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

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 carries an error, its complement is the degree of reliability. 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 Sometimes 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.
- Being economically unfeasible in time or cost.

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

Probability

There is a population from which a perfectly described character is studied. for 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 specication. It is known that the diameter follows a normal distribution* $\mathcal{N}(3.0.005)$*. What percentage of devices will be discarded?*

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

There is a population from which a character is studied of which the random variable that describes it is not known, 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 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.*

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

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. one is taken 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.*

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 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 and the infinite population. 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 :* $p^* = \frac{1}{n} \sum_{i=1}^{n} X_i$

Definition 1.1.3. *Let X be a random variable. We define the k-th moment about zero as $\alpha_k = \mathbb{E}[X^k]$ and the k-th moment about the mean as $\mu_k = \mathbb{E}[(X - \mathbb{E}[X])^2]$.*

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 about zero as $a_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$ and the k-th sample moment about 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 $\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2$, which implies that the desired result follows.

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

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

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

   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 proven in Probability Calculus and, therefore, its 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*

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

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,*

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

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, $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$. Therefore,

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

which simplifying gives $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n-1}}}$ and the desired result follows. □

If $n \geq 30$, we have that $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n-1}}} \sim N(0,1)$.

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 quasivariance 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 Fisher distribution, we have that $F_{1-\alpha,n,m} = \frac{1}{F_{\alpha,m,n}}$. The proof is here:*

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

Therefore, $1-\alpha = \mathbb{P}\left(F_{n,m} \geq \frac{1}{F_{\alpha,n,m}}\right)$ 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. 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,*

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

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

Proof. From Proposition 1.1.8, we know that

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

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

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

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

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

Example 6. *If $X \sim Poi(\lambda)$, where λ is known. Therefore, we have shown the maximum*

$$L(x_1, \dots, x_n; \lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\dots\frac{e^{-\lambda}\lambda^{x_n}}{x_n!} = \frac{e^{-n\lambda}\lambda^{x_1+\dots+x_n}}{x_1!x_2!\dots x_n!}.$$
$$LnL(x_1, \dots, x_n; \lambda) = -n\lambda + (x_1 + x_2 + \dots + x_n)Ln\lambda - Ln(x_1!x_2!\dots x_n!)$$
$$\frac{\delta LnL(x_1,\dots,x_n;\lambda)}{\delta\lambda} = -n + \frac{(x_1+x_2+\dots+x_n)}{\lambda} = 0. \implies \lambda^* = \frac{(x_1+x_2+\dots+x_n)}{n} = \overline{x}.$$

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.

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 $i = 1, \dots, k$. This can be also done with the moments about the mean instead of about the zero.

Example 7. *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 8. *If $X \sim Poi(\lambda)$, where λ is unknown. Thus, $a_1 = \bar{X}$ and $\alpha_1 = \lambda$. Therefore, $\mu^* = \bar{X}$.*

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 $X_1, \dots, X_n$ 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 (Bias)**.** *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 derive 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

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

We derive 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 1.2.1.

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

On the other hand, from the definition of the linear correlation coefficient, we can derive the following property:

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

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.

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 $[\mu_n^*] = \mu$ 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:*

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

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 $\vec{Y} = (X_1 + a, \ldots X_n + a)$. 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 derived estimator 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.

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.

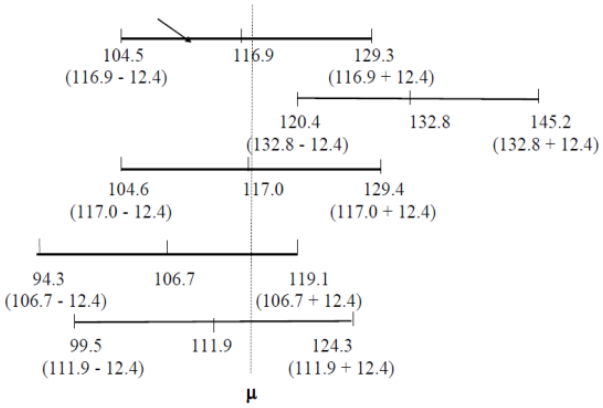

Figure 1.1: Five confidence intervals.

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 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, 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

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

□

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 $\frac{nS^2}{\sigma^2}$, 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*

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

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

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

□

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

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

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

where S_p is given in Proposition 1.1.8.

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 $X_{11}, \dots, X_{1n}$ be a simple random sample of size n from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \dots, X_{2m}$ 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.*

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

where df is given in Proposition 1.1.9.

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

In Proposition 1.1.6, we show that $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1,m-1}$. We take thus $\frac{S_1^2/(n-1)}{S_2^2/(m-1)}$ as pivot to provide an analytical expression of the CI of this case.

Proposition 1.3.7. *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 $X_{21}, \dots, X_{2m}$ 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.*

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

where df is given in Proposition 1.1.9.

Proof. We have that

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

□

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 $X_{11}, \ldots, X_{1n}$ be a simple random sample of size n from a population defined by $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{21}, \ldots, X_{2n}$ 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 $X_1, \ldots, X_n$ 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 $X_{11},\dots,X_{1n}$ be a simple random sample of size n from a population defined by $X_1 \sim Ber(p_1)$ and $X_{21},\dots,X_{2m}$ be a simple random sample of size n 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 $(1-\alpha)\cdot 100\%$

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

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

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 $(x_1, \ldots, x_n) \in S_1$, assuming that H_0 is true, that is,*

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

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.

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 $n_2<30$	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \approx t_{df}$	$\left(\bar{x}_1-\bar{x}_2 \mp t_{\frac{\alpha}{2};gl}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$	$t_p = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
$\mu_1-\mu_2$	σ_1 y σ_2 unknown Normality or $n_1 \geq 30$ and $n_2 \geq 30$	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \approx N(0,1)$	$\left(\bar{x}_1-\bar{x}_2 \mp z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$	$z_p = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
$\frac{\sigma_1^2}{\sigma_2^2}$	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) $n_1p_1^*>5, n_1q_1^*>5$ $n_2p_2^*>5, n_2q_2^*>5$	$\frac{p_1^*-p_2^*-(p_1-p_2)}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} \approx N(0,1)$	$\left(p_1^*-p_2^* \mp z_{\frac{\alpha}{2}}\sqrt{\frac{p_1^*q_1^*}{n_1}+\frac{p_2^*q_2^*}{n_2}}\right)$	$z_p = \frac{p_1^*-p_2^*}{\sqrt{\frac{p_1^*q_1^*}{n_1}+\frac{p_2^*q_2^*}{n_2}}}$

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \qquad df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

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.

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.

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.

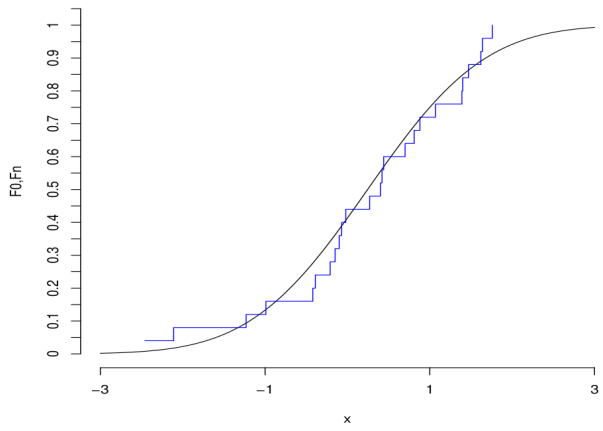

Figure 4.1: An example of F_0 (in black) and F_n (in blue).

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 $x_1, \ldots, x_n$.

