\dots dx_n = \mathbb{E}\left[\frac{\partial LnL(x_1,\dots,x_n)}{\partial\theta}\right],$$

which implies that the desired result follows. □

Lemma 1.2.2. *If θ^* is an unbiased estimator of $\gamma(\theta)$, then* $\mathbb{E}\left[\theta^*\frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right] = \gamma'(\theta)$.

Proof. Since θ^* is an unbiased estimator of $\gamma(\theta)$

$$\mathbb{E}[\theta^*] = \gamma(\theta).$$

By the definition of expected value, the lhs of the above expression is

$$\mathbb{E}[\theta^*] = \int \cdots \int \theta^* L(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

We derivate with respect to θ both sides and, assuming that the derivative and integral interchange, we have that

$$\int \cdots \int \theta^* \frac{\partial L(x_1, \ldots, x_n)}{\partial \theta} dx_1 \ldots dx_n = \gamma'(\theta).$$

Using (1.1), we have that

$$\int \cdots \int \theta^* \frac{\partial LnL(x_1, \ldots, x_n)}{\partial \theta} L(x_1, \ldots, x_n) dx_1 \ldots dx_n = \gamma'(\theta).$$

From the definition of the expected value of a function of a random variable, we have that

$$\int \cdots \int \theta^* \frac{\partial LnL(x_1, \ldots, x_n)}{\partial \theta} L(x_1, \ldots, x_n) dx_1 \ldots dx_n = \mathbb{E}\left[\theta^* \frac{\partial LnL(x_1, \ldots, x_n)}{\partial \theta}\right],$$

which implies that the desired result follows. □

We now present the result of the CRLB. As in the previous lemmas, we will assume that the derivative and integral interchange.

Theorem 1.2.1. *If θ^* is an unbiased estimator of $\gamma(\theta)$. Then,*

$$Var[\theta^*] \geq CRLB.$$

Proof. From the definition of covariance,

$$Cov\left[\theta^*, \frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial \theta}\right] = \mathbb{E}\left[\theta^* \frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial \theta}\right] - \mathbb{E}\left[\theta^*\right] \mathbb{E}\left[\frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial \theta}\right],$$

which using Lemma 1.2.1. gives

$$Cov\left[\theta^*, \frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial \theta}\right] = \mathbb{E}\left[\theta^* \frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial \theta}\right],$$

and from Lemma 1.2.2.

$$Cov\left[\theta^*, \frac{\partial LnL(x_1, \ldots, x_n; \theta)}{\partial \theta}\right] = \gamma'(\theta). \tag{1.2}$$

On the other hand, from the definition of the linear correlation coefficient, we can

derive the following property:

$$Cov\left[\theta^*, \frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right]^2 \leq Var\left[\theta^*\right] Var\left[\frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right],$$

which using the definition of variance, gives

$$Cov\left[\theta^*, \frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right]^2 \leq Var\left[\theta^*\right] \left(\mathbb{E}\left[\left(\frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right)^2\right] - \mathbb{E}\left[\frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right]^2\right),$$

and using again Lemma 1.2.1,

$$Cov\left[\theta^*, \frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right]^2 \leq Var\left[\theta^*\right] \mathbb{E}\left[\left(\frac{\partial LnL(x_1,\dots,x_n;\theta)}{\partial\theta}\right)^2\right].$$

And the desired result follows from the above expression and (1.2). □

We now provide some definitions.

Definition 1.2.5. *An estimator is efficient if its variance equals CRLB.*

Definition 1.2.6. *An estimator is optimum if it is unbiased and efficient.*

Definition 1.2.7. *Let $\{\theta_n\}_{n\in\mathbb{N}}$ be a sequence of estimators of the parameter θ. This sequence is consistent if it converges in probability to θ, i.e.,*

$$\mathbb{P}(|\theta_n - \theta| > \epsilon) = 0, \forall\epsilon > 0.$$

An alternative manner of analyzing whether an estimator is consistent is given in the following result.

Proposition 1.2.3. *Let $\{\theta_n^*\}_{n\in\mathbb{N}}$ be a sequence of estimators of the parameter θ. This sequence is consistent if, when $n \to \infty$, the following conditions hold:*

- $\mathbb{E}[\theta_n^*]$ *tends to* θ
- $Var[\theta_n^*]$ *tends to* 0.

Proof. To prove consistency we must show that, for every $\varepsilon > 0$,

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \longrightarrow 0.$$

We begin by using the inequality

$$|\hat{\theta}_n - \theta| \leq |\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| + |\mathbb{E}[\hat{\theta}_n] - \theta|.$$

Thus, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \leq \mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|\mathbb{E}[\hat{\theta}_n] - \theta| > \frac{\varepsilon}{2}\right).$$

Since $\mathbb{E}[\hat{\theta}_n] \to \theta$, we have that $\mathbb{P}\left(|\mathbb{E}[\hat{\theta}_n] - \theta| > \frac{\varepsilon}{2}\right) = 0$ when $n \to \infty$.

Now apply Chebyshev's inequality (we are now proving Chebyshev's inequality in this course) to the first term:

$$\mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| > \frac{\varepsilon}{2}\right) \leq \frac{\mathrm{Var}(\hat{\theta}_n)}{(\varepsilon/2)^2} = \frac{4\ \mathrm{Var}(\hat{\theta}_n)}{\varepsilon^2}.$$

Since $\mathrm{Var}(\hat{\theta}_n) \to 0$, this upper bound tends to zero.

Therefore,

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \longrightarrow 0.$$

□

Example 9. *Define $\mu_n^* = \frac{1}{n}\sum_{i=1}^n X_i$. We have that μ_n^* is a consistent sequence of estimators of μ because $\mathbb{E}[\mu_n^*] = \mu$ for all n and $Var[\mu_n^*] = \frac{\sigma^2}{n}$ (see Proposition 1.1.1.), which tends to zero when $n \to \infty$.*

Definition 1.2.8. *Let θ^* be a estimator of the parameter θ. This estimator is sufficient if the following property holds:*

$$f(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) g_\theta(\theta^*).$$

Example 10. *Let X be a Poisson random variable with parameter λ unknown and $X_1, \ldots, X_n$ a simple random sample. Prove that $\lambda^* = \sum_{i=1}^n X_i$ is a sufficient estimator of λ.*

$$p_\lambda(x_1, \ldots, x_n) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \ldots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1! \ldots x_n!} = \frac{1}{x_1! \ldots x_n!} e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} =$$
$$= h(x_1, \ldots, x_n) g_\lambda(\lambda^*(X_1, \ldots, X_n)).$$

We now see the concept of invariance of an estimator.

Definition 1.2.9. *Let $\vec{X} = (X_1, \ldots, X_n)$ be a simple random sample and $\theta^*(\vec{X})$ be an estimator of θ. Then, for all $c, a \in \mathbb{R}, c \neq 0$, $\vec{Y} = (cX_1 + a, \ldots cX_n + a)$. We say that an estimator is invariant to change of scale and change of origin if,*

$$\theta^*(\vec{X}) = \theta^*(\vec{Y}).$$

An estimator can be only invariant to change of scale or invariant to change of origin.

Example 11. *Show that $\bar{X}$ is not invariant to change of origin, but S^2 is invariant to change of origin. Let $\vec{Y} = (X_1 + a, \dots X_n + a)$. When $\theta^* = \bar{X}$,*

$$\theta^*(\vec{Y}) = a + \bar{X} \neq \bar{X} = \theta^*(\vec{X}).$$

When $\theta^* = S^2$,

$$\theta^*(\vec{Y}) = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} = \frac{\sum_{i=1}^n \left((X_i + a) - \left(\bar{X} + a\right)\right)^2}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \theta^*(\vec{X}).$$

The Maximum Likelihood Estimator verifies the following properties:

- it is asymptotically unbiased
- it is consistent
- it is asymptotically efficient
- it is invariant to bijective transformations
- its distribution tends to the normal distribution

The method of moments provides a simpler method to compute an estimator. However, the variance of the moments estimator (i.e., the estimator using the method of the moments) is often larger than that of the Maximum Likelihood Estimator. The reason for this is that the method of moments does not use the full information about the distribution, but only the moments.

1.3 Estimating using intervals

So far, we focused on a single value to estimate the value of an unknown parameter. Now, we will see how intervals can be used to provide estimations.

We are in the same setting as in the previous sections, i.e., we consider that we have a simple random sample $X_1, \dots, X_n$ of size n from a population defined by a cumulative density function $F(x, \theta)$, where θ is unknown.

Definition 1.3.1. *Let $\vec{X} = (X_1, \dots, X_n)$ and $\alpha \in [0, 1]$. We consider $\theta_1^*(\vec{X})$ and $\theta_2^*(\vec{X})$ such that $\theta_1^*(\vec{X}) < \theta_2^*(\vec{X})$, then if*

$$\mathbb{P}(\theta_1^*(\vec{X}) \leq \theta \leq \theta_2^*(\vec{X})) = 1 - \alpha$$

the interval $(\theta_1^*(\vec{X}), \theta_2^*(\vec{X}))$ is called the $(1 - \alpha) \cdot 100\%$ confidence interval of θ. The value $(1 - \alpha) \cdot 100\%$ is known as the confidence level of the interval. We write

$$I_\theta^{1-\alpha} = (\theta_1^*(\vec{X}), \theta_2^*(\vec{X})).$$

The method to obtain confidence intervals consists of using a value called statistical pivot $g(\vec{x}, \theta)$, whose distribution is known.

Before computing the CIs of different settings, it is important to remark that the confidence interval depends on the simple random sample, therefore, it is a random vector. The main feature of this interval is that the probability such that the unknown parameter is between these two values is fixed to $1 - \alpha$. A typical value of α is 0.05.

Example 12. *Let us consider 5 simple random samples of size 10 about some process of interest whose mean is known (i.e., μ is unknown). The value of the mean and quasi-standard deviation of each sample is given the following table.*

	1	2	3	4	5
$\bar{x}$	116,9	132,8	117	106,7	111,9
s	21,7	32,62	22,44	14,13	20,46

For each sample, one can calculate one confidence interval of μ, as it can be seen in Figure 1.1. It is remarkable that one of the CIs does not contain the value of μ.

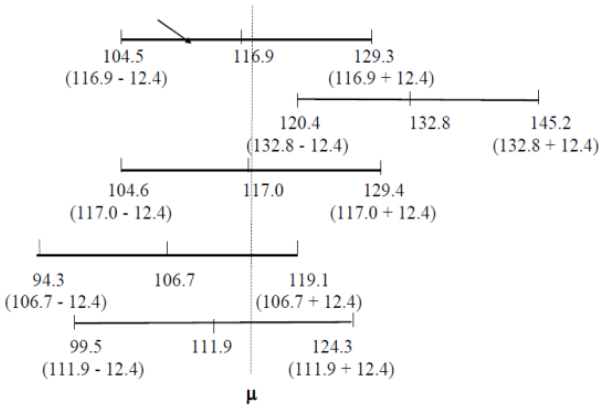


Figure 1.1: Five confidence intervals.

From this example, we conclude that there might be some CIs such that the unknown parameter is not inside. However, the definition of confidence interval establishes that

$$\mathbb{P}(\theta_1^*(\bar{X}) \leq \theta \leq \theta_2^*(\bar{X})) = 1 - \alpha,$$

which means that $(1 - \alpha) \cdot 100\%$ of the confidence intervals contain the unknown parameter.

1.3.1 CI of the mean of a normal distribution with known σ

In this case, the pivot we will use is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, which according to Proposition 1.1.2, it follows the standardized normal distribution.

Proposition 1.3.1. *When $X \sim \mathcal{N}(\mu, \sigma)$ with σ known, the $(1-\alpha)\cdot 100\%$ confidence interval of μ is*

$$I_\mu^{1-\alpha} = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

where s is the sample quasi-standard deviation.

Proof. Using that $Z \sim \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ (see Proposition 1.1.2) and the definition of the Z distribution, we have that

$$\begin{aligned}
&\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha \\
&\mathbb{P}(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha \\
&\mathbb{P}(-\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha \\
&\mathbb{P}(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha.
\end{aligned}$$

□

Using the CLT, the above result extends to big samples without requiring that the population is normal. By big samples we mean that $n \geq 30$, i.e., the size of the simple random sample is larger than 30.

Example 13. *Imagine that the university U wants study the height of the students because they would like to create a basketball team if the mean height is larger than 180 cm. Getting a large simple random sample, we can obtain the 95% confidence interval of the mean height of the students.*

Suppose that the obtained 95% CI is $(181, 192)$, this means that, with a confidence level of 95%, this university will create the basketball team.

1.3.2 CI of the mean of a normal distribution with unknown σ

In this case, the pivot we will use is $\frac{\bar{X}-\mu}{S/\sqrt{n}}$, which according to Proposition 1.1.5, it follows the Student's t distribution.

Proposition 1.3.2. *When $X \sim \mathcal{N}(\mu, \sigma)$ with σ unknown, the $(1-\alpha)\cdot 100\%$ confidence interval of μ is*

$$I_\mu^{1-\alpha} = \left(\bar{X} - t_{\alpha/2;n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{-\alpha/2;n-1}\frac{S}{\sqrt{n}}\right),$$

where S is the quasi-standard deviation of the simple random sample.

Proof. Using that $t_{n-1} \sim \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ (see Proposition 1.1.5) and the definition of the Student's t distribution, we have that

$$\begin{aligned}
&\mathbb{P}(-t_{\alpha/2;n-1} \leq t_{n-1} \leq t_{\alpha/2;n-1}) = 1-\alpha \\
&\mathbb{P}(-t_{\alpha/2;n-1} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2;n-1}) = 1-\alpha \\
&\mathbb{P}(-\bar{X} - t_{\alpha/2;n-1}\frac{S}{\sqrt{n}} \leq -\mu \leq -\bar{X} + t_{\alpha/2;n-1}\frac{S}{\sqrt{n}}) = 1-\alpha \\
&\mathbb{P}(\bar{X} - t_{\alpha/2;n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{-\alpha/2;n-1}\frac{S}{\sqrt{n}}) = 1-\alpha.
\end{aligned}$$

□

Recall that, when n is large, the Z distribution approximates well the Student's t distribution with n degrees of freedom. Therefore, for n large ($n \geq 30$), the confidence interval for this case is

$$I_{\mu}^{1-\alpha} = \left(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{-\alpha/2}\frac{s}{\sqrt{n}}\right).$$

1.3.3 CI of the variance of a normal distribution with unknown μ

In this case, the pivot we will use is $\frac{nS_n^2}{\sigma^2}$, which according to Theorem 1.1.2, it follows the chi square distribution with $n-1$ degrees of freedom.

Proposition 1.3.3. *When $X \sim \mathcal{N}(\mu, \sigma)$ with μ unknown, the $(1-\alpha)\cdot 100\%$ confidence interval of σ^2 is*

$$I_{\sigma^2}^{1-\alpha} = \left(\frac{nS_n^2}{\chi^2_{\alpha/2;n}}, \frac{nS_n^2}{\chi^2_{1-\alpha/2;n}}\right).$$

Proof. From Theorem 1.1.2, we know that $\frac{nS_n^2}{\sigma^2} \sim \chi^2_{n-1}$, therefore

$$\begin{aligned}
&\mathbb{P}(\chi^2_{1-\alpha/2;n-1} \leq \chi^2_{n-1} \leq \chi^2_{\alpha/2;n-1}) = 1-\alpha \\
&\mathbb{P}\left(\chi^2_{1-\alpha/2;n-1} \leq \frac{nS_n^2}{\sigma^2} \leq \chi^2_{\alpha/2;n-1}\right) = 1-\alpha \\
&\mathbb{P}\left(\frac{nS_n^2}{\chi^2_{\alpha/2;n-1}} \leq \sigma^2 \leq \frac{nS_n^2}{\chi^2_{1-\alpha/2;n-1}}\right) = 1-\alpha
\end{aligned}$$

□

1.3.4 CI of the difference of the means of two independent and normal distributions with known variances

Let $X_{11}, \ldots, X_{1n_1}$ be a simple random sample of size n_1 from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \ldots, X_{2n_2}$ be a simple random sample of size n_2 from a

population defined by $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, where X_1 and X_2 are independent random variables and σ_1 and σ_2 are known. In Proposition 1.1.7, we show that, when σ_1 and σ_2 are known, $\bar{X}_1 - \bar{X}_2$ is normally distributed with first parameter $\mu_1 - \mu_2$ and second $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Using the same arguments as before, taking as pivot $\bar{X}_1 - \bar{X}_2$, we can provide an analytical expression of the confidence interval for this case.

Proposition 1.3.4. *Under the above conditions,*

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

1.3.5 CI of the difference of the means of two independent and normal distributions with unknown but equal variances

For this case, we use the result of Proposition 1.1.8 and, taking as pivot $\bar{X}_1 - \bar{X}_2$ as well as the previous arguments to provide an analytical expression of the confidence interval.

Proposition 1.3.5. *Let $X_{11}, \dots, X_{1n_1}$ be a simple random sample of size n_1 from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \dots, X_{2n_2}$ be a simple random sample of size n_2 from a population defined by $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, where X_1 and X_2 are independent random variables and σ_1 and σ_2 are unknown and equal.*

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2;n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2;n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right),$$

where

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}.$$

1.3.6 CI of the difference of the means of two independent and normal distributions with unknown and unequal variances

For this case, we use the result of Proposition 1.1.9 and taking as pivot $\bar{X}_1 - \bar{X}_2$ and the previous arguments to provide an analytical expression of the confidence interval.

Proposition 1.3.6. *Let $X_{11}, \dots, X_{1n_1}$ be a simple random sample of size n_1 from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \dots, X_{2n_2}$ be a simple random sample of size n_2 from a population defined by $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, where X_1 and X_2 are independent random variables and σ_1 and σ_2 are unknown and unequal.*

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2;df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2;df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right),$$

where

$$df = \frac{\left(\frac{S_1^2}{n_1-1} + \frac{S_2^2}{n_2-1}\right)^2}{\frac{\left(\frac{S_1^2}{n_1-1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2-1}\right)^2}{n_1-1}}.$$

1.3.7 CI of the ratio of the variances of two independent and normal distributions

In Proposition 1.1.6, we show that $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$, where n_i is the sample size of population i, $i = 1, 2$. We take thus $\frac{S_1^2/(n_1-1)}{S_2^2/(n_2-1)}$ as pivot to provide an analytical expression of the CI of this case.

Proposition 1.3.7. *Let $X_{11}, \ldots, X_{1n_1}$ be a simple random sample of size n_1 from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \ldots, X_{2n_2}$ be a simple random sample of size n_2 from a population defined by $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, where X_1 and X_2 are independent random variables.*

$$I^{1-\alpha}_{\sigma_1^2/\sigma_2^2} = \left(\frac{\frac{S_1^2}{S_2^2}}{F_{\alpha/2;n_1-1;n_2-1}}, \frac{\frac{S_1^2}{S_2^2}}{F_{1-(\alpha/2);n_1-1;n_2-1}} \right).$$

Proof. We have that

$$\begin{aligned}
&\mathbb{P}\left(F_{1-(\alpha/2);n_1-1;n_2-1} \leq F_{n_1-1;n_2-1} \leq F_{\alpha/2;n_1-1;n_2-1}\right) = 1-\alpha \\
&\mathbb{P}\left(F_{1-(\alpha/2);n_1-1;n_2-1} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq F_{\alpha/2;n_1-1;n_2-1}\right) = 1-\alpha \\
&\mathbb{P}\left(\frac{\frac{S_1^2}{S_2^2}}{F_{\alpha/2;n_1-1;n_2-1}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\frac{S_1^2}{S_2^2}}{F_{1-(\alpha/2);n_1-1;n_2-1}}\right) = 1-\alpha.
\end{aligned}$$

□

1.3.8 CI of the difference of the means of two dependent and normal distributions

Let X_1 and X_2 two normal random variables that describe a characteristic of two populations, i.e., $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$. We define the following random variable

$$D = X - Y \sim \mathcal{N}(\mu_D, \sigma_D),$$

where $\mu_D = \mu_1 - \mu_2$ and σ_D cannot be computed since X and Y are not independent. Therefore, we are in the same conditions as in Section 1.3.2 (i.e., we have a single population where the variance is unknown). The pivot to be used is $\frac{\bar{D}-\mu_D}{S_D/\sqrt{n}}$, where S_D is the quasi-standard deviation of D, which follows the Student's t distribution

with $n-1$ degrees of freedom (recall that n is the size of the simple random sample, which implicitly is required to get a simple random sample of the same size in both populations)·

Proposition 1.3.8. *Let $X_{11},\dots,X_{1n}$ be a simple random sample of size n from a population defined by $X_1 \sim \mathcal{N}(\mu_1,\sigma_1)$ and $X_{21},\dots,X_{2n}$ be a simple random sample of size n from a population defined by $X_2 \sim \mathcal{N}(\mu_2,\sigma_2)$, where X_1 and X_2 are dependent random variables. Let $D = X - Y$. Then*

$$I_{\mu_D}^{1-\alpha} = \left(\bar{D} - t_{\alpha/2;n-1}\sqrt{\frac{S_D^2}{n}}, \bar{D} + t_{\alpha/2;n-1}\sqrt{\frac{S_D^2}{n}}\right).$$

1.3.9 CI of the proportion of a Bernoulli distributions

We now consider a population whose characteristic under study follows a Bernoulli distribution with unknown parameter p.

Example 14. *Suppose we have a box with red and white balls. The number of balls is huge. We aim to know the proportion of red balls. This situation follows a Bernoulli distribution with unknown p.*

We use the result of Proposition 1.1.5. However, in this case, instead of n tending to infinity, we consider that $np^* > 5$ and $nq^* > 5$, where n is the size of the simple random sample, p^* the proportion of successes of the simple random sample and $q^* = 1 - p^*$.

Therefore, the pivot we take is $\frac{p^*-p}{\sqrt{pq/n}}$, which from Proposition 1.1.5 follows the Z distribution. Using the same arguments as before, one can show the following result.

Proposition 1.3.9. *Let $X_1,\dots,X_n$ a simple random sample that follows a distribution $X \sim Ber(p)$. The $(1-\alpha)\cdot 100\%$ confidence interval of p is*

$$I_p^{1-\alpha} = \left(p^* - z_{\alpha/2;n-1}\frac{\sqrt{p^*q^*}}{\sqrt{n}}, p^* + z_{\alpha/2;n-1}\frac{\sqrt{p^*q^*}}{\sqrt{n}}\right).$$

1.3.10 CI of the difference of the proportion of two independent and Bernoulli distributions

We now consider two independent and Bernoulli distributed populations. We focus on the difference between the proportion of successes of the first population and the proportion of successes of the second population. For this case, the pivot we use is $\frac{p_1-p_2-(p_1^*-p_2^*)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}}$, which according to Proposition 1.1.11, it follows the Z distribution. The CI for this case can be computed using the same arguments as in the previous instances.

Proposition 1.3.10. *Let $X_{11},\dots,X_{1n_1}$ be a simple random sample of size n_1 from a population defined by $X_1 \sim Ber(p_1)$ and $X_{21},\dots,X_{2n_2}$ be a simple random sample of*

size n_2 from a population defined by $X_2 \sim Ber(p_2)$, where X_1 and X_2 are dependent random variables. Then

$$I_{p_1-p_2}^{1-\alpha} = \left(p_1^* - p_2^* - z_{\alpha/2}\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}, p_1^* - p_2^* + z_{\alpha/2}\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}} \right).$$

1.3.11 Properties of the CIs

We consider, for instance, the CI of Section 1.3.1. and we define the length of the CI as the difference between the extremes of the CIs, i.e., $L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. If we fix L and α, the size of the single random sample is given by

$$n = 4z_{\alpha/2}^2 \frac{\sigma^2}{n}.$$

On the other hand, if we consider the CI of Section 1.3.9., the length is $L = 2z_{\alpha/2}\frac{p^* q^*}{\sqrt{n}}$. Therefore, if we want to get the size of the sample so as to ensure that the length of the CI is fixed to L, we need to set

$$n = 4z_{\alpha/2}^2 \frac{p^* q^*}{n}.$$

From these reasoning, we get the following conclusions:

- the CI decreases with the size of the sample
- the CI increases with the confidence level $(1-\alpha)\cdot 100\%$

1.4 Fundamental exercises

Exercise 1.4.1. *(a) A company manufactures metal rods, and the length of each rod (in cm) follows a distribution that is defined below, where a is an unknown parameter. Estimate the maximum possible length a using the method of moments.*

$$f(x) = \begin{cases} \frac{2}{a^2}(a-x), & 0 < x < a, \\ f(x) = 0, & otherwise. \end{cases}$$

(b) From reliability engineering, we know that the lifetime (in hours) of a certain machine component follows a Pareto distribution with parameter a (the density function of this distribution is given below), where $a > 1$ is unknown. Estimate a using the method of moments.

$$f(x) = \begin{cases} \frac{a x_0^a}{x^{a+1}}, & x > x_0, , \\ f(x) = 0, & otherwise. \end{cases}$$

Exercise 1.4.2. *A pharmaceutical company measures the concentration of an active ingredient in tablets. Assume the concentrations follow a normal distribution $N(\mu, \sigma^2)$.*

Using a random sample of tablets, compute the maximum likelihood estimators for the mean concentration μ and variance σ^2.

Exercise 1.4.3. *A factory produces screws whose lengths are uniformly distributed between a and b. From a sample of screw lengths, estimate the minimum and maximum lengths using the maximum likelihood method.*

Exercise 1.4.4. *(a) A call center records the number of calls until the first successful sale. Assuming the number of calls follows a geometric distribution, compute the maximum likelihood estimator of the success probability θ. The probability law of the geometric distribution is given by,*

$$f(x;\theta) = \theta(1-\theta)^{x-1}, x = 1, 2, 3, \dots.$$

(b) The time (in minutes) to repair a machine follows a Gamma distribution with shape parameter 2 and scale parameter θ (its density function is provided below). Using data from a sample of observations of repair data, compute the maximum likelihood estimator of θ.

$$f(x;\theta) = \frac{1}{\theta^2} x e^{-x/\theta}.$$

Exercise 1.4.5. *The lifetime (in years) of a certain type of battery follows an exponential distribution with rate θ (its density function is provided below). Using a ranom sample of observed lifetimes, estimate θ using:*

1. *The method of moments.*
2. *The method of maximum likelihood.*

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

Exercise 1.4.6. *In a quality control process, each product is either defective (1) or not defective (0). Show that the sample proportion of defective products*

$$p^* = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of the true defect rate p. Compute its mean square error and check if it is consistent.

Exercise 1.4.7. *A researcher wants to estimate the average daily sales of a store using an estimator of the form $a\bar{X}$, where $\bar{X}$ is the sample mean of observed daily sales. Compute the value of a that minimizes the mean square error of the estimator.*

Exercise 1.4.8. *In question of a survey, each respondent either yes (1) or no (0). We aim to estimate p, which is the proportion of European citizens whose answer to that*

question is affirmative. We consider a sample of size n and the following estimator:

$$T = \sum_{i=1}^{n} X_i$$

Show that T is a sufficient statistic of p. For $n = 3$, analyze if the estimator $T = x_1 + 2x_2 + x_3$ is sufficient.

Exercise 1.4.9. *A company measures the maximum load (in kg) that a new material can withstand. The load is uniformly distributed between 0 and θ (where $\theta > 0$). Using sample data, estimate θ by:*

- *The method of moments. Check if it is unbiased.*
- *The method of maximum likelihood. Check if it is unbiased.*
- *Compare the variances of both estimators. Which one has the smaller variance?*

Exercise 1.4.10. *A financial analyst estimates the average return of a stock using an estimator*

$$A = k \sum_{i=1}^{n} iX_i,$$

where X_i are observed returns in a random sample. Which is the value of k so that the estimator is unbiased? For this value, compute:

1. *The efficiency of this estimator.*
2. *Analyze if it is a consistent estimator.*

Help: $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.

Exercise 1.4.11. *Ecologists use the capture-recapture method to estimate the size of animal populations. Initially, T animals are captured and tagged. Later, a second sample of size C is taken, and R tagged animals are found. Using $R/C \approx T/N$, estimate N, the total population size. For the following data, estimate the size of the population under study.*

- *A study on ravens: $T = 30$, $C = 20$, $R = 2$.*
- *A study on swans: $T = 50$, $C = 30$, $R = 5$.*
- *A study on drug addicts in New York: $T = 500$, $C = 1800$, $R = 20$.*

Exercise 1.4.12. *We study the viscosity of the blood of two types of rats: with adrenaline (B) and without adrenaline (A). Samples:*

- *A: 3.29, 3.91, 4.64, 3.55, 3.67, 4.18, 3.74, 4.67, 3.03, 4.61, 3.84*
- *B: 3.35, 4.26, 4.71, 3.14, 3.45, 5.01, 4.43, 4.91, 4.22, 4.83, 3.55*

Assuming normality and with a 95% *confidence level, compute the confidence interval for the difference of means. Provide conclusions.*

Exercise 1.4.13. *Climate scientists compare average temperatures recorded by 1,000 weather stations in 2000 and 2022. The mean obtained from the data from 2000 was* $57°F$*, and from 2020 it was* $57.6°F$*. Moreover, the standard deviation of difference of the data of the samples is* $s_D = 4.1°F$*. Analyze with* 95% *confidence whether the temperature in 1988 is higher than in 1950.*

Exercise 1.4.14. *A public health study surveys 1,000 teenagers under 16 and finds that 200 are regular smokers.*

1. *Compute the* 99% *confidence interval for this proportion.*
2. *Would you be surprised if another study reports the proportion as* 0.23*? Explain.*

Exercise 1.4.15. *The drug* Anturana *is tested to prevent death in a second heart attack. In a study, 733 patients took Anturana and 742 took placebo. After 8 months, 42 patients died: 29 of them took placebo and 13 of them Anturana.*

1. *Compute a* 95% *confidence interval for this difference.*
2. *If we compute a* 90% *confidence interval, would it be larger or smaller? Why?*
3. *Is there statistical evidence that Anturana reduces death rate? Explain.*
4. *Compute the minimum required sample size* ($n = n_1 = n_2$) *so that the* 95% *confidence interval has length smaller than* 0.04*. And for* 90% *confidence?*

1.5 Further exercises

Exercise 1.5.1. *A factory monitors the production of metal rods. Each rod length (in meters) is uniformly distributed between 0 and 1. From this process, we take* $k = 30$ *samples, each consisting of* 48 *rods. For each sample, we compute:*

$$Y = \frac{1}{2}(X_1 + \cdots + X_{48}) - 12.$$

1. *Analyze the distribution of* Y*.*
2. *Using statistical software, simulate these samples and compare the obtained mean and variance with the theoretical ones.*

Exercise 1.5.2. *The lifetime (in hours) of a certain electronic component follows an exponential distribution shifted by* μ*, with density:*

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

Using a sample of observed lifetimes, estimate μ and θ using the method of moments.

Exercise 1.5.3. *The duration (in minutes) of a mechanical device follows a distribution with density:*

$$f(x) = \frac{x}{a^2} e^{-x^2/(a^2)}, \quad x > 0.$$

1. *Estimate a using the method of moments.*

2. *Estimate a using the method of maximum likelihood.*

Exercise 1.5.4. *Prove that the sample mean $\bar{X}$ is a consistent estimator of the population mean μ. Interpret this result in the context of estimating the average daily energy consumption of households.*

Exercise 1.5.5. *Two laboratories measure the variability of a chemical process. The population variance σ^2 is unknown. The first lab takes a sample of size 6 and obtains $s_1^2 = 9$. The second lab takes a sample of size 8 and obtains $s_2^2 = 12$. Check that the estimator:*

$$\hat{\sigma}^2 = \frac{6s_1^2 + 8s_2^2}{12}$$

is unbiased for σ^2. Compute its estimated value.

Exercise 1.5.6. *The time (in minutes) to complete a certain industrial task follows a distribution with density:*

$$f(x;\theta) = e^{-(x-\theta)}, \quad x \geq \theta.$$

Find the maximum likelihood estimator of θ using a sample of observed completion times.

Exercise 1.5.7. *Let $X_1, \ldots, X_n$ be a simple random sample with finite variance. Suppose $a_{nk} \geq 0$ and $\sum_{k=1}^{n} a_{nk} = 1$. Consider the estimator:*

$$\theta_n = \sum_{k=1}^{n} a_{nk} X_k.$$

1. *Show that θ_n is an unbiased estimator of the population mean.*

2. *Show that $\{\theta_n\}$ is a consistent sequence of estimators.*

Exercise 1.5.8. *Suppose we take a simple random sample $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2. We consider two estimators of μ:*

$$A = \bar{X}, \qquad B = \frac{\sum_{i=1}^{n} iX_i}{\sum_{i=1}^{n} i}.$$

Are they unbiased estimators? Which one has the smaller variance? Interpret the result in the context of giving more weight to recent observations in a time series.

Exercise 1.5.9. *The tail length of a certain insect species follows a normal distribution with mean* 60 *mm and standard deviation* 1.5 *mm. A random sample of* 9 *insects is taken.*

1. *What is the distribution of the sample mean* $\bar{X}$*? Find the interval* (a, b) *such that* $P(a \leq \bar{X} \leq b) = 0.95$.
2. *Define* $B^2 = \sum_{i=1}^{9}(x_i - \bar{x})^2$.
   (a) *What is the distribution of* $\frac{B^2}{\sigma^2}$*?*
   (b) *Compute its expected value and variance.*
   (c) *Find the interval* $(0, a)$ *that contains* B^2 *with probability* 0.95.
3. *The observed sample values are: 60.4, 60.8, 59.0, 61.6, 62.6, 58.3, 59.2, 60.2, 60.5 mm.*
   (a) *Compute the sample mean and variance.*
   (b) *Are these values inside the previously computed intervals?*

Exercise 1.5.10. *Nine people with the same physical disability perform a task. The average time is* 7 *minutes with a standard deviation of* 2 *minutes. In general, the time follows a normal distribution with mean* 5 *minutes.*

1. *If a sample of size* 9 *is considered, what is the probability that the average time exceeds* 7 *minutes?*
2. *Find* A *such that* $P(5 < \bar{X} < A) = 0.95$.

Exercise 1.5.11. *The number of clients of a supermarket chain follows a normal distribution with mean* 5000 *and standard deviation* 500. *A sample of* 25 *supermarkets is taken.*

1. *Compute the probability that the sample mean is less than* 5075.
2. *Find the interval that contains the sample mean with probability* 0.95.

Exercise 1.5.12. *The time between pollination and fertilization of a plant follows a normal distribution with mean* 6 *months and standard deviation* 2 *months. A sample of* 25 *plants is taken. Compute the expectation, variance, and mean square error of the sample mean.*

Exercise 1.5.13. *Consider a Bernoulli random variable* $X \sim Bin(1, p)$ *and a sample of size* $n = 20$. *Compute the variance and standard deviation of* $\hat{p}$ *and find their maximum values. Using normal approximation:*

1. *With probability* 0.68*, what is the maximum distance between* $\hat{p}$ *and* p*? And with probability* 0.95*?*

2. *Is $n = 20$ enough so that with probability 0.95 the maximum distance between $\hat{p}$ and p is 0.05?*

Exercise 1.5.14. *Two random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are taken from different populations with unknown means μ_1, μ_2 and known variances σ_1^2, σ_2^2. Compute the sample size n so that $P(|(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)| > \epsilon) = \alpha$.*

Exercise 1.5.15. *To study the influence of sex on weight loss during exercise, 10 men and 8 women participate in a program. The percentage of weight lost is:*

- *Women: 3.0, 2.5, 3.7, 3.3, 3.8, 4.1, 3.6, 4.0*
- *Men: 2.9, 3.5, 3.9, 3.8, 3.6, 3.7, 3.8, 4.0, 3.6, 3.7*

1. *Compute the point estimate of the difference in weight loss between women and men.*
2. *Assuming normality, compute the 95% confidence interval for this difference. Can we conclude there is a significant difference? Explain.*

Exercise 1.5.16. *Two methods of heart transplantation are compared. For treatment A: $n_1 = 9$, $\bar{x}_1 = 16$ days, $S_1' = 10.1$ days. For treatment B: $n_2 = 9$, $\bar{x}_2 = 15$ days, $S_2' = 10$ days.*

1. *Compute the 95% confidence interval for the ratio of variances. What can be concluded?*
2. *Compute the 95% confidence interval for the difference in mean lifetime. Interpret the result.*

Exercise 1.5.17. *A study measures the effect of using a glucose measurement device at home. For 36 diabetics, the difference in glucose level (before minus after) has mean 2.78 mmol/l and standard deviation 6.05 mmol/l.*

1. *Is there evidence that using the device reduces glucose level? Justify.*
2. *What conditions are required for this analysis?*

Exercise 1.5.18. *Two types of families are considered: city and countryside. Let p_1 and p_2 be the proportions of families with more than two children. Samples: $n_1 = 200$, $X_1 = 152$; $n_2 = 100$, $X_2 = 62$.*

1. *Check that $\hat{p}(a) = a\frac{X_1}{n_1} + (1-a)\frac{X_2}{n_2}$ is unbiased for p when $p_1 = p_2 = p$. Find a that maximizes variance.*
2. *Estimate p_1 and p_2 with 95% confidence.*
3. *Can we conclude there is a significant difference between p_1 and p_2? Use a confidence interval.*

Exercise 1.5.19. *In 1950,* 1000 *weather stations recorded a mean temperature of* $57°F$. *In 1988, the same stations recorded* $57.6°F$. *The standard deviation of differences is* $s_d = 4.1°F$. *Test if the 1988 temperature is significantly higher at* 95% *confidence.*

Exercise 1.5.20. *A sample of* 200 *white blood cells shows* 125 *neutrophils.*

1. *Estimate the proportion of neutrophils.*
2. *Compute the* 90% *confidence interval for this proportion.*
3. *Healthy humans have* 0.6–0.7 *neutrophils. Is there evidence of imbalance?*

Exercise 1.5.21. *In 1987, a drought killed many seedlings. A sample of* 1000 *seedlings shows* 300 *dead. Compute the* 90% *confidence interval for the proportion of dead seedlings. Estimate the proportion killed by drought.*

Exercise 1.5.22. *A study of stress-related deaths finds that among* 17 *cases,* 11 *had heart problems.*

1. *Estimate the proportion of deaths due to heart problems.*
2. *Compute the* 95% *confidence interval for this proportion. Is the sample size sufficient?*
3. *Find the sample size needed so that the* 95% *confidence interval has length* 0.02.

2 Hypothesis Testing

2.1 Introduction

Hypothesis testing is one of the most important techniques that is used in statistics. It is formed by two factors:

- An hypothesis about a character of a population
- The error that can be achieved in this technique

As we are providing information about a characteristic of a population using the information obtained from a sample of that population, an error is unavoidable.

In this chapter, we will focus on parametric hypothesis testing methods, in which the hypothesis is about a parameter of the population under study (the mean, the variance or proportion, for instance).

2.2 Preliminaries

Let X a random variable with cumulative distribution function $F(x, \theta)$, where θ is an unknown parameter (or a vector of unknown parameters). We assume that X describes the characteristic of our population that we aim to study.

We denote by Ω the set of values that θ can get and we call it the parametric space. The hypothesis testing method consists of dividing Ω in two disjoint subsets. This division leads to the definition of null hypothesis (H_0) and alternative hypothesis (H_1).

- Null hypothesis (H_0): it is the set of values that are considered as true, unless the data clearly show its falsehood.

- Alternative hypothesis (H_1): it is the set of values that are not in the null hypothesis. Therefore, if we can say that the null hypothesis is false, we conclude that this set gives the set of values that the unknown parameter takes.

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1 = \bar{\Theta}_0 = \Omega - \Theta_0$$

When Θ_0 is a single point, we say that the hypothesis testing is simple, otherwise it is composed. Simple hypothesis tests are also called bilaterals. For instance,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

There are two types of composed tests: left-unilateral for instance

$$H_0 : \theta \geq \theta_0$$
$$H_1 : \theta < \theta_0$$

and right-unilateral, for instance

$$H_0 : \theta \leq \theta_0$$
$$H_1 : \theta > \theta_0.$$

In this chapter, we will see statistical techniques that determine if an hypothesis is true or not. From the definition of null and alternative hypothesis, we conclude that four different errors can be done when we carry out these tests:

	Reject H_0	Reject H_1
H_0 is true	Type-1 error	Correct
H_1 is true	Correct	Type-2 error

The probability of a type-1 error is defined as follows

$$\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true})$$

whereas the probability of a type-2 error

$$\mathbb{P}(\text{do not reject } H_0 | H_1 \text{ is true}).$$

The probability of a type-1 error is denoted by α and of a type-2 error by β. The power of a test is defined as the complementary of the probability of a type-2 error, i.e.,

$$1 - \beta = \mathbb{P}(\text{reject } H_0 | H_1 \text{ is true}).$$

Solving a hypothesis test means to conclude if the null hypothesis is rejected or not. It is important that this conclusion is done in terms of probabilities; therefore, we must always state which is the value of the significance level, i.e., the value of the probability of type-1 error, which is denoted by α. We will see two methods to solve hypothesis tests: the critical region method and the p-value. In both cases, we will assume that H_0 is true and we select a statistical pivot which, under this assumption, its distribution will be known.

2.3 Critical region method

In this method, we will take a pivot and we define a rule to reject or not H_0. This is, if assuming that H_0 is true, the pivot satisfies a given condition for the obtained simple random sample, then the null hypothesis is rejected. Otherwise, it cannot be rejected. The condition to be satisfied by the pivot will be called as critical region. The critical region is denoted by S_1. The complementary of the critical region is the acceptance region and it is denoted by S_0. Thus, if $x_1, \ldots, x_n$ is a simple random sample, then

- if $(x_1, \ldots, x_n)$ satisfies the condition of the critical region, then H_0 is rejected
- if $(x_1, \ldots, x_n)$ satisfies the condition of the acceptance region, then H_0 is not rejected

From this property, we can define again α, which is the significance level, and the power of a test $1-\beta$ as follows

$$\alpha = \mathbb{P}((x_1, \ldots, x_n) \text{ satisfies the condition of } S_1 | H_0 \text{ is true})$$

$$1-\beta = \mathbb{P}((x_1, \ldots, x_n) \text{ satisfies the condition of } S_1 | H_1 \text{ is true})$$

Example 15. *We consider a normal population with variance 1 and unknown mean. We aim to solve the following test:*

$$H_0 : \mu = 4$$
$$H_1 : \mu \neq 4.$$

We know from Proposition 1.1.2, that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Z$. *Assuming that* H_0 *is true and taking into account that* $\sigma = 1$, *we have that* $\frac{\bar{X}-4}{1/\sqrt{n}} \sim Z$.

We get a simple random sample of size n *and consider the following critical region:*

$$S_1 = \{(x_1, \ldots, x_n) | \frac{\bar{x}-4}{1/\sqrt{n}} \notin (-1, 1)\}.$$

We take a simple random sample of size $n = 9$ *and such that* $\bar{x} = 3.9$. *Thus,* $\frac{\bar{x}-4}{1/\sqrt{n}} = 0.3$, *which does not satisfy the condition of* S_1. *Therefore, from this test, we cannot reject that the population mean is 4.*

A statistical test is thus nothing but a definition of a condition. In the above example, we had that $\frac{\bar{x}-4}{1/\sqrt{n}} \notin (-1,1)$, but we could consider any other way of defining the critical region, i.e., any other way of determining when the null hypothesis can be rejected or not. There are infinite possible critical regions. Among all of them, we will be interested in the test that maximizes the power.

Definition 2.3.1. *A statistical test such that its critical region is S_1^* is said to be of maximum power if for any other statistical test with critical region S_1,*

$$\mathbb{P}((x_1,\dots,x_n) \text{ satisfies the condition of } S_1^* | H_1 \text{ is true}) > \mathbb{P}((x_1,\dots,x_n) \text{ satisfies the condition of } S_1 | H_1 \text{ is true}).$$

In the following, we will write

$$(x_1,\dots,x_n) \text{ satisfies the condition of } S_1$$

as $\mathbb{P}((x_1,\dots,x_n) \in S_1$.

Our goal is to find the statistical test that maximizes the power. For this purpose, we will use the Theorem of Neyman-Pearson.

Theorem 2.3.1 (Theorem of Neyman-Pearson)**.** *Let $\vec{x} = (x_1,\dots,x_n)$ be a simple random sample of a population defined by $f(x,\theta)$, where θ is unknown. Let us consider the following hypothesis testing:*

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

Suppose there exists a value of $k \in \mathbb{R}$ such that the critical region S_1^* verifies the following conditions:

- $f(x,\theta_1)/f(x,\theta_0) > k$, $\forall \vec{x} \in S_1^*$
- $f(x,\theta_1)/f(x,\theta_0) \leq k$, $\forall \vec{x} \notin S_1^*$
- $\alpha = \mathbb{P}((x_1,\dots,x_n) \in S_1^* | H_0 \text{ is true})$

Then, the statistical test that defines S_1 as the critical region is the test that maximizes the power among all the tests with smaller of equal significance level (i.e., probability of type-1 error).

Proof. We consider a statistical test with significance level α and power $1-\beta$, where its critical region is S_1^*. From the definition of α we have that

$$\alpha = \mathbb{P}(\vec{x} \in S_1^* | H_0 \text{ is true}) = \int_{S_1^*} f(x,\theta_0)dx,$$

and of $1-\beta$,

$$1-\beta = \mathbb{P}(\vec{x} \in S_1^* | H_1 \text{ is true}) = \int_{S_1^*} f(x, \theta_1) dx.$$

We consider another statistical test with significance level α_1 and power $1-\beta_1$ where its critical region is S_1. Hence, We show that, if $\alpha_1 \leq \alpha$, then $1-\beta_1 \leq 1-\beta$.

$$\alpha_1 = \mathbb{P}(\vec{x} \in S_1 | H_0 \text{ is true}) = \int_{S_1} f(x, \theta_0) dx,$$

$$1-\beta_1 = \mathbb{P}(\vec{x} \in S_1 | H_1 \text{ is true}) = \int_{S_1} f(x, \theta_1) dx.$$

Since $\alpha_1 \leq \alpha$, we have that

$$\int_{S_1} f(x, \theta_0) dx \leq \int_{S_1^*} f(x, \theta_0) dx.$$

We define A, B and I such that $S_1^* = I \cup A$ and $S_1 = I \cup B$. That is, I is the area that belong to both critical regions. Therefore

$$\int_{S_1} f(x, \theta_0) dx \leq \int_{S_1^*} f(x, \theta_0) dx \iff \int_{I \cup B} f(x, \theta_0) dx \leq \int_{I \cup A} f(x, \theta_0) dx \iff$$

$$\int_B f(x, \theta_0) dx \leq \int_A f(x, \theta_0) dx.$$

Since $A \subset S_1^*$, we have that $f(x, \theta_1)/f(x, \theta_0) > k$, $\forall \vec{x} \in A$. Therefore,

$$\int_A f(x, \theta_1) dx > k \int_A f(x, \theta_0) dx,$$

whereas since $B \notin S_1^*$ we have that $f(x, \theta_1)/f(x, \theta_0) \leq k$, $\forall \vec{x} \in B$. As a result,

$$\int_B f(x, \theta_1) dx \leq k \int_B f(x, \theta_0) dx.$$

Thus, combining these inequalities, we have that

$$\int_A f(x, \theta_1) dx > k \int_A f(x, \theta_0) dx \geq k \int_B f(x, \theta_0) dx \geq \int_B f(x, \theta_1) dx,$$

which means that

$$\int_A f(x, \theta_1) dx > \int_B f(x, \theta_1) dx.$$

Adding I on both sides, we have that

$$\int_{A \cup I} f(x, \theta_1) dx > \int_{B \cup I} f(x, \theta_1) dx.$$

And the desired result follows since the lhs of the above expression is $1-\beta$ and the rhs $1-\beta_1$. □

The above result states that there exists a value of k that determines the critical region of the statistical test that maximizes the power among all the tests with smaller or equal α. This value of k will be obtained using the definition of significance level. Therefore, the significance level fully determines the critical region that is obtained using the Neyman Pearson Theorem. As a result, using the obtained critical region S_1^*, we can conclude if the null hypothesi can be rejected or not. More precisely, with a significance level of α,

- if $(x_1, \dots, x_n) \in S_1^*$, then H_0 is rejected
- if $(x_1, \dots, x_n) \notin S_1^*$, then H_0 is not rejected

2.3.1 Extension to unilateral hypothesis testing

The Neyman Pearson Theorem can be applied to hypothesis testing instances where $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. However, we can extend it to

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 < \theta_0, \end{aligned}$$

which gives the left-unilateral hypothesis testing (note that we have assumed that $\theta_1 < \theta_0$) where the condition of H_0 $\theta \geq \theta_0$ is replaced by $\theta = \theta_0$. Similarly, the extension to the right-unilateral can be done with

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 > \theta_0. \end{aligned}$$

Regarding the extension to unilateral, it can be done by assuming that $\theta_1 < \theta_0$ or $\theta_1 > \theta_0$. It is important to note that this extension can only be carried out in cases where the monotone likelihood ratio property is satisfied. All the instances of this course will satisfy this property.

2.3.2 Likelihood ratio test

A more general method to solve an hypothesis testing consists of the likelihood ratio test. It works for an arbitrary number of unknown parameters and/or when the most powerful test does not exist. This method also provides a critical region,

Let X be a random variable with density $f(x; \theta_1, \dots, \theta_m)$, where $\theta_1, \dots, \theta_m$ are

unknown. We consider

$$H_0 : (\theta_1, \ldots, \theta_m) \in W$$
$$H_1 : (\theta_1, \ldots, \theta_m) \notin W.$$

The likelihood ratio is given by $L(x, \theta_1, \ldots, \theta_m)$. We define

$$L(W) = \max_{(\theta_1,\ldots,\theta_m)\in W} L(x, \theta_1, \ldots, \theta_m)$$

and $L(\Omega) = \max_{(\theta_1,\ldots,\theta_m)\in\Omega} L(x, \theta_1, \ldots, \theta_m)$.

The likelihood ratio is given by $\lambda = L(W)/L(\Omega)$. We have that $\lambda \in (0, 1]$. The critical region is

$$S_1 = \{(x_1, \ldots, x_n) | \lambda < c\},$$

where c can be computed using that

$$\alpha = \sup_{(\theta_1,\ldots,\theta_m)\in W} \mathbb{P}(\lambda < c).$$

Remark 2.3.1. *We consider a normal population with known σ. For this case, we apply Neymann Pearson to the hypothesis test*

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

and we obtain that the most powerful critical region is

$$S_1 = \{(x_1, \ldots, x_n) | \bar{x} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\} \cup \{(x_1, \ldots, x_n) | \bar{x} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}.$$

We apply the likelihood test ratio to this hypothesis test now and we observe that we get the same result as Neymann Pearson.

First, since $W = \{\mu_0\}$, we have that

$$L(W) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2},$$

whereas $L(\Omega)$ is the maximum of the likelihood function, which is achieved when $\mu = \bar{x}$, i.e., at the maximum likelihood estimator. Thus,

$$L(\Omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

From $\lambda = \frac{L(W)}{L(\Omega)}$, and, after some simplification, we get

$$\lambda = e^{\frac{-n}{2\sigma^2}(\bar{x}-\mu_0)^2}.$$

Now,

$$\lambda < c \iff \frac{-n}{2\sigma^2}(\bar{x}-\mu_0)^2 < Ln(c) \iff \frac{n}{\sigma^2}(\bar{x}-\mu_0)^2 > k = 2Ln(c).$$

Therefore, the critical region is

$$S_1 = \{(x_1,\dots,x_n)|\frac{n}{\sigma^2}(\bar{x}-\mu_0)^2 > k\}.$$

When H_0 is true, we have that $X \sim \mathcal{N}(\mu_0,\sigma)$, therefore $\bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$. Thus,

$$\alpha = \mathbb{P}(\frac{n}{\sigma^2}(\bar{x}-\mu_0)^2 > k|\mu=\mu_0) = 1-\mathbb{P}(\frac{n}{\sigma^2}(\bar{x}-\mu_0)^2 < k|\mu=\mu_0) = 1-\mathbb{P}(Z^2 < k|\mu=\mu_0) = 1-\mathbb{P}(-k' < Z <$$

As a consequence,

$$1-\alpha = \mathbb{P}(-k' < Z < k'),$$

which means that $k' = z_{\alpha/2}$. And, the critical region is

$$S_1 = \{(x_1,\dots,x_n)|\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > z^2_{\alpha/2}\},$$

which coincides with the critical region obtained using Neymann Pearson.

2.4 Method of p-value

We focus on the method of the p-value. Using this method, we compute the value of the p-value and we compare it with the value of the significance level (which is α) as follows:

- if the p-value is smaller than α, then we reject H_0 with a significance level of α
- otherwise, we cannot reject H_0 with a significance level of α.

But, what is the p-value? Intuitively, it is the probability that the the pivot we select, which is a function of the values obtained in the simple random sample, follows the distribution that should follow under the assumption that H_0 is true. If this probability is small, we conclude that the assumption that H_0 is true does not hold. A more precise definition of a p-value is presented now.

Definition 2.4.1. *The p-value is the probability that the H_0 is rejected, i.e., that $(x_1,\dots,x_n) \in S_1$, assuming that H_0 is true, that is,*

$$\mathbb{P}((x_1,\dots,x_n) \in S_1|H_0 \text{ is true}).$$

The value of the p-value depends using two factors: the hypothesis testing under consideration and the statistical pivot. Moreover, the statistical pivot depends on the

sample data and the conditions of the hypothesis testing (i.e., which are the unknown parameters and other information of the population distribution). In the following table, we present the information required to compute the p-values (as well as the confidence intervals) of all the instances of the Sections 1.3.1-1.3.10.

The calculation of the p-value is different for different types of the hypothesis testing we have seen in Section 2.2 (bilateral, left-unilateral and right-unilateral). In fact, in any bilateral case, when the pivot is T and its distribution is D, we have that

$$p - value = 2\mathbb{P}(D \geq |T|).$$

For the left-unilateral case, when pivot is T and its distribution is D, we have

$$p - value = \mathbb{P}(D < T).$$

Finally, for the right-unilateral case, when pivot is T and its distribution is D, we have

$$p - value = \mathbb{P}(D > T).$$

For this reason, the bilateral case is also called two-tailed, whereas the unilateral one-tailed. Note that the bilateral case is

$$\mathbb{P}(D < -|T| \cup D > |T|) = 2\mathbb{P}(D \geq |T|).$$

2.5 Hypothesis testing vs confidence intervals

We remark, at this point, that all the instances we have studied in this section can be analyzed using Confidence Intervals. The reason for this is that we are learning parametric hypothesis testing. In the next chapter, we will see another hypothesis testing which considers an arbitrary number of independent population (which cannot be solved using CIs), whereas in the last one non-parametric hypothesis testing, i.e., we focus on other things than parameters (which cannot be solved using CIs either).

It is important that, for the analysis carried out in the hypothesis testings of this section, we must obtain the same conclusions as for CIs when we use the same value of α and the same sample data. For instance, in the case of Section 1.3.1., if the CI of μ obtained with confidence-level of 95% is (5.1,6.3) for a simple random sample, the conclusions of the hypothesis testing

$$H_0 : \mu = 4$$
$$H_1 : \mu \neq 4$$

using the same sample data must be that, with a significance level of 5%, the null hypothesis is rejected.

2.6 Fundamental exercises

Exercise 2.6.1. *In a normal population with a mean μ and a variance of 81, a simple random sample of size 16 is taken. Construct, with a significance level of 5%, the most powerful critical region for a null hypothesis where the population mean is 80, against the alternative hypothesis where the population mean is 90. Perform the test if the sample mean is 87.*

Exercise 2.6.2. *A and B are two people who play heads or tails with a coin. After playing 100 times, it has been observed that the number of heads was 64. Therefore, B believes that the coin is biased. However, A believes that the coin is not biased.*

1. *Formulate the null and alternative hypotheses to test whether the coin is biased.*
2. *Build the most powerful critical region.*
3. *With a significance level of 5%, can we conclude that A is incorrect?*
4. *Calculate the power of the test when the null hypothesis is such that the probability getting head is 2/3.*

Exercise 2.6.3. *We want to compare the lead concentration in two houses. The pipes in one house (house 1) are made of 50% lead and 50% tin, while in the other house (house 2) they are made entirely of tin. Water samples are taken from the kitchen and bathroom of both houses, yielding the following data:*

- *House 1:* $n_1 = 25$, $\bar{x}_1 = 350$ *ppb*, $s_1 = 8$ *ppb*
- *House 2:* $n_2 = 13$, $\bar{x}_2 = 275$ *ppb*, $s_2 = 4$ *ppb*

We suspect that the mean lead concentration in house 1 is higher than in house 2. Suppose the lead concentration follows a normal distribution. Solve the problem using the p-value ($\alpha = 0.05$).

Exercise 2.6.4. *To investigate knee injuries in soccer players, we want to compare two types of shoes. Among the 200 players who used the new type of shoes, 10 players suffered knee injuries. Among the 2050 players who used the traditional shoes, there were 160 injuries. Can we conclude, at a significance level of 5%, that the new type of shoes reduces the number of knee injuries? What about at significance levels of 1% and 10%? Solve the problem using the p-value.*

Exercise 2.6.5. *We want to analyze electric valves purchased from different factories. Let X be the random variable representing the lifespan of a valve. We know that X follows a normal distribution with mean and standard deviation depending on the factory. Specifically:*

In the first factory: $X_1 \approx N(\mu_1, \sigma_1)$ *with* $\sigma_1 = 20$

In the second factory: $X_2 \approx N(\mu_2, \sigma_2)$ *with* $\sigma_2 = 24$

In the first factory, a random sample of size 100 is taken, and the average valve lifespan in the sample is 1252. In the second factory, a sample of size 150 is taken, and the average lifespan is 1247. What can we say, with a significance level of 1%*, about the equality of the mean lifespan of valves in the different factories? Answer using the p-value.*

Exercise 2.6.6. *In 2010, 1000 space stations measured the temperature of the water and land on Earth, obtaining an average temperature of 14 degrees Celsius. In 2022, the same space stations took the same measurements, and the average temperature was 15 degrees Celsius. It is also known that the difference in measurements between the two years has a standard deviation of* $s_D = 0.5$ *degrees Celsius. Can we conclude, based on this data, that the temperature in 2022 is higher than in 2010? Answer using the p-value.*

2.7 Further exercises

Exercise 2.7.1. *Feed producers claim that their product increases milk production in cows. So far, each farmer has obtained an average milk production per cow of 2800 liters, with a standard deviation of 100 liters. Before purchasing the feed, a farmer feeds 64 cows throughout the year and obtains a milk production per cow of 2900 liters.*

1. *To test if the feed increases milk production, what hypotheses should be tested?*
2. *Construct the most powerful critical region for a significance level of* $\alpha = 0.05$*.*
3. *Can we conclude that the feed increases the average milk production?*

Exercise 2.7.2. *We are interested in the growth of a type of leaf, which is a variable that follows a normal distribution with a standard deviation of 4 mm. Construct, with a significance level of* 5%*, the most powerful critical region for a null hypothesis where the mean growth is 35 mm, against the alternative hypothesis where the mean growth is 40 mm. Calculate the sample size and the acceptance region so that the probability of type I error is 0.05 and the type II error is 0.01.*

Exercise 2.7.3. *Let's assume a normal population with a standard deviation of 10. We want to choose between the following hypotheses:* $\mu = 110$ *and* $\mu = 120$*. To do this, we decide to use a statistical test with an acceptance region of* $\bar{x} > K$*. Calculate* n *and* K *so that the significance level is 0.01 and the power of the test is 0.95.*

Exercise 2.7.4. *Let* X *be the random variable that measures the amount of rainfall in a country. Suppose* X *follows a normal distribution with a mean of 20 and a variance of 4. It is suspected that, due to climate change, the average amount of rainfall per year has decreased. Measurements of the amount of rainfall per year are taken, resulting in the following values:*

$$14.1 \quad 23.7 \quad 17.4 \quad 21.1 \quad 20.9 \quad 25.2 \quad 18.4 \quad 22.1 \quad 16.2 \quad 21.2$$

1. *Formulate the necessary hypotheses and construct the most powerful critical region.*

2. *With a significance level of* 1%*, what would be your response?*

3. *Obtain* β *as a function of* μ_1*. What would be the value of* μ_1 *if* $\beta = 0.05$*?*

Exercise 2.7.5. *The concentration of carbon dioxide in the air follows a normal distribution with a mean of* 0.035%*. It is suspected that the concentration is higher at the Earth's surface. Consider* $\alpha = 0.05$*.*

1. *Formulate the null and alternative hypotheses for the required test to test this suspicion.*

2. *Construct the most powerful critical region.*

3. *Samples of 144 air measurements have been randomly taken, with a sample mean of* 0.09% *and a quasi-standard deviation of* 0.25%*. What is the p-value of the test?*

4. *Do you believe the suspicion is true? What significance level did you use? Justify your response using the p-value and the most powerful critical region.*

Exercise 2.7.6. *The fat content in different types of cheese is to be tested. On one hand, it is known that the fat content for a certain type of cheese follows a normal distribution with a mean of* $\mu_0 = 25\%$*. However, for the remaining types of cheese, it is suspected that their mean is* $\mu > \mu_0$ *(for all types of cheese, the standard deviation of fat content is* 6%*).*

1. *Formulate the hypothesis test and calculate the most powerful critical region, knowing that the significance level is* 0.01*.*

2. *Calculate the probability of type II error.*

3. *If a simple random sample of size* $n = 50$ *is taken, and the observed mean fat content in the sample is* 30%*, what conclusions can you draw? Calculate the p-value.*

Exercise 2.7.7. *We want to control* p*, which represents the proportion of defective light bulbs in an industrial production. In the process control, we can choose between the following hypotheses:* $H_0 : p = p_0 = 0.1$ *and* $H_1 : p < p_0$*. A simple random sample of size* $n = 100$ *is taken.*

1. *Calculate the most powerful critical region for a significance level of 0.05.*

2. *If 8 defective light bulbs were observed in this sample, what conclusions can you draw?*

3. *Calculate the power of the test as a function of* p*.*

Exercise 2.7.8. *We have a box with red and white balls. The proportion of white balls is p, and the proportion of blue balls is $1-p$. We perform the following test:*

$$H_0 : p = \frac{1}{2}, \qquad H_1 : p \neq \frac{1}{2}.$$

We randomly draw n balls from the box and apply the following criterion: we will reject H_0 if the number of white balls is 0 or n.

1. *What is the minimum number of balls to be drawn so that the probability of type I error is 0.1 or lower?*

2. *If $n = 10$ balls were drawn, calculate the power of the test $1-\beta(p)$. Calculate $1-\beta(0)$, $1-\beta(1)$, $1-\beta(1/2)$, $1-\beta(3/4)$, $1-\beta(1/4)$, $1-\beta(1/8)$, and $1-\beta(7/8)$. Plot the power function graphically.*

Exercise 2.7.9. *We consider a random variable $N(\mu, 1)$. A random sample of size 25 is taken to estimate μ. Specifically, to test the null hypothesis $\mu = 1$ against the alternative hypothesis $\mu \neq 1$, a test is considered with a critical region of $|\bar{x} - 1| > k$. Calculate the power of this test so that the probability of type I error is 0.05.*

Exercise 2.7.10. *We want to test a large batch of devices. Let's assume that X is the random variable that measures the characteristic of interest in these devices, and it follows a normal distribution with a mean μ and a standard deviation $\sigma = 2$. We are uncertain between the following values of μ: 20 and 22. We randomly select 16 devices.*

1. *Formulate the hypothesis test under consideration in which the null hypothesis is that μ is 20.*

2. *For a significance level of 0.01, what is the most powerful critical region for this test?*

3. *In the obtained sample, the mean is observed to be equal to 21. Which hypothesis would you choose? What is the p-value?*

4. *What is the power of the test?*

Exercise 2.7.11. *We consider a variable $N(\mu, 4)$ and a hypothesis test $H_0 : \mu = 10$ and $H_1 : \mu = 15$. A simple random sample of size 20 is taken, and the following criterion is adopted:*

$$\bar{x} < 13 \Rightarrow \textit{Do not reject } H_0, \qquad \bar{x} \geq 13 \Rightarrow \textit{Reject } H_0.$$

Calculate the significance level and the power of the test. Do you think the criterion followed is appropriate? Justify your response.

Exercise 2.7.12. *We study the reception of parts from an industrial process. Let's assume that p is the proportion of defective parts. The following test is conducted:*

$$H_0 : p = 0.03, \qquad H_1 : p = 0.06.$$

If H_0 is true, the merchandise is accepted. However, if H_1 is true, the merchandise is discarded.

A random sample of 200 parts is taken, where h is the proportion of defective parts in the sample, and the following criterion is used:

$$h < 0.05 \Rightarrow \textit{Accept the merchandise}, \qquad h \geq 0.05 \Rightarrow \textit{Discard the merchandise}.$$

Calculate the probability of type I error, the probability of type II error, and the power of the test defined by this criterion. Do you think the proposed criterion is appropriate? Justify your response.

Exercise 2.7.13. *We have two types of visually indistinguishable threads. The resistance (in grams) is $\mu_0 = 120$ for one type of thread and $\mu_1 = 130$ for the other. Let X be the random variable that measures the resistance of the threads, following a normal distribution with a standard deviation of $\sigma = 5$. We select several threads, all of the same type, but without knowing which type they belong to. We want to determine the type of thread selected using the following hypothesis test:*

$$H_0 : \mu = 120, \qquad H_1 : \mu = 130.$$

1. *Construct the most powerful critical region.*
2. *The resistance of the 10 selected threads is analyzed, and the sample mean is observed to be 125 grams.*
   (a) *What is the critical region for a type I error probability of 0.01?*
   (b) *What conclusions can you draw?*
   (c) *What is the value of β?*
3. *What should be the sample size so that the probability of type I and type II errors is 0.01?*

Exercise 2.7.14. *The time to complete a route follows a standard distribution with a mean of 10 hours and a standard deviation of 1 hour. This route is completed 10 times.*

1. *Construct the most powerful critical region, with a significance level of 1%, for the test in which the null hypothesis is that the mean is 10 hours and the alternative hypothesis is 8 hours.*
2. *After completing the route 50 times, the average time was 8.8 hours. What conclusions can you draw? What is the p-value?*
3. *Calculate the power of the test conducted in the part (a).*

Exercise 2.7.15. *Let $X_1, \ldots, X_n$ be a simple random sample from a Binomial distribution $\text{Bin}(1, p)$, with $n \geq 30$. Suppose we want to perform the following test:*

$$H_0 : p = 0.48, \qquad H_1 : p = 0.52.$$

1. *Calculate the necessary sample size for the test, with the critical region defined as*

$$S_1 = \left\{ \vec{x} : \sum_{i=1}^{n} x_i > k \right\},$$

with type I and type II error probabilities of 0.01.

2. *Suppose a simple random sample of size 500 is taken, and the sample proportion is 0.51. With a significance level of* 1%*, what conclusions can you draw?*

3. *Calculate the power of the test from the previous part.*

Exercise 2.7.16. *We formulate a hypothesis test where the null hypothesis states that the percentage of males in a species is* 25%*, against the alternative hypothesis that the percentage of males is* 30%*. We select a simple random sample of size 100, and if there are more than 30 males in that sample, the null hypothesis is rejected. Find the probability of type I and type II errors.*

Exercise 2.7.17. *Let X be a random variable with the following density function ($\theta > 0$):*

$$f(x) = \begin{cases} \frac{1}{2\theta^3} x^2 e^{-x/\theta}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Suppose we want to test the null hypothesis $\theta = 4$ against the alternative hypothesis $\theta = 2$, with a sample size of 5.

1. *Find the distribution of $\frac{2X}{\theta}$.*

2. *Calculate the uniformly most powerful critical region when the significance level is 0.1.*

3. *What is the power of the test?*

Exercise 2.7.18. *Two machines produce the same item, and a simple random sample is taken from each machine to determine if the proportion of defective items is the same in both machines. From the first machine, 300 units are sampled, and 20 are found to be defective, while from the second machine, 500 units are sampled, and 40 are found to be defective. Is this result consistent with the hypothesis that the proportion of defective items is the same in both machines? Answer using the p-value and the critical region.*

Exercise 2.7.19. *We want to compare the heating consumption of country A and country B. It is assumed that the heating consumption in each country follows a normal distribution. We take a sample from houses in each country, resulting in the following data:*

- *Country A: $n_1 = 150$, $\bar{x}_1 = 600$, and $s_1^2 = 134.23$*

- *Country B:* $n_2 = 75$, $\bar{x}_2 = 500$, *and* $s_2^2 = 216.22$

Can we say that the difference in consumption between the two countries is significant? Use $\alpha = 0.05$.

Exercise 2.7.20. *We want to compare the sales of two products. To do this, we randomly select 30 stores and define* X *as the random variable representing the number of sales for one product, and* Y *as the random variable representing the number of sales for the other product. We assume normality for these random variables. We have the following values:*

$$x_1, x_2, \ldots, x_{30}, \qquad y_1, y_2, \ldots, y_{30}.$$

Based on these values, the following statistics are obtained:

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = 120, \qquad \sum_{i=1}^{30}(x_i - \bar{x})^2 = 28000,$$

$$\bar{y} = \frac{1}{30}\sum_{i=1}^{30} y_i = 130, \qquad \sum_{i=1}^{30}(y_i - \bar{y})^2 = 25750.$$

Is the equality of variances in the number of sales for both products accepted? If so, estimate the common variance and calculate its 99% *confidence interval. Is the difference in means of the number of sales between the two products significant? Answer using the p-value and the critical region.*

Exercise 2.7.21. *An insurance company wants to analyze the proportion of deaths in the twenty years following the purchase of a life insurance policy with them. There are two types of insurance policies: fixed premium and variable premium, depending on the client's health status (worse health results in a higher premium). In the fixed premium policy, 676 insured individuals are randomly selected, and there were 38 deaths. In the variable premium policy, 240 insured individuals are selected, and there were 26 deaths. The company wants to know if the percentage of deaths is the same for both types of policies. What would you answer to that question? Answer using the p-value.*

3

Analysis of Variance

3.1 Introduction

In previous chapters, we have seen hypothesis testings to compare the mean of two independent populations using the t-test. In practical applications it is often necessary to compare the mean of more than two populations. This is, indeed, the goal of this chapter.

Definition 3.1.1. *We define a factor as the characteristic that differentiates each population.*

For instance, when we consider population of different countries, each of the countries is a factor. ´

In this chapter, we will assume that there are $k \geq 2$ independent population, which are normal with the same variance. This means that $\forall k$, $X_k \sim \mathcal{N}(\mu_k, \sigma)$. We will get a simple random sample of each population (that of population k will be denoted as n_k) and we are interested in solving the following hypothesis testing:

$$H_0 : \mu_1 = \cdots = \mu_k$$
$$H_1 : \exists i \neq j \text{ such that } \mu_i \neq \mu_j.$$

This means that we are testing whether all the population means are equal or not. We will see how to solve this hypothesis testing. As in the previous chapter, we will be able to conclude if H_0 can be rejected, in which case we can conclude that there is a difference on the population means, or if H_0 cannot be rejected. To solve this hypothesis testing, we will use the ANOVA table. We will first present some important concepts.

We consider $k \geq 2$ independent populations. We denote by $X_{i,j}$ the observation i of the population j, with $i = 1, \ldots, n_j$ and $j = 1, \ldots, k$. Let $n = \sum_{j=1}^{k} n_j$. We have that $X_{i,j} \sim \mathcal{N}(\mu_i, \sigma)$ and

$$\bar{X} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{i,j}}{n},$$

as well as $\bar{X}_j = \frac{\sum_{i=1}^{n_j} X_{i,j}}{n_j}$, for $j = 1, \ldots, k$.

We formulate a model where $X_{i,j} = \mu_j + e_{i,j}$, where $e_{i,j}$ is the error or the residual. We know that $e_{i,j} \sim \mathcal{N}(0, \sigma)$.

3.2 The ANOVA table

Definition 3.2.1. *We define:*

- *Sum of Squares of Errors:* $SS_E = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{i,j} - \bar{X}_j)^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} e_{i,j}^2$
- *Sum of Squares Between factors:* $SS_B = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2$
- *Total Sum of Squares:* $SS_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{i,j} - \bar{X})^2$

We note that

$$SS_E = \sum_{j=1}^{k} (n_j - 1) S_j^2 = \sum_{j=1}^{k} n_j S_{j,n_j}^2,$$

where S_j^2 is the quasi-variance of population j and S_{j,n_j}^2 the variance of population j.

Theorem 3.2.1 (Fundamental Equation of the Analysis of Variance)**.**

$$SS_T = SS_E + SS_B$$

Proof.

$$\begin{aligned}
SS_T &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{i,j}-\bar{X})^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{i,j}-\bar{X}_i+\bar{X}_i-\bar{X})^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{i,j}-\bar{X}_i)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_i-\bar{X})^2+2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{i,j}-\bar{X}_i)(\bar{X}_i-\bar{X}) \\
&= SS_B+SS_E+2\sum_{j=1}^{k}(\bar{X}_i-\bar{X})\sum_{i=1}^{n_j}(X_{i,j}-\bar{X}_i) \\
&= SS_B+SS_E+2\sum_{j=1}^{k}(\bar{X}_i-\bar{X})\left(\sum_{i=1}^{n_j}X_{i,j}-n_j\bar{X}_i\right) \\
&= SS_B+SS_E.
\end{aligned}$$

□

Proposition 3.2.1. *When H_0 is true, $SS_E/\sigma^2 \sim \chi^2_{n-k}$, $SS_B/\sigma^2 \sim \chi^2_{k-1}$ and $SS_T/\sigma^2 \sim \chi^2_{n-1}$.*

Proof. We show each result separately.

- We focus on SS_E. Since $SS_E = \sum_{j=1}^{k}(n_j-1)S_j^2$ and $\frac{(n_j-1)S_j^2}{\sigma} \sim \chi^2_{n_j-1}$. Finally, the sum of k random variables such that $\chi^2_{n_j-1}$ is χ^2_{n-k}. Therefore, $\sum_{j=1}^{k}\frac{(n_j-1)S_j^2}{\sigma} \sim \chi^2_{n-k}$.

  Remark 3.2.1. *The above result is true even if H_0 is not true.*

- We now focus on SS_B.

$$SS_B = \sum_{j=1}^{k} n_j(\bar{X}_j-\bar{X})^2 = \sum_{j=1}^{k} n_j(\bar{X}_j-\mu)^2 - n(\bar{X}-\mu)^2$$

  We divide by σ^2:

$$\frac{SS_B}{\sigma^2} = \sum_{j=1}^{k}\left(\frac{\bar{X}_i-\mu}{\sigma/\sqrt{n_j}}\right)^2 - \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

  If H_0 is true, the first term is the sum of k Z^2 distributions (therefore, it follows the χ^2_k distribution) and the second the Z^2 distribution, ie., the χ^2_1 distribution. As a result, the subtraction of them follows the χ^2_{k-1} distribution.

- Finally, we use the result of Theorem 3.2.1 to conclude that $\chi^2_{n-k} + \chi^2_{k-1} \sim \chi^2_{n-1}$.

□

Definition 3.2.2. *We define:*

- *Mean Squares of Errors:* $MS_E = \frac{SS_E}{n-k}$
- *Mean Squares Between factors:* $MS_B = \frac{SS_B}{k-1}$
- *Total Mean Squares:* $MS_T = \frac{SS_T}{n-1}$

Proposition 3.2.2. *We have that* $MS_E \sim \chi^2_{n-k}$, $MS_B \sim \chi^2_{k-1}$ *and* $MS_T \sim \chi^2_{n-1}$ *are unbiased estimators of* σ^2 *when* H_0 *is true.*

Proof. The result follows using Proposition 3.2.1 and, as a consequence, $\mathbb{E}[SS_B] = \sigma^2(k-1)$, $\mathbb{E}[SS_E] = \sigma^2(n-k)$ and $\mathbb{E}[SS_T] = \sigma^2(n-1)$. Therefore, $\mathbb{E}[MS_B] = \sigma^2$, $\mathbb{E}[MS_E] = \sigma^2$ and $\mathbb{E}[MS_T] = \sigma^2$ □

We now present the ANOVA table.

	Sum of Squares	Degrees of Freedom	Mean Squares
Between Groups	SS_B	k-1	MS_B
Errors	SS_E	n-k	MS_E
Total	SS_T	n-1	MS_T

Table 3.1: The ANOVA table.

The statistical pivot of this test is $\frac{MS_B}{MS_E}$, which follows the Fischer distribution with $k-1$ and $n-k$ degrees of freedom when H_0 is true since, in that case, $\frac{MS_B}{MS_E}$ is a ratio where the numerator follows the χ^2_{k-1} distribution and the denominator the χ^2_{n-k} distribution.

The idea of the test is to reject H_0 if SS_B is very large compared with SS_E. Note that SS_B measures the sample mean differences between groups. More precisely, if we denote by F_p the value of $\frac{MS_B}{MS_E}$ obtained from the sample, we will reject H_0 when F_p is larger than $F_{\alpha;k-1;n-k}$, that is, the critical region is

$$S_1 = \{(x_{1,1}, \dots, x_{1,n_1}, x_{2,1}, \dots, x_{k,n_k}) | F_p > F_{\alpha;k-1;n-k}\}.$$

Using the same reasoning, the p-value is computed as $\mathbb{P}(F_{k-1,n-k} > F_p)$ and, as in the previous chapter, we will reject H_0 if the p-value does not exceed the significance level.

Remark 3.2.2. *Recall that, when we do the Analysis of Variance test, we are assuming some properties of the data, i.e., normal distribution and equal variance. We need to check that these assumptions holds.*

3.3 Multiple comparisons

When the Analysis of Variance test conclusion consists of rejecting H_0, we know that there are differences on the means of the population. But, which is different? Or, are they all different? We now address this kind of questions with the multiple comparison methods.

A possibility is to use the t-test we have studied in the previous chapter to analysis the difference of by pairs. However, since this approach needs to do $\binom{k}{2}$ tests, it can be very tedious when the number of population is large. Another disadvantage of this approach consists of the so-called family-wise error rate. This means that, when we compute a confidence interval in a pair, we get conclusions with a confidence-level of 0.95, whereas where when we consider, for instance, 15 tests, the total confidence-level is $0.95^{15} = 0.54$, which is very low.

There are several methods that compare the mean of each pair of populations very easily and such that the problem of the family-wise error rate is not achieved. Some examples are Tukey, Scheffe and Bonferroni. In all of them, the output is the p-value of the t-test of each pair and the confidence interval of 95% confidence level.

3.4 Fundamental exercises

Exercise 3.4.1. *Let's assume that we measure the number of pieces produced by four machines on four different days. The obtained data is as follows:*

	Machine 1	*Machine 2*	*Machine 3*	*Machine 4*
Day 1	103	109	104	128
Day 2	115	106	98	117
Day 3	101	116	117	121
Day 4	105	124	99	130

(a) Construct the ANOVA table (Hint: $SS_B = 957.2$ and $SS_E = 647.8$). *At a significance level of* 5%*, can we conclude that there is a significant difference in the average number of pieces produced per day by each machine? What about at a significance level of* 1%*?*

(b) In the following table, we observe the output of conducting multiple comparisons. What conclusions can you draw from these results? How would you rank the machines from lowest to highest average production?

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: `aov(formula = production ~ machine, data = mydata)`

Linear Hypotheses (simultaneous tests):

Contrast	*Estimate*	*Std. Error*	*t value*	*Pr(|t|)*
$2-1=0$	7.750	5.195	1.492	0.4716
$3-1=0$	−1.500	5.195	−0.289	0.9912
$4-1=0$	18.000	5.195	3.465	0.0211*
$3-2=0$	−9.250	5.195	−1.781	0.3284
$4-2=0$	10.250	5.195	1.973	0.2509
$4-3=0$	19.500	5.195	3.754	0.0128*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
Adjusted p-values reported — single-step method.

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: `aov(formula = production ~ machine, data = mydata)`

Quantile = 2.9669
95% family-wise confidence level

Simultaneous confidence intervals for contrasts:

Contrast	*Estimate*	*lwr*	*upr*
$2-1=0$	7.7500	−7.6637	23.1637
$3-1=0$	−1.5000	−16.9137	13.9137
$4-1=0$	18.0000	2.5863	33.4137
$3-2=0$	−9.2500	−24.6637	6.1637
$4-2=0$	10.2500	−5.1637	25.6637
$4-3=0$	19.5000	4.0863	34.9137

Exercise 3.4.2. *A smoker is worried about the mean nicotine level of the three tobacco brands (A, B and C) she smokes. She takes a sample of brand and measure the nicotine level of each cigarette, getting the following data:*

A : 28; 20; 26; 30; 32; 34; 25;

B : 29; 31; 28; 30; 24; 26; 25; 25; 28; 28;

C : 32; 30; 28; 30; 25; 27; 30; 26; 24; 22;

If we analyze the data in R, we obtain the following partial results:

Shapiro–Wilk normality test

data: brandA
$W = 0.98219$, *p-value* $= 0.8993$.

Shapiro–Wilk normality test

data: brandB
$W = 0.97221$, *p-value* $= 0.9021$.

Shapiro–Wilk normality test

data: brandB
$W = 1.614$, *p-value* $= 0.7085$.

Levene's Test for Homogeneity of Variance (center = mean)

Source	*Df*	*F value*	$\Pr(>F)$
group	2	2.4886	0.1034
residuals	25		

(a) From these results, what can we say about the necessary assumptions to perform the Analysis of Variance test?

(b) Taking into account that $SS_B = 0.129$ *and* $SS_E = 292.3$*, construct the ANOVA table, solve the problem with* $\alpha = 0.05$ *and get conclusions.*

Exercise 3.4.3. *We want to study whether there is a significant difference in the effect of three analgesics. For this purpose, we take a random sample of people with headaches and each of them has been given a different type of analgesic. Look at the time (in hours) when the analgesic takes effect. The data obtained are the following:*

$$A: 2.8;\ 3.4;\ 3.1;\ 4.2;\ 2.9;$$
$$B: 3.6;\ 4.1;\ 3.9;\ 4.3;$$
$$C: 2.8;\ 1.9;\ 2.7;\ 3.1;$$

(a) Assuming normality of the time in which the analgesic takes effect and that the variances are equal, solve the hypothesis test with a significance level of 5%*, the equality of the mean time to take effect of the three types of analgesics.* (Hint: $SS_B = 3.646$ and $SS_E = 2.323$).

(b) Using the R output below, what conclusions would you draw? What is the best pain reliever? Reason the answer.

Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

95% family-wise confidence level

Simultaneous confidence intervals for contrasts:

Contrast	*Estimate*	*lwr*	*upr*
$B - A = 0$	0.6950	−0.1902	1.5802
$C - A = 0$	−0.6550	−1.5402	0.2302
$C - B = 0$	−1.3500	−2.2831	−0.4169

Exercise 3.4.4. *We want to analyze the amount of amino acids (in mmol/gr) in four different species of arthropods. The data obtained are the following:*

Specie 1: 431.1; 440.2; 443.2; 445.5; 448.6; 451.2; 442.3;

Specie 2: 477.1; 479; 481.3; 487.8; 489.6; 474; 489.3;

Specie 3: 385.5; 387.9; 389.6; 391.4; 399.1; 403.7; 394.5;

Specie 4: 366.8; 369.8; 371.4; 373.2; 377.2; 379.4; 381.3;

(a) Assuming normality and equality of variances of the quantity of amino acids of the four species of arthropods, test the hypothesis that the average amount of amino acids is the same for the four species. Consider $\alpha = 0.05$ *and use that* $SS_B = 50653.07$ *and* $SS_E = 914.1$.

(b) Using the R output below, what conclusions do you draw? Sort the four species of arthropods based on their number mean amino acids. Reason the answer.

Multiple Comparisons of Means: Tukey Contrasts

Simultaneous tests for contrasts:

Contrast	*Estimate*	*Std. Error*	*t value*	$\Pr(>t)$
$2 - 1 = 0$	39.429	3.299	11.952	$< 1 \times 10^{-5}$
$3 - 1 = 0$	−50.057	3.299	−15.174	$< 1 \times 10^{-5}$
$4 - 1 = 0$	−69.000	3.299	−20.917	$< 1 \times 10^{-5}$
$3 - 2 = 0$	−89.486	3.299	−27.127	$< 1 \times 10^{-5}$
$4 - 2 = 0$	−108.429	3.299	−32.869	$< 1 \times 10^{-5}$
$4 - 3 = 0$	−18.943	3.299	−5.742	2.3×10^{-5}

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
(Adjusted p values reported — single-step method.)

3.5 Further exercises

Exercise 3.5.1. *The temperature of a mixture has been measured using four different thermometers. The obtained data is as follows:*
THERM1: 63; 63; 62; 65; 66;
THERM2: 64; 64; 63; 64; 65;
THERM3: 58; 59; 59; 68;
THERM4: 61; 61; 62; 60; 63;

(a) Taking into account that $SS_B = 34.42$ and $SS_E = 84$, can you reject the hypothesis that the four thermometers measure the same mean temperature? Build the ANOVA table.

(b) Calculate the error of each measurement. Graphically represent these errors of the measurements of each thermometer and analyze if graphically the equality of variances can be contrasted. Is there reason to remove any of the thermometers from the analysis? Reason the answer.

(c) If we remove the third thermometer, the ANOVA table obtained is the one below. What conclusions can you draw from it?

ANOVA table (R output)

Source	*Df*	*Sum Sq*	*Mean Sq*	*F value*	$\Pr(>F)$
termometro	2	20.93	10.47	6.978	0.00977
Residuals	12	18.00	1.50		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Exercise 3.5.2. *Prove the following equalities:*

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\big(y_{ij}-\bar{y}_{i\cdot}\big)^2=\sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-\sum_{i=1}^{I}n_i\,\bar{y}_{i\cdot}^2,$$

$$\sum_{i=1}^{I}n_i\big(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\big)^2=\sum_{i=1}^{I}n_i\,\bar{y}_{i\cdot}^2-n\,\bar{y}_{\cdot\cdot}^2,\qquad \text{where } n=\sum_{i=1}^{I}n_i.$$

Exercise 3.5.3. *Prove that, under the conditions of the Analysis of Variance, the variance of the residuals is the linear combination of an unbiased estimator of the variance of each group.*

Exercise 3.5.4. *An analysis of variance is considered with five modalities where the estimator of the mean of each group is 20, 22, 24, 26 and 28. The number of observations in each group is 10 and the estimate of the variance of the mean of each group is 6. Construct the ANOVA table and draw conclusions from it.*

Exercise 3.5.5. *Assuming $n_1 = n_2$, obtain the contrast (test) of the hypothesis of the comparison of population means of two independent populations with unknown and equal variances, as a particular case of analysis of variance with two groups.*

Exercise 3.5.6. *Calculate the variance of the maximum likelihood estimator of the variance of the model residuals of the analysis of variance.*

Exercise 3.5.7. *When performing the analysis of variance, if the number of observations in each group is equal, $n_i = k$, show that the following statement is true:*

$$\sum_{i=1}^{I}\sum_{j=1}^{I}\left(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}\right)^2 = 2I\sum_{i=1}^{I}\left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2.$$

Exercise 3.5.8. *Suppose that in an industrial process there are eight production lines and each day a sample is taken from each line. For the coefficient of determination (that is, the ratio SS_E/SS_T) to be 10%, with a significance level of 5%, what is the number of days to be sampled?* (Hint: Consider $F_{0.05;7,n-8} \approx 3$.)

4

Non-Parametric Statistics

4.1 Introduction

The diagnosis of the model allows us to know whether the assumption we have made hold. For instance, in the previous chapters, we have assumed sometimes that the characteristic under study of a population follows a given distribution (the normal distribution). In this chapter, we will see how we can do the diagnosis of the models we have seen. More precisely,

- In Section 4.2, we will study hypothesis testings where the goal is to know whether the data under consideration follows a given distribution.
- In Section 4.3, we will focus on the homogeneity of a sample and whether two random variables are independent.
- In Section 4.4, we will study statistical tests that focus on the rank, sign and order.

4.2 Goodness of fit tests

In order to contrast hypotheses and construct confidence intervals, the choice of the estimator and the precision of its estimate are crucial and depend on the pre-established model. A statistical procedure is said to be robust to a hypothesis when the , even if there are small changes in the hypothesis, the entire procedure is more or less valid.

In general, inference techniques for the mean are robust; it doesn't matter what the distribution of the population is, the mean is an unbiased estimator of the expectation,

its variance being σ^2/n, and using the central limit theorem, its distribution is asymptotically normal. Therefore, the contrasts and confidence intervals based on Student's t distribution are more or less independent of the population distribution. In other words, the 95% confidence interval will include 95% of the population mean, in the long run.

But, when the assumptions about the population distribution are false, inference to the mean, although valid, is not optimal. Procedures that assume normality are not accurate when this condition is not met, and the result is that very long intervals or contrasts of low power are obtained.

Inference for variance is very sensitive to the normality assumption. The S^2 estimator is an unbiased estimator of σ^2, but the variance of S^2 depends on the population distribution. When the normality of the population is in doubt, confidence intervals and hypothesis-contrasts for the variance are not recommended.

As a result of all this, it is considered of great important for the diagnosis of the model to perform goodness of fit tests, i.e., to verify whether the data follows a desired distribution.

4.2.1 Pearson's chi square test

It is the oldest contrast for goodness-of-fit and can be used in both discrete and continuous distributions. It is used in the following cases: when we hypothesize that the population has a determined distribution function or that all possible events have concrete probabilities, and when we want to check whether the hypothesis is reasonable through the sample taken from the population.

Suppose that we have a population divided in k classes $Z_1, \dots, Z_k$, where Z_i represents the population of class i, $i = 1, \dots, k$. We denote by $p_1, \dots, p_k$ the set of probabilities that we want to check. This is, our goal is to know if the probability of population i is p_i, for $i = 1, \dots, k$. More precisely,

$$H_0 : \text{the probability of class } 1, \dots, k \text{ is } p_1, \dots, p_k, \text{ respectively}$$
$$H_1 : \text{the probability of class } 1, \dots, k \text{ is NOT } p_1, \dots, p_k.$$

We suppose we have a simple random sample of size n. For this intance, we define the expected value of class i as follows

$$e_i = n\ p_i,\ \ i = 1, \dots, k.$$

Remark 4.2.1. *It is easy to see that* $\sum_{i=1}^{k} e_i = n$.

We require that $e_i > 5$ for all the classes. Otherwise, the classes are merged until this condition is satisfied.

The number of observations on the sample of class-i is denoted as o_i. Therefore, we have that $\sum_i o_i = n$.

The statistical pivot we use to solve this hypothesis test is the following:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}.$$

It can be shown that, when H_0 is true, $\chi^2 \sim \chi^2_{k-l-1}$, where l is the number of parameters that must be estimated so as to fully define the probabilities $p_1, \ldots, p_k$ and k is the number of classes (after merging them, if necessary).

The critical region of this test is $(\chi^2_{k-l-1;\alpha}, \infty)$ whereas the p-value is $p = \mathbb{P}(\chi^2_{k-l-1} > \chi^2)$. The intuition behind these expressions is that, when the value of the statistical pivot is large, there is a large difference between the values of e_i and o_i, therefore the assumption that states that the probabilities are $p_1, \ldots, p_k$ is false.

Proposition 4.2.1.

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - n.$$

Proof.

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{o_i^2 + e_i^2 - 2o_i e_i}{e_i} = \sum_{i=1}^{k} \frac{o_i^2}{e_i} + \sum_{i=1}^{k} e_i - 2\sum_{i=1}^{k} o_i = \sum_{i=1}^{k} \frac{o_i^2}{e_i} + n - 2n$$
$$= \sum_{i=1}^{k} \frac{o_i^2}{e_i} - n.$$

□

Remark 4.2.2. *We see why χ^2 follows the chi square distribution when H_0 is true. We focus on a class i of the population and we divide the population in two, whether an observation belongs to class i or not. In this case, we have that O_i, which is the random variable of the number of observations of class i in a sample of size n, follows a binomial distribution where the probability of success is p_i. When n is large and p_i small, it can be approximated by the Poisson distribution with parameter np_i, in which case the expected value and the variance is np_i. Thus,*

$$\frac{O_i - np_i}{np_i} \sim \mathcal{N}(0, 1)$$

and, therefore, the statistical pivot is the sum of the square of normal standard distributions, i.e., it follows the chi square distribution.

Another important remark is presented now. When we do not reject H_0 (when the sample data does not belong to S_1 or when the p-value is larger than the significance level), we conclude that, with this method, we cannot say that the data does not follow the probability distribution under consideration. This does not mean that the data

follows the probability distribution. In fact, we can check other tests different from this one to check this property.

4.2.2 Kolmogorov-Smirnov test

With this test we will check whether the data follows a given cumulative distribution function $F_0(x)$ (the theoretical distribution). This test is only valid for continuous distributions.

Let $F_n(x)$ be the cumulative distribution function (or empirical distribution) of a sample data of size n. We define the null and the alternative hypothesis for this test as follows:

$$H_0 : F_n(x) = F_0(x)$$
$$H_1 : F_n(x) \neq F_0(x).$$

We order the sample data so as to get the following

$$x_1 \leq x_2 \leq \cdots \leq x_n,$$

where x_i is the i-th data of the sample.

We compute the empirical distribution of the sample data as follows:

- if $x < x_1$, $F_n(x) = 0$
- if $x_k \leq x \leq x_{k+1}$, $F_n(x) = \frac{k}{n}$
- Si $x \geq x_n$, $F_n(x) = 1$.

In Figure 4.1, we show an example of the functions F_0 (in black) and F_n (in blue). As it can be seen in this illustration, F_0 is stepwise constant non-decreasing.

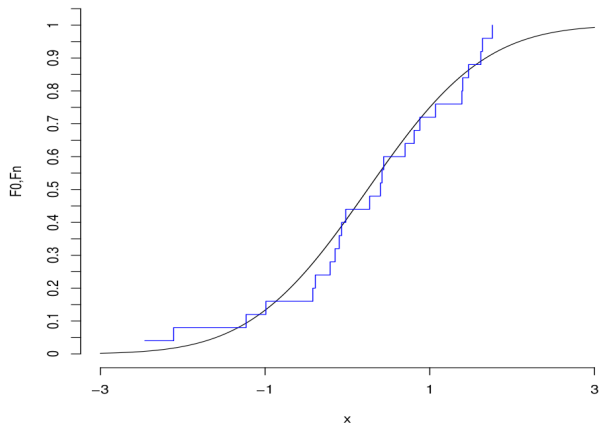


Figure 4.1: An example of F_0 (in black) and F_n (in blue).

The statistical pivot we will use in this test is the maximum of the absolute value of the difference between the empirical distribution and the theoretical one, i.e.,

$$D_n = \max |F_n(x) - F_0(x)|.$$

The computation of the statistical pivot needs to compare $F_n(x)$ and $F_0(x)$ for all $x \in [x_1, x_n]$. This can be extremely difficult in practice. To overcome this difficulty, as it can be seen in Figure 4.2, we only need to compare these values at the points $x_1, \ldots, x_n$.

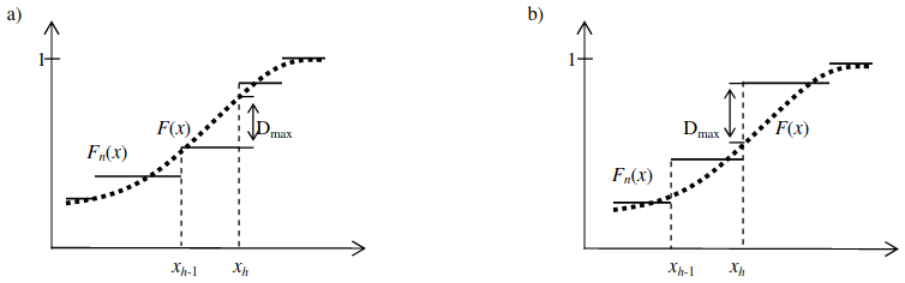


Figure 4.2: The computation of D_n.

The distribution of the statistical pivot D_n is given in Table 7. As in the case of the χ^2 test, we will reject H_0 if the value of the statistical pivot is large (which means that the difference between the empirical and theoretical distributions are not always very close). Specifically, let d_n be the value of the statistical pivot of the sample data, we have that

- if $d_n \geq D_{\alpha;n}$, then the p-value is smaller than the significance level α and, as a result, we reject H_0 (and otherwise we cannot reject H_0)
- the critical region is $S_1 = \{(x_1, \ldots, x_n) | d_n \geq D_{\alpha;n}\}$ and therefore, if the sample belongs to the critical region, we reject H_0 (and otherwise we cannot reject H_0)

Remark 4.2.3. *We have said that D_n follows the distribution of Table 7. However, when we check the normality, it is preferable to consider the values of the Table 8 instead.*

4.3 Independence and homogeneity tests

In this section we will explain hypothesis contrasts for comparing two or more distributions. This will allow us to analyze a problem that often appears in practice, that is, to decide whether or not there is a dependency between two discrete variables. This test can be used to answer questions such as: Is there a relationship...

- Between getting a job and being a woman?
- Hypertension and smoking?

- Between the color and the smell of flowers?

- Between the purchased car and the job?

Consider a discrete random variable A with r classes and another random variable B with s classes. A contingency table is a matrix with r rows and s columns that represents the frecuencies of all the possible outcomes. In a simple random sample of size n, we denote by $o_{i,j}$ the number of observations of class i of random variable A and class j of random variable B. We define $z_{\cdot,j}$ as the marginal of class j, i.e., $z_{\cdot,j} = \sum_{i=1}^{r} o_{i,j}$. Likewise, we define $f_{i,\cdot}$ as the marginal class i, i.e., $f_{i,\cdot} = \sum_{j=1}^{s} o_{i,j}$.

We present an example of the previous definitions in a contingency table.

	1	2	3	...	s	
1	o_{11}	o_{12}	o_{13}	...	o_{1s}	$f_{1\cdot}$
2	o_{21}	o_{22}	o_{23}	...	o_{2s}	$f_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	o_{r1}	o_{r2}	o_{r3}	...	o_{rs}	$f_{r\cdot}$
	$z_{\cdot 1}$	$z_{\cdot 2}$	$z_{\cdot 3}$	...	$z_{\cdot s}$	n

Figure 4.3: A contingency table.

4.3.1 Methodology

Two type of tests will be studied using contingency tables:

- Homogeneity test: We consider that one of the random variables is under control of the designer. The goal is to know whether the other random variable satisfies a homogeneity condition, i.e., it comes from the same population. In this case,

$$H_0 : \text{The homogeneity is achieved}$$
$$H_1 : \text{The homogeneity is NOT achieved.}$$

- Independent test. We aim to know if two discrete random variables are indepent.

In this case,

$$H_0 : \text{The random variables are independent}$$
$$H_1 : \text{The random variables are dependent.}$$

The good news is that the methodology in both tests is the same. We explain it with the independent test only.

We first compute the expected value of each of the possible values of the contingency table, which is

$$e_{i,j} = \mathbb{P}(A = a_i \cap B = b_j).$$

Assuming that H_0 is true, we have

$$\mathbb{P}(A = a_i \cap B = b_j) = \mathbb{P}(A = a_i)\mathbb{P}(B = b_j).$$

and moreover,

$$\mathbb{P}(A = a_i) = \frac{f_{i,\cdot}}{n}, \quad \mathbb{P}(B = b_i) = \frac{z_{\cdot,j}}{n}.$$

Therefore, the expected values can be computed as $e_{i,j} = \frac{f_{i,\cdot} z_{\cdot,j}}{n}$, for $i = 1, \ldots, r$ and $s = 1, \ldots, s$. We require that $e_{i,j} > 5$ for all i,j and, in the negative case, we merge the groups until we get it.

The statistical pivot we will use to solve the test is

$$\chi^2_p = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}.$$

Remark 4.3.1. *The above statistical pivot can be alternatively written as follows:*

$$\chi^2_p = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{o^2_{i,j}}{e_{i,j}} - n.$$

The distribution of the statistical pivot follows the chi square distribution with $(r-1)(s-1)$ degrees of freedom. Note that r and s is the number of rows and columns of the contingency table after merging groups, if necessary.

We remark that large values of χ^2_p imply that there are some values of i and j for which the difference between $o_{i,j}$ and $e_{i,j}$ is large. Therefore, when this occurs, we know that the expected values different from the observed ones, which leads to a dependency of the random variables. More precisely, we define the p-value as

$$p = \mathbb{P}(\chi^2_{(r-1)(s-1)} > \chi^2_p).$$

As in the previous cases, we will reject H_0 (and, therefore, we conclude that the random variables are dependent) when the p-value is smaller than the significance level

α.

4.3.2 Correction of Yates

When we consider a contingency table of size 2×2, we need to take into account that the statistical pivot does not always follow the χ^2 distribution with one degree of freedom. When this occurs, we consider the Correction of Yates. This consists of taking the following as the statistical pivot

$$\chi^2_Y = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(|o_{i,j} - e_{i,j}| - \frac{1}{2})^2}{e_{i,j}}.$$

which follows the χ^2 distribution with one degree of freedom when

$$|o_{1,1}o_{2,2} - o_{1,2}o_{2,1}| > n/2.$$

That is, when the above condition is satisfied we use χ^2_Y instead of χ^2_p and, if the above condition is not satisfied, we use χ^2_p.

Remark 4.3.2. *In the 2×2 tables, we need to check also that $e_{i,j} > 5$ for all i,j and apply the Correction of Yates if necessary. However, when $e_{i,j} > 5$ for all i,j is not satisfied, there is an alternative method that is called the exact test of Fisher, which we are not studying in this course.*

4.4 Tests of position

4.4.1 The test of signs

In section 2, we studied hypothesis testing for the mean, in which case we assumed that the population distribution is normal or that the size of the sample is large ($n \geq 30$) so that the Central Limit Theorem applies. In practice, these assumption are not met very often. In this part, we will study how to deal with this hypothesis testing when normality is not achieved using the test of signs, which is the simplest test for this task.

The idea of this test is to consider that half of the population should satisfy that its value is larger than μ_0 and half smaller, when H_0 is true. Therefore, if the number of values that are larger and smaller than μ_0 is very different, then null hypothesis is rejected.

We will perform this test following the next steps:

1. Formulate the null and the alternative hypothesis

2. We compute the difference of the values of the sample and μ_0 to compute the variable D, i.e., $D = X - \mu_0$.

3. We compute the statistical pivot. The statistical pivot we consider is denoted as S^+, which is the number of positive values of D. We denote by s the statistic we obtain in the sample.

   When H_0 is true, the distribution of S^+ is binomial with parameters n and $1/2$, i.e., $S^+ \sim Bin(n, 1/2)$.

4. We compute the p-value, which will depend on the type of hypothesis testing formulated.

   If $H_1 : \mu < \mu_0$, then $p = \mathbb{P}(S^+ \leq s)$

   If $H_1 : \mu > \mu_0$, then $p = \mathbb{P}(S^+ \geq s)$

   If $H_1 : \mu \neq \mu_0$, then when $s < n/2$, $p = 2\mathbb{P}(S^+ \leq s)$ and when $s \geq n/2$, $p = 2\mathbb{P}(S^+ \geq s)$.

5. We compare the p-value with the significance level. If the p-value is smaller than the significance level, the null hypothesis is rejected. Otherwise, we cannot reject H_0.

Remark 4.4.1. *The probability of getting that $D = 0$ is zero since the values are sampled from a continuous random variable. However, this occurs often in practice. In these cases, there are several solutions. In this course, we will remove these values from the sample and, therefore, the size of the sample gets reduced.*

Example 16. *We obtain the following sample:* $1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2$ *and* 1.7 *and we want to check if they are sampled from a population whose mean is 1.8 with* $\alpha = 0.05$*. The population is not normal.*

We formulate the following hypothesis testing:

$$H_0 : \mu = 1.8$$
$$H_1 : \mu \neq 1.8$$

x_i	1.5	2.2	0.9	1.3	2.0	1.6	1.8	1.5	2.0	1.2	1.7
$d_i = x_i$ - 1.8	-0.3	0.4	-0.9	-0.5	0.2	-0.2	0	-0.3	0.2	-0.6	-0.1
	–	+	–	–	+	–		–	+	–	–

From the above table, we get that s=3 (there are three positive values) and we observe that there is single value equal to 1.8 (which implies that one value is removed from the

sample), which means that $n = 10$. *Since, in this case,* $s < n/2$ *the p-value is computed as follows:*

$$p = 2\mathbb{P}(S^+ \leq 3),$$

where $S \sim Bin(10, 1/2)$. *Thus,*

$$p = 2\mathbb{P}(S^+ \leq 3) = 2(\sum_{i=0}^{3} \mathbb{P}(S^+ = i)) = 0.3438,$$

which is larger than 0.05. *As a result, we do not reject* H_0, *which means that the sample can come from a population with mean 1.8.*

Remark 4.4.2. *This test can also be used to analyze the difference between paired data. In this case, we consider that the values of D are obtained by computing the difference between both random variables.*

4.4.2 Test of Wilcoxon of signed ranks

The test of signs uses only the signs of the differences (between μ_0 and the observations when it is a sample and the difference between pairs of observations when it is pairwise data) and ignores the magnitudes of these differences. The non-parametric test that uses both the sign and the size of the differences was proposed by Wilcoxon in 1945, the Wilcoxon test of signed ranks.

As in the test of signs, we will compute a new variable whose are obtained by performing the difference between the sample values and μ_0. Moreover, the values of D equal to zero are removed from the sample. We will order the obtained values of D in increasing order. To each of the positions, a value is assigned. This value will be the rank. If several values have the same value of D, their rank is the mean of them.

We denote by T_+ the sum of the ranks whose value of D is positive and by T_- the sum of the ranks whose value of D is negative. A large value of T_+ in the sample leads to conclude that the null hypothesis $H_0 : \mu < \mu_0$ is not true. For $H_0 : \mu > \mu_0$, a small value of T_+ says that H_0 is not true. Finally, when $H_0 : \mu = \mu_0$, large or small values of T_+ in the sample lead to conclude that H_0 is not true.

We have that

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

When H_0 is true, T_+ and T_- are equally distributed and therefore,

$$\mathbb{E}[T_+] = \mathbb{E}[T_-] = \frac{n(n+1)}{4},$$

and

$$Var[T_+] = Var[T_-] = \frac{n(n+1)(2n+1)}{24}.$$

We carry out this test following the next steps:

1. Formulate the null and the alternative hypothesis

2. We compute the difference of the values of the sample and μ_0 to obtain the variable D, i.e., $D = X - \mu_0$.

3. We order the values of D in an increasing order according to $|D|$.

4. We assign the rank to each of the values. The rank of the value i is denoted by $h(i)$.

5. We compute the statistical pivot of the test, which is the maximum between the sum of the ranks of the positive values and the sum of the ranks of the negative values, i.e., $t_0 = \max(t_+, t_-)$, where t_+ and t_- are the values of the random variables T_+ and T_-, respectively, obtained in the sample.

   When H_0 is true, the distribution of T_+ and T_- is the same, in which case we denote it by T. The distribution of T is given on Table 9.

6. We compute the p-value, which will depend on the type of hypothesis testing formulated.

   In case of a unilateral test, then $p = \mathbb{P}(T \geq t_0)$

   In case of a bilateral test, then $p = 2\mathbb{P}(T \geq t_0)$

7. We compare the p-value with the significance level. If the p-value is smaller than the significance level, the null hypothesis is rejected. Otherwise, we cannot reject H_0.

Example 17. *We solve the test that in the previous section has been solved using the Test of signs. Thus, from the following sample*

$$1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7,$$

we want to check if the data is sampled from a population whose mean is 1.8 with $\alpha = 0.05$. *The population is not normal.*

We formulate the hypothesis testing:

$$H_0 : \mu = 1.8$$
$$H_1 : \mu \neq 1.8$$

x_i	1.5	2.2	0.9	1.3	2.0	1.6	1.8	1.5	2.0	1.2	1.7
$d_i = x_i - 1.8$	-0.3	0.4	-0.9	-0.5	0.2	-0.2	0	-0.3	0.2	-0.6	-0.1
$h(i)$	5.5	7	10	8	3	3		5.5	3	9	1

We observe that $t_+ = 7 + 3 + 3 = 13$ *and* $t_- = 5.5 + 10 + 8 + 3 + 5.5 + 9 + 1 = 42$. *Therefore,* $t_0 = max(13, 42) = 45$. *Since* $n = 10$, *we now compute the p-value as follows*

$$p = 2\mathbb{P}(T \geq 42) = 2 \cdot 0.08 = 0.16.$$

We observe that the p-value obtained using this test is smaller than the p-value obtained with the test of signs. In both cases, the p-value is larger than the significance level and, therefore, we cannot reject that $\mu = 1.8$.

We consider a different example now.

Example 18. *We consider the following sample*

$$119, 120, 125, 122, 118, 117, 126, 114, 115, 123, 121, 120, 124, 127, 126,$$

and we want to check if the data is sampled from a population whose mean is 120 with $\alpha = 0.05$. *The population is not normal.*

We formulate the hypothesis testing:

$$H_0 : \mu = 120$$
$$H_1 : \mu \neq 120$$

We obtain that $t_+ = 61$ *and* $t_- = 30$. *Therefore,* $t_0 = max(61, 30) = 61$. *Since* $n = 13$, *we now compute the p-value as follows*

$$p = 2\mathbb{P}(T \geq 61) = 2 \cdot 0.153 = 0.306.$$

We observe that the p-value is larger than the significance level and, therefore, we cannot reject that $\mu = 120$.

Remark 4.4.3. *When* $n \geq 15$, *the statistic* t_+ *approximates the normal distribution. Specifically,*

$$t_+ \sim \mathcal{N}(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}).$$

Therefore, in Table 9, we have only shown the values of the probabilities when $n \leq 15$. *In the rest of the cases, we use the normal approximation to compute the p-value.*

Remark 4.4.4. *We observe that, in Table 9, we only consider* $t \geq \frac{n(n+1)}{4}$. *When this*

does not occur, we use the following property:

$$\mathbb{P}(T \leq t) = \mathbb{P}\left(T \geq \frac{n(n+1)}{2} - t\right)$$

This property is given after Table 9.

4.4.3 Wilcoxon test of the sum of ranks (Mann-Whitney)

The two methods that we have discussed before, the sign test and the Wilcoxon signed-ranks test, are non-parametric methods that are used instead of the parametric Student's t test defined in one-sample or paired data.

The non-parametric method for testing the equality of the means of two independent populations was proposed by Wilcoxon, it is called the Wilcoxon test of the sum of ranks, and it is used instead of the Student's t test defined for two samples, especially when the normality of the population is in doubt.

We aim to study the following hypothesis testing:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Without loss of generality, we assume that $n_1 \leq n_2$ (i.e., the population one is chosen to be that of the smallest sample size). We order the values of both samples and we assign a rank to each of them (with the same method as for the Wilcoxon test). The statistical pivot we will consider is W_1, which is the sum of the ranks of the first sample (that of the smallest sample size). We have that, when H_0 is true,

$$\mathbb{E}[W_1] = \frac{n_1(n_1+n_2+1)}{2}, \qquad Var[W_1] = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

Let W_2 the sum of the ranks of the second sample. Thus,

$$W_1 + W_2 = \sum_{i=1}^{n_1+n_2} \frac{(n_1+n_2)(n_1+n_2+1)}{2}.$$

As before, we will use the value of the statistic obtained from the sample, which is denoted by w_1, to compute the p-value. Then, the p-value will be compared with the significance level to conclude whether we can reject H_0 or not.

We carry out this test following the next steps:

1. Formulate the null and the alternative hypothesis.

   In general, this test is used to check whether two samples follow the same distribution.

2. We order the values of both samples in an increasing order.

3. We assign the rank to each of the values. The rank of the value i is denoted by $h(i)$. When several values coincide, we assign the mean of them to all.

4. We compute the statistical pivot of the test which is the sum of the ranks of the first sample. When H_0 is true, the distribution of W_1 is given in Table 10.

5. We compute the p-value, which will depend on the type of hypothesis testing formulated.

   In case of a unilateral right-tailed test, then $p = \mathbb{P}(W_1 \geq w_1)$

   In case of a unilateral left-tailed test, then $p = \mathbb{P}(W_1 \leq w_1)$

   In case of a bilateral test, then

   - when $w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$, then $p = 2\mathbb{P}(W_1 \geq w_1)$
   - when $w_1 < \frac{n_1(n_1+n_2+1)}{2}$, then $p = 2\mathbb{P}(W_1 \leq w_1)$

6. We compare the p-value with the significance level. If the p-value is smaller than the significance level, the null hypothesis is rejected. Otherwise, we cannot reject H_0.

Example 19. *We aim to check whether there is a significance difference in the mean of the populations where the following two samples come from:*

- *Sample of population X: 75, 82, 28, 82, 94, 78, 76, 64*
- *Sample of population Y: 78, 95, 63, 37, 48, 74, 65, 77, 63*

We observe that the sample size of X is smaller than the sample size of Y. Therefore, we choose X to be the first sample and Y the second one. Like this, we have that $n_1 = 8$ *and* $n_2 = 9$,, *which implies that* $n_1 \leq n_2$ *is verified.*

We aim to study the following hypothesis testing:

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y,$$

where μ_X *is the mean (or expected value) of population X and* μ_Y *of population Y.*

In the following table, we present the computations of this test. In the first row, there are the values of the both samples in order. In the second row, there are the ranks

28	37	48	63	63	64	65	74	75	76	77	78	78	82	82	94	95
1	2	3	4.5	4.5	6	7	8	9	10	11	12.5	12.5	14.5	14.5	16	17
X	*Y*	*Y*	*Y*	*Y*	*X*	*Y*	*Y*	*X*	*X*	*Y*	*X*	*Y*	*X*	*X*	*X*	*Y*

assigned to each of the values. In the last row, it is indicated to which population is each value.

From this table, we get that $w_1 = 1+6+9+10+12.5+14.5+14.5+16 = 83.5$ *In our case, we have a bilateral test and since* $\frac{n_1(n_1+n_2+1)}{2} = 72$*, we have that* $w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$*, which implies that*

$$p = 2\mathbb{P}(W_1 \geq 83.5)$$

and using the values of Table 10, we have that

$$p = 2\mathbb{P}(W_1 \geq 83.5) \geq 2\mathbb{P}(W_1 \geq 84) = 2 \cdot 0.138 = 0.277$$

As a result, we cannot reject H_0 *with* $\alpha = 0.05$*, i.e., with a significance level of* 0.05 *we cannot ensure that there is a significance difference on the mean of the populations of these samples.*

Remark 4.4.5. *When* $n_1 \geq 10$ *and* $n_2 \geq 10$*, the statistic* w_1 *approximates the normal distribution. Specifically,*

$$w_1 \sim N\left(\mathbb{E}[W_1], Var[W_1]\right),$$

where $\mathbb{E}[W_1]$ *and* $Var[W_1]$ *have been defined above.*

Therefore, in Table 10, we have only shown the values of the probabilities when $n_1 \leq 10$ *and* $n_2 \leq 10$*. In the rest of the cases, we use the normal approximation to compute the p-value.*

Remark 4.4.6. *We observe that, in Table 10, we only consider*

$$w_1 \geq \frac{n_1(n_1 + n_2 + 1)}{2}$$

. When this does not occur, we use the following property:

$$\mathbb{P}(W_1 \leq w_1) = \mathbb{P}(W_1 \geq n_1(n_1 + n_2 + 1) - w_1).$$

This property is given after Table 10.

Example 20. *We have two samples, each one of them is picked from a different population. We aim to know if there is a significance difference in the mean of these populations:*

The sample size of both population is equal. We choose the first one to be X.

We formulate the following hypothesis testing:

X	0.464	0.060	1.486	1.022	1.394	0.906	1.179	-1.501	-0.69
Y	0.672	1.187	1.785	1.194	0.742	2.579	2.090	1.448	0.543

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y,$$

where μ_X is the mean of population X and μ_Y of population Y.

We get that $w_1 = 1+2+3+4+8+9+10+13+15 = 65$. In our case, we have a bilateral test and since $\frac{n_1(n_1+n_2+1)}{2} = 85.5$, we have that $w_1 < 85.5$ and, therefore, using that $\frac{n_1(n_1+n_2+1)}{2} = 171$, we compute the p-value as follows:

$$\begin{aligned} p =& 2\mathbb{P}(W_1 \geq 65) = 2(1 - \mathbb{P}(W_1 \leq 65)) = 2(1 - \mathbb{P}(W_1 \geq 171 - 65) = 2(1 - \mathbb{P}(W_1 \geq 106)) \\ &= 2 \cdot 0.039 = 0.078. \end{aligned}$$

As a result, we cannot reject H_0 with $\alpha = 0.05$, i.e., with a significance level of 0.05 we cannot ensure that there is a significance difference on the means of the population of these samples.

4.4.4 Kruskal-Wallis test

The idea of using the sum of ranks to compare two populations based on independent random samples can be extended to more than two populations. The statistical test for this was developed by W. H. Kruskal and W. A. Wallis in 1952, which is why it is called the (Kruskall-Wallis) test.

Suppose that we have random samples of sizes n_1, n_2, ...,n_k that are selected from k independent populations, respectively. Let $N = n_1 + n_2 + ... + n_k$. We want to check whether these populations have the same mean, i.e.,

$$H_0 : \mu_i = \mu_j, \forall i \neq j$$
$$H_1 : \exists i, j \ s.t. \mu_i \neq \mu_j.$$

Each value is then assigned a rank from 1 to N. In the case of ties, they are replaced by the mean, in the same way as in the Wilcoxon test.

Let r_i be the sum of ranks of the values drawn from the i-th population, where $i = 1, 2, \ldots, k$. The statistical pivot of this test is the following one:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\frac{r_i^2}{n_i} \right) - 3(N+1).$$

When H_0 is true, we have that $H \sim \chi^2_{k-1}$. Therefore, the p-value is given by:

$$p = \mathbb{P}(\chi^2_{k-1} > H).$$

The p-value is then compared with the considered significance level and, as in the previous cases, if it is smaller than α, H_0 is rejected.

Example 21. *We aim to analyze the effect of three diets in animals with $\alpha = 0.05$. The data obtained by sampling is given in the following table.*

1	104	108	107	106
2	112	115	118	116
3	120	124	114	112

We observe that $n_1 = n_2 = n_3 = 4$ and therefore $n = 12$. We formulate the following hypothesis testing:

$H_0 : \mu_1 = \mu_2 = \mu_3$
H_1 : *There is a difference on the mean of the populations*,

In the following table, we present the computations of this test. In the first row, there are the values of the both samples in order. In the second row, it is indicated to which population is each value and in the last row, there are the ranks assigned to each of the values.

104	106	107	108	112	112	114	115	116	118	120	124
1	1	1	1	2	3	3	2	2	2	3	3
1	2	3	4	5.5	5.5	7	8	9	10	11	12

From the above table, we get that $r_1 = 1+2+3+4 = 10$, $r_2 = 5.5+8+9+10 = 32.5$ and $r_3 = 5.5 + 7 + 11 + 12 = 35.5$. Therefore, the statistical pivot is

$$H = \frac{12}{12 \cdot 13} \sum_{i=1}^{k} \left(\frac{r_i^2}{n_i} \right) - 3 \cdot 13 = \frac{12}{12 \cdot 13} \left(\frac{10^2}{4} + \frac{32.5^2}{4} + \frac{35.5^2}{4} \right) - 3 \cdot 13 = 7.4711.$$

Hence,

$$p = \mathbb{P}(\chi^2_2 > 7.4711) < \mathbb{P}(\chi^2_2 > 7.378) = 0.025,$$

which implies that H_0 is rejected with $\alpha = 0.05$, i.e., they do not have all the same mean.

4.5 Fundamental exercises

Exercise 4.5.1. *The following table shows the results obtained after rolling a dice 100 times. What can you say about the fairness of this dice ($\alpha = 0.05$)?*

Result	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*10*	*14*	*25*	*16*	*20*	*15*

Exercise 4.5.2. *Analyze, at a significance level of 1%, whether the following numbers follow a uniform distribution between 0 and 1: 0.49; 0.62; 0.57; 0.53; 0.41; 0.46; 0.32; 0.80; 0.86; 0.68; 0.56; 0.28; 0.63; 0.53; 0.48; 0.21.*

Exercise 4.5.3. *The following table shows the grades obtained by students at the secondary level in Mathematics and Social Sciences. What can it be said about the relationship between the grade and the subject? ($\alpha = 0.05$)*

	Low	*Good*	*Excellent*
Mathematics	*10*	*18*	*7*
Social Sciences	*3*	*20*	*27*

Exercise 4.5.4. *From the data below, we aim to whether the proportion in which each type of bleeding occurs is the same for women who have taken drugs during pregnancy and those who have not.*

	Took drugs	*Did not take drugs*	*Total*
normal bleeding	*28*	*12*	*40*
not normal bleeding	*332*	*228*	*560*
Total	*360*	*240*	*600*

Exercise 4.5.5. *The following table shows the ages of drivers over 21 years of age and the number of accidents they have suffered. We want to study whether age influences the number of accidents.*

	No accident	*One accident*	*More than one accident*
21-30	*748*	*74*	*5*
31-40	*821*	*60*	*6*
41-50	*786*	*51*	*3*
51-60	*720*	*66*	*4*
61-70	*672*	*50*	*7*

(a) Check that age and number of accidents are independent ($\alpha = 0.05$).

(b) If we consider two age groups, one older than 50 and the other younger, can it be said that the age of those who have accidents and those who do not is the same?

Exercise 4.5.6. *The following data indicates the blood glucose level (in mg/kg) of rabbits before and two hours after giving them a pain reliever.*

Rabbit Id.	*Before*	*After*
1	*158*	*206*
2	*119*	*134*
3	*122*	*204*
4	*89*	*105*
5	*111*	*96*
6	*135*	*171*
7	*138*	*212*
8	*122*	*134*
9	*127*	*177*
10	*127*	*136*
11	*137*	*137*
12	*120*	*117*
13	*118*	*127*
14	*126*	*140*
15	*134*	*153*
16	*134*	*147*
17	*125*	*131*
18	*124*	*131*

Analyze whether the analgesic has a significance effect using the test of signs and the Wilcoxon test of sum of signs. If the required conditions were met, carry out the parametric test to perform this analysis? What are these conditions? Use $\alpha = 0.05$.

Exercise 4.5.7. *The following table shows the time until the battery runs out of two types of calculators:*

- *Type A: 4.5; 4.6; 5.3; 4.6; 4.3; 5.0; 5.2; 4.8; 5.1;*
- *Type B: 3.8; 3.5; 4.3; 3.2; 4.0; 3.9; 4.5; 4.2; 4.5;*

Using the Mann-Whitney test with a significance level of 5%*, analyze the lifetime of type A calculators is higher than type B.*

Exercise 4.5.8. *At the beginning of an academic year, a teacher has 4 large groups of students and wants to know if the knowledge is the same across groups. For this purpose, the teacher takes a random sample of each group and the scores obtained by each of the selected students is shown here:*

- *Group 1: 165, 187, 173, 179, 181, 169*
- *Group 2: 175, 169, 183, 181, 172, 179, 190*
- *Group 3: 159, 178, 167, 162, 183, 176*
- *Group 4: 194, 189, 180, 188*

Can you say, with $\alpha = 0.05$, that there is a significance difference on the mean scores obtained between groups?

4.6 Further exercises

Exercise 4.6.1. *In World War II, the map of London was divided in parts of $1/4$ km^2 square and the number of bombs in each part was counted. The results obtained are presented in the following table. With a significance level of 5%, what can you say about the idea that the number of bombs dropped follows a Poisson distribution?*

Number of bombs	*0*	*1*	*2*	*3*	*4*	*5*
Number of squares	*200*	*190*	*80*	*23*	*7*	*0*

Exercise 4.6.2. *The following table shows the number of draws (not winning and not losing) for a football team over two consecutive seasons. What can you say about the hypothesis that these data follow a Poisson distribution ($\alpha = 0.05$)?*

Amount of ties	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*4*	*15*	*20*	*12*	*8*	*3*	*2*

Exercise 4.6.3. *We have asked the number of keys carried in the pocket to 100 people taken at random. What can you say about the hypothesis that these data follow a Poisson distribution ($\alpha = 0.05$)?*

Number of keys	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Frequency	*4*	*9*	*20*	*15*	*21*	*11*	*8*	*6*	*4*	*2*

Exercise 4.6.4. *A box has red, green, white, yellow, and blue balls. With a significance level of 5%, we want to check whether the proportions of the balls of each color follow the next probability distribution: P(Red)=P(Green)=0.15, P(White)=P(Yellow)=0.25 and P(Blue)=0.2.*

We take 200 balls with replacement and observe: 35 red, 25 green, 38 white, 45 yellow and 57 blue. What can you say about the hypothesis ($\alpha = 0.05$)?

Exercise 4.6.5. *100 guinea pigs of different colors have been crossed and 44 red, 25 black and 31 white have been obtained. According to classical genetic models, these numbers should appear in the ratio 9 red, 3 black, and 4 white. With a significance level of 1%, what can you say?*

Exercise 4.6.6. *In the table below, the number of war declarations involving armed actions with more than 50,000 troops per year, from the 17th century to the end of the 20th century, is shown. Study the hypothesis that these data follow a Poisson distribution ($\alpha = 0.05$).*

Number of declarations of war	*0*	*1*	*2*	*3*	*4*
Number of years	*215*	*120*	*45*	*15*	*5*

Exercise 4.6.7. *10 9-volt batteries have been taken and their duration measured (in hours): 27.8, 14.1, 27.6, 71.4, 47.5, 51.3, 36.5, 48.4, 61.2, 54.2*

- *Study the hypothesis that these data follow an exponential distribution ($\alpha = 0.05$).*
- *It is suspected that the duration follows an exponential distribution with parameter 50. What would you say?*

Exercise 4.6.8. *The height (in cms) of 12 newborns is measured: 46, 54, 57, 44, 45, 42, 43, 54, 51, 46, 43, 55. What can we say about the normality of this data ($\alpha = 0.05$)?*

Exercise 4.6.9. *Michelson obtained the following adjusted (299000 subtracted) data for the speed of light: 850, 740, 900, 1070, 930, 850, 950, 980, 880, 1000, 980, 930, 650, 760, 980*

Newcomb later obtained: 883, 816, 778, 796, 682, 711, 611, 599, 1051, 781, 578, 796, 774, 820, 772

- *At the 5% level, test whether the mean speed of light is the same (assume normality).*
- *Test whether each data set is normal. Based on that, would you trust the previous test?*

Exercise 4.6.10. *It is suspected that the number of children born with deformation is not uniform in the 5 regions of a city. The data are:*

Region	*Deformed births*	*Total births*
1	*62*	*1740*
2	*38*	*1498*
3	*25*	*942*
4	*38*	*1312*
5	*50*	*1326*

Formulate the required statistical test and solve it ($\alpha = 0.05$).

Exercise 4.6.11. *200 people were asked whether they prefer butter or margarine, divided by gender:*

	Butter	*Margarine*
Men	*40*	*60*
Women	*70*	*30*

Analyze whether preference depends on sex ($\alpha = 0.05$).

Exercise 4.6.12. *Suppose the null hypothesis $H_0 : p_1 = p_2$, that is, a two-sided test of the success rate of two populations binomial and independent. Prove that the statistic used in topic two to test the equality of two proportions is the square of the chi-square statistic used in this unit.*

Exercise 4.6.13. *We want to measure the efficiency of streptokinase in patients who have suffered a heart attack. For which, they have taken part both sick as a control group. In the heart attack group, 2 patients out of 15 died within a year: while In the control group, of 19 patients, 4 died.*

- *Compare the result when solving the test assuming normality and using the chi-square test of independence.*
- *Based on the above data, it was considered that in the control group the probability that a person would die after a year is $p_1 = 0.2$ and let us suppose that the probability that among the heart attack patients a person dies after a year with probability p_2. To contrast the difference in the death rate in the two groups, a level of significance of 5% and a power of 90%. Assuming that the sample size in both groups is the same, express the necessary sample size based on p_2.*

Exercise 4.6.14. *Consider the test of ranks and signs of Wilcoxon for a sample size of $n = 4$.*

- *Give all the possible combinations for the ranks 1,2,3 and 4 as well as the value of W^+ in each case.*
- *Calculate the probability distribution of W^+ using the result of the previous section.*
- *Using the previous section, calculate $\mathbb{E}[W^+]$ and $Var[W^+]$ and compare it with the normal approximation values.*

Exercise 4.6.15. *Suppose that to perform the test of ranks and signs of Wilcoxon we use the following statistical pivot:*

$$D = X - \mu_0.$$

Compute $\mathbb{E}[W^+]$ and $Var[W^+]$ using the following random variable:

$$U_i = \begin{cases} 1 & D_i > 0 \\ 0 & D_i < 0 \end{cases}$$

for all $i = 1, \dots, n$ and $W^+ = U_1 + 2U_2 + \cdots + nU_n$. Bear in mind that, since D is a continuous random variable, we have that $\mathbb{P}(D = 0) = 0$.

Exercise 4.6.16. *The beats per minute of children over one year of age are analyzed and the following sample of 10 children is taken: 119, 120, 125, 122, 118, 117, 126,*

114, 115, 123, 121, 120, 124, 127, 126. Suppose that the beats per minute follow a distribution continuous and symmetric. You want to check if the median of the beats per minute of children over one year old is 120. Use $\alpha = 0.05$.

Exercise 4.6.17. *The time to give birth to children with spontaneous death at birth and children with birth without problems is analyzed, obtaining the following data:*

- *Deaths: 60, 25, 6, 8,* < 5*, 10, 25, 15, 10.*
- *Without problem: 13, 20, 15, 7, 75, 120, 10, 100, 9, 25.*

Check if there is a difference in the distribution function of the time of giving birth between the two groups. Use $\alpha = 0.05$.

Exercise 4.6.18. *We want to compare the efficacy of traditional medicine (medications) and psychotherapy in the treatment of nervousness. For this, data is obtained on the level of improvement (on a scale from 1 to 10, where 1 means that the improvement is maximum and 10 that it is minimum) after two months of treatment using each of the methods. You get the following data:*

Traditional medicine		*Psychotherapy*	
Improvement Level	*Frequency*	*Improvement Level*	*Frequency*
1	*2*	*1*	*0*
2	*4*	*2*	*2*
3	*7*	*3*	*5*
4	*3*	*4*	*3*
5	*2*	*5*	*8*
6	*1*	*6*	*4*
7	*0*	*7*	*2*
8	*0*	*8*	*1*
9	*0*	*9*	*1*
10	*0*	*10*	*0*

- *Why do you think the test of* χ^2 *is not adequate in this case?*
- *What non-parametric test would you use to perform this analysis? Carry out the contrast and interpret the results obtained.*

A

Formulas for the exam

Probability distributions

X	$f(x)$ or $p(x)$	$E(X)$	$Var(X)$
$B(n,p)$	$\binom{n}{k}p^k q^{n-k}$	np	npq
$P(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
$H(N,M,n)$	$\frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$	np	$npq\left(\frac{N-n}{N-1}\right)$
$G(p)$	q^xp	$\frac{q}{p}$	$\frac{q}{p^2}$
$U(a,b)$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
$Exp(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$N(\mu,\sigma)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2
$\Gamma(r,\lambda)$	$\frac{\Lambda}{\Gamma(r)}\left(\lambda x\right)^{r-1}e^{-\lambda x}$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
$\beta(\alpha,\beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
t(n)	$\frac{1}{\beta(p,q)}\frac{1}{\sqrt{n}}\frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$	0	$\frac{n}{n-2}$

χ^2_n is $\Gamma(\frac{n}{2},\frac{1}{2})$.

ANOVA

	SS	df	MS	F
Between Groups	$\sum_{i=1}^{k} n_i(\overline{y}_{i,\cdot} - \overline{y}_{\cdot,\cdot})^2$	$k-1$	$\frac{SS_B}{k-1}$	$\frac{MS_B}{MS_E}$
Error	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i,j} - \overline{y}_{i,\cdot})^2$	$n-k$	$\frac{SS_E}{n-k}$	
Sum	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i,j} - \overline{y}_{\cdot,\cdot})^2$	$n-1$	$\frac{SS_T}{n-1}$	

Correction of Yates

- When $|o_{11}o_{22} - o_{12}o_{21}| \leq n/2$, $\chi^2_p = \frac{n(o_{11}o_{22}-o_{12}o_{21})^2}{z_{\cdot 1}\cdot z_{\cdot 2}\cdot f_{1\cdot}\cdot f_{2\cdot}}$
- When $|o_{11}o_{22} - o_{12}o_{21}| > n/2$, $\chi^2_Y = \frac{n(|o_{11}o_{22}-o_{12}o_{21}|-\frac{n}{2})^2}{z_{\cdot 1}\cdot z_{\cdot 2}\cdot f_{1\cdot}\cdot f_{2\cdot}}$

Non parametric tests

Wilcoxon test of ranks

$n \geq 15$, $t_+ \sim N\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$

Wilcoxon test of sum of ranks (Mann-Whitney)

$n_1 \geq 10$ and $n_2 \geq 10$, $w_1 \sim N\left(\frac{n_1(n_1+n_2+1)}{2}, \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}\right)$

Kruskal-Wallis

$H = \frac{12V}{n(n+1)} = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{r_i^2}{n_i} - 3(n+1) \sim \chi^2_{\alpha;k-1}$

B

Table for HT and CIs

In the next page, we present the table that we will use to compute CIs and to solve HTs.

Parameter	Conditions	Distribution	$(1-\alpha)\cdot 100\%$ CI	Pivot
μ	σ known Normality (or $n \geq 30$)	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$	$\left(\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$	$z_p = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
μ	σ unknown Normality and $n < 30$	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx t_{n-1}$	$\left(\bar{x} \mp t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}\right)$	$t_p = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
μ	σ unknown Normality y $n \geq 30$	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1)$	$\left(\bar{x} \mp z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$	$z_p = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
σ^2	Normality	$\frac{(n-1)s^2}{\sigma^2} \approx \chi^2_{n-1}$	$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}\right)$	$\chi^2_p = \frac{(n-1)s^2}{\sigma_0^2}$
$\mu_1-\mu_2$	σ_1 y σ_2 known Normality (or $n_1 \geq 30$ and $n_2 \geq 30$)	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \approx N(0,1)$	$\left(\bar{x}_1-\bar{x}_2 \mp z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	$z_p = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
$\mu_1-\mu_2$	σ_1 y σ_2 unknown Normality and $\sigma_1=\sigma_2$ $n_1<30$ or $n_2<30$	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \approx t_{n_1+n_2-2}$	$\left(\bar{x}_1-\bar{x}_2 \mp t_{\frac{\alpha}{2};n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$	$t_p = \frac{\bar{X}_1-\bar{X}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
$\mu_1-\mu_2$	σ_1 y σ_2 unknown Normality and $\sigma_1\neq\sigma_2$ $n_1<30$ or $n_2<30$	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \approx t_{df}$	$\left(\bar{x}_1-\bar{x}_2 \mp t_{\frac{\alpha}{2};df}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$	$t_p = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
$\mu_1-\mu_2$	σ_1 y σ_2 unknown Normality and $n_1 \geq 30, n_2 \geq 30$	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \approx N(0,1)$	$\left(\bar{x}_1-\bar{x}_2 \mp z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$	$z_p = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
$\frac{\sigma_1^2}{\sigma_2^2}$	Normality	$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \approx F_{(n_1-1),(n_2-1)}$	$\left(\frac{s_1^2/s_2^2}{F_{\frac{\alpha}{2};(n_1-1),(n_2-1)}}, \frac{s_1^2/s_2^2}{F_{1-\frac{\alpha}{2};(n_1-1),(n_2-1)}}\right)$	$F_p = \frac{s_1^2}{s_2^2}$
p	Bin(1,p) $np^*>5$ y $nq^*>5$	$\frac{p^*-p}{\sqrt{pq/n}} \approx N(0,1)$	$\left(p^* \mp z_{\frac{\alpha}{2}}\sqrt{\frac{p^*q^*}{n}}\right)$	$z_p = \frac{p^*-p_0}{\sqrt{\frac{p_0q_0}{n}}}$
p_1-p_2	Bin(1,p) $n_1p_1^*>5$,$n_1q_1^*>5$ $n_2p_2^*>5$,$n_2q_2^*>5$	$\frac{p_1^*-p_2^*-(p_1-p_2)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \approx N(0,1)$	$\left(p_1^*-p_2^* \mp z_{\frac{\alpha}{2}}\sqrt{\frac{p_1^*q_1^*}{n_1}+\frac{p_2^*q_2^*}{n_2}}\right)$	$z_p = \frac{p_1^*-p_2^*}{\sqrt{\frac{p_1^*q_1^*}{n_1}+\frac{p_2^*q_2^*}{n_2}}}$

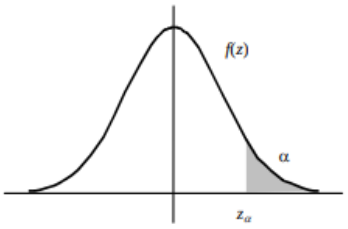
$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \qquad df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

C

Tables of Distributions

C.1 Standardized Normal Distribution: Z=$\mathcal{N}(0,1)$

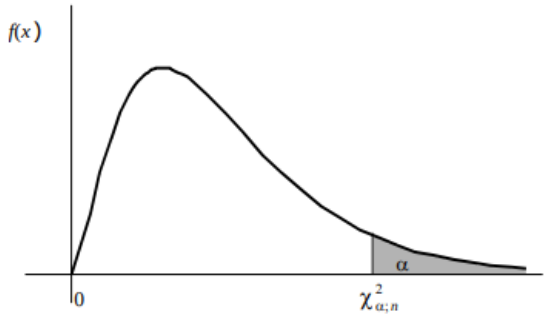
$$\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$



z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

Table C.1: Standarized Normal Distribution

C.2 Chi Squared Distribution: χ^2_n

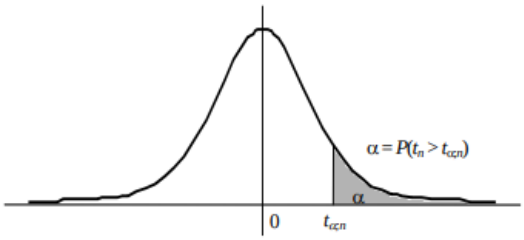


α / n	0,995	0,99	0,98	0,975	0,95	0,90	0,10	0,05	0,025	0,02	0,01
1	0,000	0,000	0,001	0,001	0,004	0,016	2,706	3,841	5,024	5,412	6,635
2	0,010	0,020	0,040	0,051	0,103	0,211	4,605	5,991	7,378	7,824	9,210
3	0,072	0,115	0,185	0,216	0,352	0,584	6,251	7,815	9,348	9,837	11,345
4	0,207	0,297	0,429	0,484	0,711	1,064	7,779	9,488	11,143	11,668	13,277
5	0,412	0,554	0,752	0,831	1,145	1,610	9,236	11,070	12,832	13,388	15,086
6	0,676	0,872	1,134	1,237	1,635	2,204	10,645	12,592	14,449	15,033	16,812
7	0,989	1,239	1,564	1,690	2,167	2,833	12,017	14,067	16,013	16,622	18,475
8	1,344	1,647	2,032	2,180	2,733	3,490	13,362	15,507	17,535	18,168	20,090
9	1,735	2,088	2,532	2,700	3,325	4,168	14,684	16,919	19,023	19,679	21,666
10	2,156	2,558	3,059	3,247	3,940	4,865	15,987	18,307	20,483	21,161	23,209
11	2,603	3,053	3,609	3,816	4,575	5,578	17,275	19,675	21,920	22,618	24,725
12	3,074	3,571	4,178	4,404	5,226	6,304	18,549	21,026	23,337	24,054	26,217
13	3,565	4,107	4,765	5,009	5,892	7,041	19,812	22,362	24,736	25,471	27,688
14	4,075	4,660	5,368	5,629	6,571	7,790	21,064	23,685	26,119	26,873	29,141
15	4,601	5,229	5,985	6,262	7,261	8,547	22,307	24,996	27,488	28,259	30,578
16	5,142	5,812	6,614	6,908	7,962	9,312	23,542	26,296	28,845	29,633	32,000
17	5,697	6,408	7,255	7,564	8,672	10,085	24,769	27,587	30,191	30,995	33,409
18	6,265	7,015	7,906	8,231	9,390	10,865	25,989	28,869	31,526	32,346	34,805
19	6,844	7,633	8,567	8,907	10,117	11,651	27,204	30,144	32,852	33,687	36,191
20	7,434	8,260	9,237	9,591	10,851	12,443	28,412	31,410	34,170	35,020	37,566
21	8,034	8,897	9,915	10,283	11,591	13,240	29,615	32,671	35,479	36,343	38,932
22	8,643	9,542	10,600	10,982	12,338	14,041	30,813	33,924	36,781	37,659	40,289
23	9,260	10,196	11,293	11,689	13,091	14,848	32,007	35,172	38,076	38,968	41,638
24	9,886	10,856	11,992	12,401	13,848	15,659	33,196	36,415	39,364	40,270	42,980
25	10,520	11,524	12,697	13,120	14,611	16,473	34,382	37,652	40,646	41,566	44,314
26	11,160	12,198	13,409	13,844	15,379	17,292	35,563	38,885	41,923	42,856	45,642
27	11,808	12,878	14,125	14,573	16,151	18,114	36,741	40,113	43,195	44,140	46,963
28	12,461	13,565	14,847	15,308	16,928	18,939	37,916	41,337	44,461	45,419	48,278
29	13,121	14,256	15,574	16,047	17,708	19,768	39,087	42,557	45,722	46,693	49,588
30	13,787	14,953	16,306	16,791	18,493	20,599	40,256	43,773	46,979	47,962	50,892

Table C.2: Chi Squared Distribution

$$n > 30 \Rightarrow \chi^2_{\alpha,n} = \frac{1}{2}\left(z_\alpha + \sqrt{2n-1}\right)^2$$

C.3 Student's t Distribution: t_n



α / n	0,40	0,3	0,2	0,1	0,05	0,025	0,01	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,706	31,821	63,656	318,289	636,578
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,328	31,600
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,214	12,924
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,894	6,869
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,689
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,660
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,526	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,416
100	0,254	0,526	0,845	1,290	1,660	1,984	2,364	2,626	3,174	3,390
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,340
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,107	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,327	2,576	3,091	3,291

Table C.3: Student's t Distribution

C.4 Binomial Distribution: Bin(n,p)

$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}$

n	k	p	0,01	0,05	0,10	0,15	0,20	0,25	0,30	1 / 3	0,35	0,40	0,45	0,49	0,50
2	0		0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4444	0,4225	0,3600	0,3025	0,2601	0,2500
	1		0,0198	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4444	0,4550	0,4800	0,4950	0,4998	0,5000
	2		0,0001	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1111	0,1225	0,1600	0,2025	0,2401	0,2500
3	0		0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2963	0,2746	0,2160	0,1664	0,1327	0,1250
	1		0,0294	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4444	0,4436	0,4320	0,4084	0,3823	0,3750
	2		0,0003	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2222	0,2389	0,2880	0,3341	0,3674	0,3750
	3		0,0000	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0370	0,0429	0,0640	0,0911	0,1176	0,1250
4	0		0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1975	0,1785	0,1296	0,0915	0,0677	0,0625
	1		0,0388	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3951	0,3845	0,3456	0,2995	0,2600	0,2500
	2		0,0006	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,2963	0,3105	0,3456	0,3675	0,3747	0,3750
	3		0,0000	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,0988	0,1115	0,1536	0,2005	0,2400	0,2500
	4		0,0000	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0123	0,0150	0,0256	0,0410	0,0576	0,0625
5	0		0,9510	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1317	0,1160	0,0778	0,0503	0,0345	0,0313
	1		0,0480	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3292	0,3124	0,2592	0,2059	0,1657	0,1563
	2		0,0010	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3292	0,3364	0,3456	0,3369	0,3185	0,3125
	3		0,0000	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1646	0,1811	0,2304	0,2757	0,3060	0,3125
	4		0,0000	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0412	0,0488	0,0768	0,1128	0,1470	0,1563
	5		0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0041	0,0053	0,0102	0,0185	0,0282	0,0313
6	0		0,9415	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0878	0,0754	0,0467	0,0277	0,0176	0,0156
	1		0,0571	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2634	0,2437	0,1866	0,1359	0,1014	0,0938
	2		0,0014	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3292	0,3280	0,3110	0,2780	0,2436	0,2344
	3		0,0000	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2195	0,2355	0,2765	0,3032	0,3121	0,3125
	4		0,0000	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0823	0,0951	0,1382	0,1861	0,2249	0,2344
	5		0,0000	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0165	0,0205	0,0369	0,0609	0,0864	0,0938
	6		0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0014	0,0018	0,0041	0,0083	0,0138	0,0156
7	0		0,9321	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0585	0,0490	0,0280	0,0152	0,0090	0,0078
	1		0,0659	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,2048	0,1848	0,1306	0,0872	0,0604	0,0547
	2		0,0020	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,3073	0,2985	0,2613	0,2140	0,1740	0,1641
	3		0,0000	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2561	0,2679	0,2903	0,2918	0,2786	0,2734
	4		0,0000	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1280	0,1442	0,1935	0,2388	0,2676	0,2734
	5		0,0000	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0384	0,0466	0,0774	0,1172	0,1543	0,1641
	6		0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0064	0,0084	0,0172	0,0320	0,0494	0,0547
	7		0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0006	0,0016	0,0037	0,0068	0,0078
8	0		0,9227	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0390	0,0319	0,0168	0,0084	0,0046	0,0039
	1		0,0746	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1561	0,1373	0,0896	0,0548	0,0352	0,0313
	2		0,0026	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2731	0,2587	0,2090	0,1569	0,1183	0,1094
	3		0,0001	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2731	0,2786	0,2787	0,2568	0,2273	0,2188
	4		0,0000	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1707	0,1875	0,2322	0,2627	0,2730	0,2734
	5		0,0000	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0683	0,0808	0,1239	0,1719	0,2098	0,2188
	6		0,0000	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0171	0,0217	0,0413	0,0703	0,1008	0,1094
	7		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0024	0,0033	0,0079	0,0164	0,0277	0,0313
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0002	0,0007	0,0017	0,0033	0,0039
9	0		0,9135	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0260	0,0207	0,0101	0,0046	0,0023	0,0020
	1		0,0830	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1171	0,1004	0,0605	0,0339	0,0202	0,0176
	2		0,0034	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2341	0,2162	0,1612	0,1110	0,0776	0,0703
	3		0,0001	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2731	0,2716	0,2508	0,2119	0,1739	0,1641
	4		0,0000	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2048	0,2194	0,2508	0,2600	0,2506	0,2461
	5		0,0000	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1024	0,1181	0,1672	0,2128	0,2408	0,2461
	6		0,0000	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0341	0,0424	0,0743	0,1160	0,1542	0,1641
	7		0,0000	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0073	0,0098	0,0212	0,0407	0,0635	0,0703
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0009	0,0013	0,0035	0,0083	0,0153	0,0176
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0008	0,0016	0,0020
10	0		0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0173	0,0135	0,0060	0,0025	0,0012	0,0010
	1		0,0914	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0867	0,0725	0,0403	0,0207	0,0114	0,0098
	2		0,0042	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1951	0,1757	0,1209	0,0763	0,0494	0,0439
	3		0,0001	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2601	0,2522	0,2150	0,1665	0,1267	0,1172
	4		0,0000	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2276	0,2377	0,2508	0,2384	0,2130	0,2051
	5		0,0000	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1366	0,1536	0,2007	0,2340	0,2456	0,2461
	6		0,0000	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0569	0,0689	0,1115	0,1596	0,1966	0,2051
	7		0,0000	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0163	0,0212	0,0425	0,0746	0,1080	0,1172
	8		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0030	0,0043	0,0106	0,0229	0,0389	0,0439
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005	0,0016	0,0042	0,0083	0,0098
	10		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0010

Table C.4: Binomial Distribution

C.5 Poisson Distribution: $\mathcal{P}(\lambda)$

$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$

λ \ k	0	1	2	3	4	5	6	7	8	9	10	11	12
0,1	0,9048	0,0905	0,0045	0,0002									
0,2	0,8187	0,1637	0,0164	0,0011	0,0001								
0,3	0,7408	0,2222	0,0333	0,0033	0,0003								
0,4	0,6703	0,2681	0,0536	0,0072	0,0007	0,0001							
0,5	0,6065	0,3033	0,0758	0,0126	0,0016	0,0002							
0,6	0,5488	0,3293	0,0988	0,0198	0,0030	0,0004							
0,7	0,4966	0,3476	0,1217	0,0284	0,0050	0,0007	0,0001						
0,8	0,4493	0,3595	0,1438	0,0383	0,0077	0,0012	0,0002						
0,9	0,4066	0,3659	0,1647	0,0494	0,0111	0,0020	0,0003						
1,0	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005	0,0001					
1,1	0,3329	0,3662	0,2014	0,0738	0,0203	0,0045	0,0008	0,0001					
1,2	0,3012	0,3614	0,2169	0,0867	0,0260	0,0062	0,0012	0,0002					
1,3	0,2725	0,3543	0,2303	0,0998	0,0324	0,0084	0,0018	0,0003	0,0001				
1,4	0,2466	0,3452	0,2417	0,1128	0,0395	0,0111	0,0026	0,0005	0,0001				
1,5	0,2231	0,3347	0,2510	0,1255	0,0471	0,0141	0,0035	0,0008	0,0001				
1,6	0,2019	0,3230	0,2584	0,1378	0,0551	0,0176	0,0047	0,0011	0,0002				
1,7	0,1827	0,3106	0,2640	0,1496	0,0636	0,0216	0,0061	0,0015	0,0003	0,0001			
1,8	0,1653	0,2975	0,2678	0,1607	0,0723	0,0260	0,0078	0,0020	0,0005	0,0001			
1,9	0,1496	0,2842	0,2700	0,1710	0,0812	0,0309	0,0098	0,0027	0,0006	0,0001			
2,0	0,1353	0,2707	0,2707	0,1804	0,0902	0,0361	0,0120	0,0034	0,0009	0,0002			
2,2	0,1108	0,2438	0,2681	0,1966	0,1082	0,0476	0,0174	0,0055	0,0015	0,0004	0,0001		
2,4	0,0907	0,2177	0,2613	0,2090	0,1254	0,0602	0,0241	0,0083	0,0025	0,0007	0,0002		
2,6	0,0743	0,1931	0,2510	0,2176	0,1414	0,0735	0,0319	0,0118	0,0038	0,0011	0,0003	0,0001	
2,8	0,0608	0,1703	0,2384	0,2225	0,1557	0,0872	0,0407	0,0163	0,0057	0,0018	0,0005	0,0001	
3,0	0,0498	0,1494	0,2240	0,2240	0,1680	0,1008	0,0504	0,0216	0,0081	0,0027	0,0008	0,0002	0,0001
3,2	0,0408	0,1304	0,2087	0,2226	0,1781	0,1140	0,0608	0,0278	0,0111	0,0040	0,0013	0,0004	0,0001
3,4	0,0334	0,1135	0,1929	0,2186	0,1858	0,1264	0,0716	0,0348	0,0148	0,0056	0,0019	0,0006	0,0002
3,6	0,0273	0,0984	0,1771	0,2125	0,1912	0,1377	0,0826	0,0425	0,0191	0,0076	0,0028	0,0009	0,0003
3,8	0,0224	0,0850	0,1615	0,2046	0,1944	0,1477	0,0936	0,0508	0,0241	0,0102	0,0039	0,0013	0,0004
4,0	0,0183	0,0733	0,1465	0,1954	0,1954	0,1563	0,1042	0,0595	0,0298	0,0132	0,0053	0,0019	0,0006
5,0	0,0067	0,0337	0,0842	0,1404	0,1755	0,1755	0,1462	0,1044	0,0653	0,0363	0,0181	0,0082	0,0034
6,0	0,0025	0,0149	0,0446	0,0892	0,1339	0,1606	0,1606	0,1377	0,1033	0,0688	0,0413	0,0225	0,0113
7,0	0,0009	0,0064	0,0223	0,0521	0,0912	0,1277	0,1490	0,1490	0,1304	0,1014	0,0710	0,0452	0,0263
8,0	0,0003	0,0027	0,0107	0,0286	0,0573	0,0916	0,1221	0,1396	0,1396	0,1241	0,0993	0,0722	0,0481
9,0	0,0001	0,0011	0,0050	0,0150	0,0337	0,0607	0,0911	0,1171	0,1318	0,1318	0,1186	0,0970	0,0728
10,0	0,0000	0,0005	0,0023	0,0076	0,0189	0,0378	0,0631	0,0901	0,1126	0,1251	0,1251	0,1137	0,0948

λ \ k	13	14	15	16	17	18	19	20	21	22	23	24
3,6	0,0001											
3,8	0,0001											
4,0	0,0002	0,0001										
5,0	0,0013	0,0005	0,0002									
6,0	0,0052	0,0022	0,0009	0,0003	0,0001							
7,0	0,0142	0,0071	0,0033	0,0014	0,0006	0,0002	0,0001					
8,0	0,0296	0,0169	0,0090	0,0045	0,0021	0,0009	0,0004	0,0002	0,0001			
9,0	0,0504	0,0324	0,0194	0,0109	0,0058	0,0029	0,0014	0,0006	0,0003	0,0001		
10,0	0,0729	0,0521	0,0347	0,0217	0,0128	0,0071	0,0037	0,0019	0,0009	0,0004	0,0002	0,0001

Table C.5: Poisson Distribution

C.6 Fischer-Schenedor Distribution: F_{n_1,n_2}

$\mathbb{P}(F_{n_1,n_2} > F_{\alpha;n_1,n_2}) = \alpha$

n_1		1	2	3	4	5	6	8	10	12	24	∞
n_2	α											
1	0,01	4052,2	4999,5	5403,4	5624,6	5763,6	5859,0	5981,1	6055,8	6106,3	6234,6	6365,5
1	0,025	647,8	799,5	864,2	899,6	921,8	937,1	956,7	968,6	976,7	997,2	1018,2
1	0,05	161,45	199,50	215,71	224,58	230,16	233,99	238,88	241,88	243,91	249,05	254,30
1	0,10	39,86	49,50	53,59	55,83	57,24	58,20	59,44	60,19	60,71	62,00	63,32
1	0,20	9,47	12,00	13,06	13,64	14,01	14,26	14,58	14,77	14,90	15,24	15,58
2	0,01	98,50	99,00	99,17	99,25	99,30	99,33	99,37	99,40	99,42	99,46	99,50
2	0,025	38,51	39,00	39,17	39,25	39,30	39,33	39,37	39,40	39,41	39,46	39,50
2	0,05	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,40	19,41	19,45	19,50
2	0,10	8,53	9,00	9,16	9,24	9,29	9,33	9,37	9,39	9,41	9,45	9,49
2	0,20	3,56	4,00	4,16	4,24	4,28	4,32	4,36	4,38	4,40	4,44	4,48
3	0,01	34,12	30,82	29,46	28,71	28,24	27,91	27,49	27,23	27,05	26,60	26,13
3	0,025	17,44	16,04	15,44	15,10	14,88	14,73	14,54	14,42	14,34	14,12	13,90
3	0,05	10,13	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,74	8,64	8,53
3	0,10	5,54	5,46	5,39	5,34	5,31	5,28	5,25	5,23	5,22	5,18	5,13
3	0,20	2,68	2,89	2,94	2,96	2,97	2,97	2,98	2,98	2,98	2,98	2,98
4	0,01	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,55	14,37	13,93	13,46
4	0,025	12,22	10,65	9,98	9,60	9,36	9,20	8,98	8,84	8,75	8,51	8,26
4	0,05	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,91	5,77	5,63
4	0,10	4,54	4,32	4,19	4,11	4,05	4,01	3,95	3,92	3,90	3,83	3,76
4	0,20	2,35	2,47	2,48	2,48	2,48	2,47	2,47	2,46	2,46	2,44	2,43
5	0,01	16,26	13,27	12,06	11,39	10,97	10,67	10,29	10,05	9,89	9,47	9,02
5	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,52	6,28	6,02
5	0,05	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,68	4,53	4,37
5	0,10	4,06	3,78	3,62	3,52	3,45	3,40	3,34	3,30	3,27	3,19	3,11
5	0,20	2,18	2,26	2,25	2,24	2,23	2,22	2,20	2,19	2,18	2,16	2,13
6	0,01	13,75	10,92	9,78	9,15	8,75	8,47	8,10	7,87	7,72	7,31	6,88
6	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,37	5,12	4,85
6	0,05	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	4,00	3,84	3,67
6	0,10	3,78	3,46	3,29	3,18	3,11	3,05	2,98	2,94	2,90	2,82	2,72
6	0,20	2,07	2,13	2,11	2,09	2,08	2,06	2,04	2,03	2,02	1,99	1,95
7	0,01	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,62	6,47	6,07	5,65
7	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,67	4,41	4,14
7	0,05	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,57	3,41	3,23
7	0,10	3,59	3,26	3,07	2,96	2,88	2,83	2,75	2,70	2,67	2,58	2,47
7	0,20	2,00	2,04	2,02	1,99	1,97	1,96	1,93	1,92	1,91	1,87	1,83
8	0,01	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,81	5,67	5,28	4,86
8	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,20	3,95	3,67
8	0,05	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,28	3,12	2,93
8	0,10	3,46	3,11	2,92	2,81	2,73	2,67	2,59	2,54	2,50	2,40	2,29
8	0,20	1,95	1,98	1,95	1,92	1,90	1,88	1,86	1,84	1,83	1,79	1,74
9	0,01	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,26	5,11	4,73	4,31
9	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,87	3,61	3,33
9	0,05	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,07	2,90	2,71
9	0,10	3,36	3,01	2,81	2,69	2,61	2,55	2,47	2,42	2,38	2,28	2,16
9	0,20	1,91	1,93	1,90	1,87	1,85	1,83	1,80	1,78	1,76	1,72	1,67
10	0,01	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,85	4,71	4,33	3,91
10	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,62	3,37	3,08
10	0,05	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,91	2,74	2,54
10	0,10	3,29	2,92	2,73	2,61	2,52	2,46	2,38	2,32	2,28	2,18	2,06
10	0,20	1,88	1,90	1,86	1,83	1,80	1,78	1,75	1,73	1,72	1,67	1,62

Table C.6: Fischer-Schenedor Distribution: Part 1

n_2	n_1 / α	1	2	3	4	5	6	8	10	12	24	∞
11	0,01	9,65	7,21	6,22	5,67	5,32	5,07	4,74	4,54	4,40	4,02	3,60
11	0,025	6,72	5,26	4,63	4,28	4,04	3,88	3,66	3,53	3,43	3,17	2,88
11	0,05	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,85	2,79	2,61	2,40
11	0,10	3,23	2,86	2,66	2,54	2,45	2,39	2,30	2,25	2,21	2,10	1,97
11	0,20	1,86	1,87	1,83	1,80	1,77	1,75	1,72	1,69	1,68	1,63	1,57
12	0,01	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,30	4,16	3,78	3,36
12	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,51	3,37	3,28	3,02	2,72
12	0,05	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,69	2,51	2,30
12	0,10	3,18	2,81	2,61	2,48	2,39	2,33	2,24	2,19	2,15	2,04	1,90
12	0,20	1,84	1,85	1,80	1,77	1,74	1,72	1,69	1,66	1,65	1,60	1,54
13	0,01	9,07	6,70	5,74	5,21	4,86	4,62	4,30	4,10	3,96	3,59	3,17
13	0,025	6,41	4,97	4,35	4,00	3,77	3,60	3,39	3,25	3,15	2,89	2,60
13	0,05	4,67	3,81	3,41	3,18	3,03	2,92	2,77	2,67	2,60	2,42	2,21
13	0,10	3,14	2,76	2,56	2,43	2,35	2,28	2,20	2,14	2,10	1,98	1,85
13	0,20	1,82	1,83	1,78	1,75	1,72	1,69	1,66	1,64	1,62	1,57	1,51
14	0,01	8,86	6,51	5,56	5,04	4,69	4,46	4,14	3,94	3,80	3,43	3,00
14	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,29	3,15	3,05	2,79	2,49
14	0,05	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,53	2,35	2,13
14	0,10	3,10	2,73	2,52	2,39	2,31	2,24	2,15	2,10	2,05	1,94	1,80
14	0,20	1,81	1,81	1,76	1,73	1,70	1,67	1,64	1,62	1,60	1,55	1,48
15	0,01	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,80	3,67	3,29	2,87
15	0,025	6,20	4,77	4,15	3,80	3,58	3,41	3,20	3,06	2,96	2,70	2,40
15	0,05	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,54	2,48	2,29	2,07
15	0,10	3,07	2,70	2,49	2,36	2,27	2,21	2,12	2,06	2,02	1,90	1,76
15	0,20	1,80	1,80	1,75	1,71	1,68	1,66	1,62	1,60	1,58	1,53	1,46
16	0,01	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,69	3,55	3,18	2,75
16	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,12	2,99	2,89	2,63	2,32
16	0,05	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,42	2,24	2,01
16	0,10	3,05	2,67	2,46	2,33	2,24	2,18	2,09	2,03	1,99	1,87	1,72
16	0,20	1,79	1,78	1,74	1,70	1,67	1,64	1,61	1,58	1,56	1,51	1,43
17	0,01	8,40	6,11	5,19	4,67	4,34	4,10	3,79	3,59	3,46	3,08	2,65
17	0,025	6,04	4,62	4,01	3,66	3,44	3,28	3,06	2,92	2,82	2,56	2,25
17	0,05	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,45	2,38	2,19	1,96
17	0,10	3,03	2,64	2,44	2,31	2,22	2,15	2,06	2,00	1,96	1,84	1,69
17	0,20	1,78	1,77	1,72	1,68	1,65	1,63	1,59	1,57	1,55	1,49	1,42
18	0,01	8,29	6,01	5,09	4,58	4,25	4,01	3,71	3,51	3,37	3,00	2,57
18	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,01	2,87	2,77	2,50	2,19
18	0,05	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,34	2,15	1,92
18	0,10	3,01	2,62	2,42	2,29	2,20	2,13	2,04	1,98	1,93	1,81	1,66
18	0,20	1,77	1,76	1,71	1,67	1,64	1,62	1,58	1,55	1,53	1,48	1,40
19	0,01	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,43	3,30	2,92	2,49
19	0,025	5,92	4,51	3,90	3,56	3,33	3,17	2,96	2,82	2,72	2,45	2,13
19	0,05	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,38	2,31	2,11	1,88
19	0,10	2,99	2,61	2,40	2,27	2,18	2,11	2,02	1,96	1,91	1,79	1,63
19	0,20	1,76	1,75	1,70	1,66	1,63	1,61	1,57	1,54	1,52	1,46	1,39
20	0,01	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,37	3,23	2,86	2,42
20	0,025	5,87	4,46	3,86	3,51	3,29	3,13	2,91	2,77	2,68	2,41	2,09
20	0,05	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,28	2,08	1,84
20	0,10	2,97	2,59	2,38	2,25	2,16	2,09	2,00	1,94	1,89	1,77	1,61
20	0,20	1,76	1,75	1,70	1,65	1,62	1,60	1,56	1,53	1,51	1,45	1,37

Table C.7: Fischer-Schenedor Distribution: Part 2

n_1		1	2	3	4	5	6	8	10	12	24	∞
n_2	α											
21	0,01	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,31	3,17	2,80	2,36
21	0,025	5,83	4,42	3,82	3,48	3,25	3,09	2,87	2,73	2,64	2,37	2,04
21	0,05	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,32	2,25	2,05	1,81
21	0,10	2,96	2,57	2,36	2,23	2,14	2,08	1,98	1,92	1,87	1,75	1,59
21	0,20	1,75	1,74	1,69	1,65	1,61	1,59	1,55	1,52	1,50	1,44	1,36
23	0,01	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,21	3,07	2,70	2,26
23	0,025	5,75	4,35	3,75	3,41	3,18	3,02	2,81	2,67	2,57	2,30	1,97
23	0,05	4,28	3,42	3,03	2,80	2,64	2,53	2,37	2,27	2,20	2,01	1,76
23	0,10	2,94	2,55	2,34	2,21	2,11	2,05	1,95	1,89	1,84	1,72	1,55
23	0,20	1,74	1,73	1,68	1,63	1,60	1,57	1,53	1,51	1,49	1,42	1,34
25	0,01	7,77	5,57	4,68	4,18	3,85	3,63	3,32	3,13	2,99	2,62	2,17
25	0,025	5,69	4,29	3,69	3,35	3,13	2,97	2,75	2,61	2,51	2,24	1,91
25	0,05	4,24	3,39	2,99	2,76	2,60	2,49	2,34	2,24	2,16	1,96	1,71
25	0,10	2,92	2,53	2,32	2,18	2,09	2,02	1,93	1,87	1,82	1,69	1,52
25	0,20	1,73	1,72	1,66	1,62	1,59	1,56	1,52	1,49	1,47	1,41	1,32
27	0,01	7,68	5,49	4,60	4,11	3,78	3,56	3,26	3,06	2,93	2,55	2,10
27	0,025	5,63	4,24	3,65	3,31	3,08	2,92	2,71	2,57	2,47	2,19	1,85
27	0,05	4,21	3,35	2,96	2,73	2,57	2,46	2,31	2,20	2,13	1,93	1,67
27	0,10	2,90	2,51	2,30	2,17	2,07	2,00	1,91	1,85	1,80	1,67	1,49
27	0,20	1,73	1,71	1,66	1,61	1,58	1,55	1,51	1,48	1,46	1,40	1,30
29	0,01	7,60	5,42	4,54	4,04	3,73	3,50	3,20	3,00	2,87	2,49	2,03
29	0,025	5,59	4,20	3,61	3,27	3,04	2,88	2,67	2,53	2,43	2,15	1,81
29	0,05	4,18	3,33	2,93	2,70	2,55	2,43	2,28	2,18	2,10	1,90	1,64
29	0,10	2,89	2,50	2,28	2,15	2,06	1,99	1,89	1,83	1,78	1,65	1,47
29	0,20	1,72	1,70	1,65	1,60	1,57	1,54	1,50	1,47	1,45	1,39	1,29
30	0,01	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,98	2,84	2,47	2,01
30	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,65	2,51	2,41	2,14	1,79
30	0,05	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	2,09	1,89	1,62
30	0,10	2,88	2,49	2,28	2,14	2,05	1,98	1,88	1,82	1,77	1,64	1,46
30	0,20	1,72	1,70	1,64	1,60	1,57	1,54	1,50	1,47	1,45	1,38	1,28
40	0,01	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,80	2,66	2,29	1,80
40	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,53	2,39	2,29	2,01	1,64
40	0,05	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08	2,00	1,79	1,51
40	0,10	2,84	2,44	2,23	2,09	2,00	1,93	1,83	1,76	1,71	1,57	1,38
40	0,20	1,70	1,68	1,62	1,57	1,54	1,51	1,47	1,44	1,41	1,34	1,24
60	0,01	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,63	2,50	2,12	1,60
60	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,41	2,27	2,17	1,88	1,48
60	0,05	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99	1,92	1,70	1,39
60	0,10	2,79	2,39	2,18	2,04	1,95	1,87	1,77	1,71	1,66	1,51	1,29
60	0,20	1,68	1,65	1,59	1,55	1,51	1,48	1,44	1,41	1,38	1,31	1,18
120	0,01	6,85	4,79	3,95	3,48	3,17	2,96	2,66	2,47	2,34	1,95	1,38
120	0,025	5,15	3,80	3,23	2,89	2,67	2,52	2,30	2,16	2,05	1,76	1,31
120	0,05	3,92	3,07	2,68	2,45	2,29	2,18	2,02	1,91	1,83	1,61	1,25
120	0,10	2,75	2,35	2,13	1,99	1,90	1,82	1,72	1,65	1,60	1,45	1,19
120	0,20	1,66	1,63	1,57	1,52	1,48	1,45	1,41	1,37	1,35	1,27	1,12
∞	0,01	6,64	4,61	3,78	3,32	3,02	2,80	2,51	2,32	2,18	1,79	1,01
∞	0,025	5,02	3,69	3,12	2,79	2,57	2,41	2,19	2,05	1,94	1,64	1,01
∞	0,05	3,84	3,00	2,60	2,37	2,21	2,10	1,94	1,83	1,75	1,52	1,01
∞	0,10	2,71	2,30	2,08	1,94	1,85	1,77	1,67	1,60	1,55	1,38	1,01
∞	0,20	1,64	1,61	1,55	1,50	1,46	1,43	1,38	1,34	1,32	1,23	1,00

Table C.8: Fischer-Schenedor Distribution: Part 3

C.7 Distribution of the Statistic of the Kolmogorov-Smirnov Test:

$\mathbb{P}(D_n > D_{\alpha;n}) = \alpha$

	(α)				
n	0,2	0,15	0,10	0,05	0,01
1	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
>35	$\frac{1,07}{\sqrt{n}}$	$\frac{1,14}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

C.8 Distribution of the Statistic of the Kolmogorov-Smirnov Test (Normality):

$\mathbb{P}(D_n > D_{\alpha;n}) = \alpha$

	(α)				
n	0,2	0,15	0,10	0,05	0,01
4	0,300	0,319	0,352	0,381	0,417
5	0,285	0,299	0,315	0,337	0,405
6	0,265	0,277	0,294	0,319	0,364
7	0,247	0,258	0,276	0,300	0,348
8	0,233	0,244	0,261	0,285	0,331
9	0,223	0,233	0,249	0,271	0,311
10	0,215	0,224	0,239	0,258	0,294
11	0,206	0,217	0,230	0,249	0,284
12	0,199	0,212	0,223	0,242	0,275
13	0,190	0,202	0,214	0,234	0,268
14	0,183	0,194	0,207	0,227	0,261
15	0,177	0,187	0,201	0,220	0,257
16	0,173	0,182	0,195	0,213	0,250
17	0,169	0,177	0,189	0,206	0,245
18	0,166	0,173	0,184	0,200	0,239
19	0,163	0,169	0,179	0,195	0,235
20	0,160	0,166	0,174	0,190	0,231
25	0,149	0,153	0,165	0,180	0,203
30	0,131	0,136	0,144	0,161	0,187
> 35	$\frac{0,736}{\sqrt{n}}$	$\frac{0,768}{\sqrt{n}}$	$\frac{0,805}{\sqrt{n}}$	$\frac{0,886}{\sqrt{n}}$	$\frac{1,031}{\sqrt{n}}$

C.9 Wilcoxon Test of Signed Ranks

When $t \geq \frac{n(n+1)}{4}$

n	t	$P(T \geq t)$	n	t	$P(T \geq t)$	n	t	$P(T \geq t)$	n	t	$P(T \geq t)$
2	2	0,500	**8**	18	0,527		34	0,278		59	0,009
	3	0,250		19	0,473		35	0,246		60	0,007
3	3	0,625		20	0,422		36	0,216		61	0,005
	4	0,375		21	0,371		37	0,188		62	0,003
	5	0,250		22	0,320		38	0,161		63	0,002
	6	0,125		23	0,273		39	0,138		64	0,001
4	5	0,562		24	0,230		40	0,116		65	0,001
	6	0,438		25	0,191		41	0,097		66	0,000
	7	0,312		26	0,156		42	0,080	**12**	39	0,515
	8	0,188		27	0,125		43	0,065		40	0,485
	9	0,125		28	0,098		44	0,053		41	0,455
	10	0,062		29	0,074		45	0,042		42	0,425
5	8	0,500		30	0,055		46	0,032		43	0,396
	9	0,406		31	0,039		47	0,024		44	0,367
	10	0,312		32	0,027		48	0,019		45	0,339
	11	0,219		33	0,020		49	0,014		46	0,311
	12	0,156		34	0,012		50	0,010		47	0,285
	13	0,094		35	0,008		51	0,007		48	0,259
	14	0,062		36	0,004		52	0,005		49	0,235
	15	0,031		23	0,500		53	0,003		50	0,212
	11	0,500	**9**	24	0,455		54	0,002		51	0,190
6	12	0,422		25	0,410		55	0,001		52	0,170
	13	0,344		26	0,367		33	0,517		53	0,151
	14	0,281		27	0,326	**11**	34	0,483		54	0,133
	15	0,219		28	0,285		35	0,449		55	0,117
	16	0,156		29	0,248		36	0,416		56	0,102
	17	0,109		30	0,213		37	0,382		57	0,088
	18	0,078		31	0,180		38	0,350		58	0,076
	19	0,047		32	0,150		39	0,319		59	0,065
	20	0,031		33	0,125		40	0,289		60	0,055
	21	0,016		34	0,102		41	0,260		61	0,046
	14	0,531		35	0,082		42	0,232		62	0,039
7	15	0,469		36	0,064		43	0,207		63	0,032
	16	0,406		37	0,049		44	0,183		64	0,026
	17	0,344		38	0,037		45	0,160		65	0,021
	18	0,289		39	0,027		46	0,139		66	0,017
	19	0,234		40	0,020		47	0,120		67	0,013
	20	0,188		41	0,014		48	0,103		68	0,010
	21	0,148		42	0,010		49	0,087		69	0,008
	22	0,109		43	0,006		50	0,074		70	0,006
	23	0,078		44	0,004		51	0,062		71	0,005
	24	0,055		45	0,002		52	0,051		72	0,003
	25	0,039		28	0,500		53	0,042		73	0,002
	26	0,023	**10**	29	0,461		54	0,034		74	0,002
	27	0,016		30	0,423		55	0,027		75	0,001
	28	0,008		31	0,385		56	0,021		76	0,001
				32	0,348		57	0,016		77	0,000
				33	0,312		58	0,012		78	0,000

When $t \geq \frac{n(n+1)}{4}$

n	t	$P(T \geq t)$	n	t	$P(T \geq t)$	n	t	$P(T \geq t)$	n	t	$P(T \geq t)$
13	46	0,500		87	0,001		88	0,012		82	0,115
	47	0,473		88	0,001		89	0,010		83	0,104
	48	0,446		89	0,000		90	0,008		84	0,094
	49	0,420		90	0,000		91	0,007		85	0,084
	50	0,393		91	0,000		92	0,005		86	0,076
	51	0,368	**14**	53	0,500		93	0,004		87	0,068
	52	0,342		54	0,476		94	0,003		88	0,060
	53	0,318		55	0,452		95	0,003		89	0,053
	54	0,294		56	0,428		96	0,002		90	0,047
	55	0,271		57	0,404		97	0,002		91	0,042
	56	0,249		58	0,380		98	0,001		92	0,036
	57	0,227		59	0,357		99	0,001		93	0,032
	58	0,207		60	0,335		100	0,001		94	0,028
	59	0,188		61	0,313		101	0,000		95	0,024
	60	0,170		62	0,292		102	0,000		96	0,021
	61	0,153		63	0,271		103	0,000		97	0,018
	62	0,137		64	0,251		104	0,000		98	0,015
	63	0,122		65	0,232		105	0,000		99	0,013
	64	0,108		66	0,213	**15**	60	0,511		100	0,011
	65	0,095		67	0,196		61	0,489		101	0,009
	66	0,084		68	0,179		62	0,467		102	0,008
	67	0,073		69	0,163		63	0,445		103	0,006
	68	0,064		70	0,148		64	0,423		104	0,005
	69	0,055		71	0,134		65	0,402		105	0,004
	70	0,047		72	0,121		66	0,381		106	0,003
	71	0,040		73	0,108		67	0,360		107	0,003
	72	0,034		74	0,097		68	0,339		108	0,002
	73	0,029		75	0,086		69	0,319		109	0,002
	74	0,024		76	0,077		70	0,300		110	0,001
	75	0,020		77	0,068		71	0,281		111	0,001
	76	0,016		78	0,059		72	0,262		112	0,001
	77	0,013		79	0,052		73	0,244		113	0,001
	78	0,011		80	0,045		74	0,227		114	0,000
	79	0,009		81	0,039		75	0,211		115	0,000
	80	0,007		82	0,034		76	0,195		116	0,000
	81	0,005		83	0,029		77	0,180		117	0,000
	82	0,004		84	0,025		78	0,165		118	0,000
	83	0,003		85	0,021		79	0,151		119	0,000
	84	0,002		86	0,018		80	0,138		120	0,000
	85	0,002		87	0,015		81	0,126			
	86	0,001									

Observation:

$$\mathbb{P}(T \leq t) = \mathbb{P}\left(T \geq \frac{n(n+1)}{2} - t\right)$$

C.10 Test of Mann-Whitney

$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 1$)			**($n_1 = 2$)**			**($n_1 = 2$)**
1	2	0,500	4	10	0,133	10	18	0,182
2	2	0,667		11	0,067		19	0,136
	3	0,333	5	8	0,571		20	0,091
3	3	0,500		9	0,429		21	0,061
	4	0,250		10	0,286		22	0,030
4	3	0,600		11	0,190		23	0,015
	4	0,400		12	0,095			
	5	0,200		13	0,048			**($n_1 = 3$)**
5	4	0,500	6	9	0,571	3	11	0,500
	5	0,333		10	0,429		12	0,350
	6	0,167		11	0,321		13	0,200
6	4	0,571		12	0,214		14	0,100
	5	0,429		13	0,143		15	0,050
	6	0,286		14	0,071	4	12	0,571
	7	0,143		15	0,036		13	0,429
7	5	0,500	7	10	0,556		14	0,314
	6	0,375		11	0,444		15	0,200
	7	0,250		12	0,333		16	0,114
	8	0,125		13	0,250		17	0,057
8	5	0,556		14	0,167		18	0,029
	6	0,444		15	0,111	5	14	0,500
	7	0,333		16	0,056		15	0,393
	8	0,222		17	0,028		16	0,286
	9	0,111	8	11	0,556		17	0,196
9	6	0,500		12	0,444		18	0,125
	7	0,400		13	0,356		19	0,071
	8	0,300		14	0,267		20	0,036
	9	0,200		15	0,200		21	0,018
	10	0,100		16	0,133	6	15	0,548
10	6	0,545		17	0,089		16	0,452
	7	0,455		18	0,044		17	0,357
	8	0,364		19	0,022		18	0,274
	9	0,273	9	12	0,545		19	0,190
	10	0,182		13	0,455		20	0,131
	11	0,091		14	0,364		21	0,083
				15	0,291		22	0,048
		($n_1 = 2$)		16	0,218		23	0,024
2	5	0,667		17	0,164		24	0,012
	6	0,333		18	0,109	7	17	0,500
	7	0,167		19	0,073		18	0,417
3	6	0,600		20	0,036		19	0,333
	7	0,400		21	0,018		20	0,258
	8	0,200	10	13	0,545		21	0,192
	9	0,100		14	0,455		22	0,133
4	7	0,600		15	0,379		23	0,092
	8	0,400		16	0,303		24	0,058
	9	0,267		17	0,242		25	0,033

$$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 3$)			**($n_1 = 4$)**			**($n_1 = 4$)**
7	26	0,017	4	18	0,557	7	38	0,003
	27	0,008		19	0,443	8	26	0,533
8	18	0,539		20	0,343		27	0,467
	19	0,461		21	0,243		28	0,404
	20	0,388		22	0,171		29	0,341
	21	0,315		23	0,100		30	0,285
	22	0,248		24	0,057		31	0,230
	23	0,188		25	0,029		32	0,184
	24	0,139		26	0,014		33	0,141
	25	0,097	5	20	0,548		34	0,107
	26	0,067		21	0,452		35	0,077
	27	0,042		22	0,365		36	0,055
	28	0,024		23	0,278		37	0,036
	29	0,012		24	0,206		38	0,024
	30	0,006		25	0,143		39	0,014
9	20	0,500		26	0,095		40	0,008
	21	0,432		27	0,056		41	0,004
	22	0,364		28	0,032		42	0,002
	23	0,300		29	0,016	9	28	0,530
	24	0,241		30	0,008		29	0,470
	25	0,186	6	22	0,543		30	0,413
	26	0,141		23	0,457		31	0,355
	27	0,105		24	0,381		32	0,302
	28	0,073		25	0,305		33	0,252
	29	0,050		26	0,238		34	0,207
	30	0,032		27	0,176		35	0,165
	31	0,018		28	0,129		36	0,130
	32	0,009		29	0,086		37	0,099
	33	0,005		30	0,057		38	0,074
10	21	0,531		31	0,033		39	0,053
	22	0,469		32	0,019		40	0,038
	23	0,406		33	0,010		41	0,025
	24	0,346		34	0,005		42	0,017
	25	0,287	7	24	0,536		43	0,010
	26	0,234		25	0,464		44	0,006
	27	0,185		26	0,394		45	0,003
	28	0,143		27	0,324		46	0,001
	29	0,108		28	0,264	10	30	0,527
	30	0,080		29	0,206		31	0,473
	31	0,056		30	0,158		32	0,420
	32	0,038		31	0,115		33	0,367
	33	0,024		32	0,082		34	0,318
	34	0,014		33	0,055		35	0,270
	35	0,007		34	0,036		36	0,227
	36	0,003		35	0,021		37	0,187
				36	0,012		38	0,152
				37	0,006		39	0,120

$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 4$)			**($n_1 = 5$)**			**($n_1 = 5$)**
10	40	0,094	7	38	0,216	9	51	0,041
	41	0,071		39	0,172		52	0,030
	42	0,053		40	0,134		53	0,021
	43	0,038		41	0,101		54	0,014
	44	0,0,027		42	0,074		55	0,009
	45	0,018		43	0,053		56	0,006
	46	0,012		44	0,037		57	0,003
	47	0,007		45	0,024		58	0,002
	48	0,004		46	0,015		59	0,001
	49	0,002		47	0,009		60	0,000
	50	0,001		48	0,005	10	40	0,523
				49	0,003		41	0,477
		($n_1 = 5$)		50	0,001		42	0,430
5	28	0,500	8	35	0,528		43	0,384
	29	0,421		36	0,472		44	0,339
	30	0,345		37	0,416		45	0,297
	31	0,274		38	0,362		46	0,257
	32	0,210		39	0,311		47	0,220
	33	0,155		40	0,262		48	0,185
	34	0,111		41	0,218		49	0,155
	35	0,075		42	0,177		50	0,127
	36	0,048		43	0,142		51	0,103
	37	0,028		44	0,111		52	0,082
	38	0,016		45	0,085		53	0,065
	39	0,008		46	0,064		54	0,050
	40	0,004		47	0,047		55	0,038
6	30	0,535		48	0,033		56	0,028
	31	0,465		49	0,023		57	0,020
	32	0,396		50	0,015		58	0,014
	33	0,331		51	0,009		59	0,010
	34	0,268		52	0,005		60	0,006
	35	0,214		53	0,003		61	0,004
	36	0,165		54	0,002		62	0,002
	37	0,123		55	0,001		63	0,001
	38	0,089	9	38	0,500		64	0,001
	39	0,063		39	0,449		65	0,000
	40	0,041		40	0,399			
	41	0,026		41	0,350			**($n_1 = 6$)**
	42	0,015		42	0,303	6	39	0,531
	43	0,009		43	0,259		40	0,469
	44	0,004		44	0,219		41	0,409
	45	0,002		45	0,182		42	0,350
7	33	0,500		46	0,149		43	0,294
	34	0,438		47	0,120		44	0,242
	35	0,378		48	0,095		45	0,197
	36	0,319		49	0,073		46	0,155
	37	0,265		50	0,056		47	0,120

$$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 6$)			**($n_1 = 6$)**			**($n_1 = 6$)**
6	48	0,090	8	60	0,030	10	60	0,184
	49	0,066		61	0,021		61	0,157
	50	0,047		62	0,015		62	0,132
	51	0,032		63	0,010		63	0,110
	52	0,021		64	0,006		64	0,090
	53	0,013		65	0,004		65	0,074
	54	0,008		66	0,002		66	0,059
	55	0,004		67	0,001		67	0,047
	56	0,002		68	0,001		68	0,036
	57	0,001		69	0,000		69	0,028
7	42	0,527	9	48	0,523		70	0,021
	43	0,473		49	0,477		71	0,016
	44	0,418		50	0,432		72	0,011
	45	0,365		51	0,388		73	0,008
	46	0,314		52	0,344		74	0,005
	47	0,267		53	0,303		75	0,004
	48	0,223		54	0,264		76	0,002
	49	0,183		55	0,228		77	0,001
	50	0,147		56	0,194		78	0,001
	51	0,117		57	0,164		79	0,000
	52	0,090		58	0,136		80	0,000
	53	0,069		59	0,112		81	0,000
	54	0,051		60	0,091			
	55	0,037		61	0,072			**($n_1 = 7$)**
	56	0,026		62	0,057	7	53	0,500
	57	0,017		63	0,044		54	0,451
	58	0,011		64	0,033		55	0,402
	59	0,007		65	0,025		56	0,355
	60	0,004		66	0,018		57	0,310
	61	0,002		67	0,013		58	0,267
	62	0,001		68	0,009		59	0,228
	63	0,001		69	0,006		60	0,191
8	45	0,525		70	0,004		61	0,159
	46	0,475		71	0,002		62	0,130
	47	0,426		72	0,001		63	0,104
	48	0,377		73	0,001		64	0,082
	49	0,331		74	0,000		65	0,064
	50	0,286		75	0,000		66	0,049
	51	0,245	10	51	0,521		67	0,036
	52	0,207		52	0,479		68	0,027
	53	0,172		53	0,437		69	0,019
	54	0,141		54	0,396		70	0,013
	55	0,114		55	0,356		71	0,009
	56	0,091		56	0,318		72	0,006
	57	0,071		57	0,281		73	0,003
	58	0,054		58	0,246		74	0,002
	59	0,041		59	0,214		75	0,001

$$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 7$)			**($n_1 = 7$)**			**($n_1 = 7$)**
7	76	0,001	9	76	0,045	10	94	0,001
	77	0,000		77	0,036		95	0,000
8	56	0,522		78	0,027		96	0,000
	57	0,478		79	0,021		97	0,000
	58	0,433		80	0,016		98	0,000
	59	0,389		81	0,011			
	60	0,347		82	0,008			**($n_1 = 8$)**
	61	0,306		83	0,006	8	68	0,520
	62	0,268		84	0,004		69	0,480
	63	0,232		85	0,003		70	0,439
	64	0,198		86	0,002		71	0,399
	65	0,168		87	0,001		72	0,360
	66	0,140		88	0,001		73	0,323
	67	0,116		89	0,000		74	0,287
	68	0,095		90	0,000		75	0,253
	69	0,076		91	0,000		76	0,221
	70	0,060	10	63	0,519		77	0,191
	71	0,047		64	0,481		78	0,164
	72	0,036		65	0,443		79	0,139
	73	0,027		66	0,406		80	0,117
	74	0,020		67	0,370		81	0,097
	75	0,014		68	0,335		82	0,080
	76	0,010		69	0,300		83	0,065
	77	0,007		70	0,268		84	0,052
	78	0,005		71	0,237		85	0,041
	79	0,003		72	0,209		86	0,032
	80	0,002		73	0,182		87	0,025
	81	0,001		74	0,157		88	0,019
	82	0,001		75	0,135		89	0,014
	83	0,000		76	0,115		90	0,010
	84	0,000		77	0,097		91	0,007
9	60	0,500		78	0,081		92	0,005
	61	0,459		79	0,067		93	0,003
	62	0,419		80	0,054		94	0,002
	63	0,379		81	0,044		95	0,001
	64	0,340		82	0,035		96	0,001
	65	0,303		83	0,028		97	0,001
	66	0,268		84	0,022		98	0,000
	67	0,235		85	0,017		99	0,000
	68	0,204		86	0,012		100	0,000
	69	0,176		87	0,009	9	72	0,519
	70	0,150		88	0,007		73	0,481
	71	0,126		89	0,005		74	0,444
	72	0,105		90	0,003		75	0,407
	73	0,087		91	0,002		76	0,371
	74	0,071		92	0,002		77	0,336
	75	0,057		93	0,001		78	0,303

$$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 8$)			**($n_1 = 8$)**			**($n_1 = 9$)**
9	79	0,271	10	93	0,073	9	107	0,031
	80	0,240		94	0,061		108	0,025
	81	0,212		95	0,051		109	0,020
	82	0,185		96	0,042		110	0,016
	83	0,161		97	0,034		111	0,012
	84	0,138		98	0,027		112	0,009
	85	0,118		99	0,022		113	0,007
	86	0,100		100	0,017		114	0,005
	87	0,084		101	0,013		115	0,004
	88	0,069		102	0,010		116	0,003
	89	0,057		103	0,008		117	0,002
	90	0,046		104	0,006		118	0,001
	91	0,037		105	0,004		119	0,001
	92	0,030		106	0,003		120	0,001
	93	0,023		107	0,002		121	0,000
	94	0,018		108	0,002		122	0,000
	95	0,014		109	0,001		123	0,000
	96	0,010		110	0,001		124	0,000
	97	0,008		111	0,000		125	0,000
	98	0,006		112	0,000		126	0,000
	99	0,004		113	0,000	10	90	0,516
	100	0,003		114	0,000		91	0,484
	101	0,002		115	0,000		92	0,452
	102	0,001		116	0,000		93	0,421
	103	0,001					94	0,390
	104	0,000			**($n_1 = 9$)**		95	0,360
	105	0,000	9	86	0,500		96	0,330
	106	0,000		87	0,466		97	0,302
	107	0,000		88	0,432		98	0,274
	108	0,000		89	0,398		99	0,248
10	76	0,517		90	0,365		100	0,223
	77	0,483		91	0,333		101	0,200
	78	0,448		92	0,302		102	0,178
	79	0,414		93	0,273		103	0,158
	80	0,381		94	0,245		104	0,139
	81	0,348		95	0,218		105	0,121
	82	0,317		96	0,193		106	0,106
	83	0,286		97	0,170		107	0,091
	84	0,257		98	0,149		108	0,078
	85	0,230		99	0,129		109	0,067
	86	0,204		100	0,111		110	0,056
	87	0,180		101	0,095		111	0,047
	88	0,158		102	0,081		112	0,039
	89	0,137		103	0,068		113	0,033
	90	0,118		104	0,057		114	0,027
	91	0,102		105	0,047		115	0,022
	92	0,086		106	0,039		116	0,017

$w_1 \geq \frac{n_1(n_1+n_2+1)}{2}$

n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$	n_2	w_1	$P(W_1 \geq w_1)$
		($n_1 = 9$)			**($n_1 = 10$)**			**($n_1 = 10$)**
10	117	0,014	10	108	0,427	10	132	0,022
	118	0,011		109	0,398		133	0,018
	119	0,009		11	0,370		134	0,014
	120	0,007		111	0,342		135	0,012
	121	0,005		112	0,315		136	0,009
	122	0,004		113	0,289		137	0,007
	123	0,003		114	0,264		138	0,006
	124	0,002		115	0,241		139	0,004
	125	0,001		116	0,218		140	0,003
	126	0,001		117	0,197		141	0,003
	127	0,001		118	0,176		142	0,002
	128	0,000		119	0,157		143	0,001
	129	0,000		120	0,140		144	0,001
	130	0,000		121	0,124		145	0,001
	131	0,000		122	0,109		146	0,001
	132	0,000		123	0,095		147	0,000
	133	0,000		124	0,083		148	0,000
	134	0,000		125	0,072		149	0,000
	135	0,000		126	0,062		150	0,000
				127	0,053		151	0,000
		($n_1 = 10$)		128	0,045		152	0,000
10	105	0,515		129	0,038		153	0,000
	106	0,485		130	0,032		154	0,000
	107	0,456		131	0,026		155	0,000

Observation:

$$\mathbb{P}(W_1 \leq w_1) = \mathbb{P}(W_1 \geq n_1(n_1 + n_2 + 1) - w_1)$$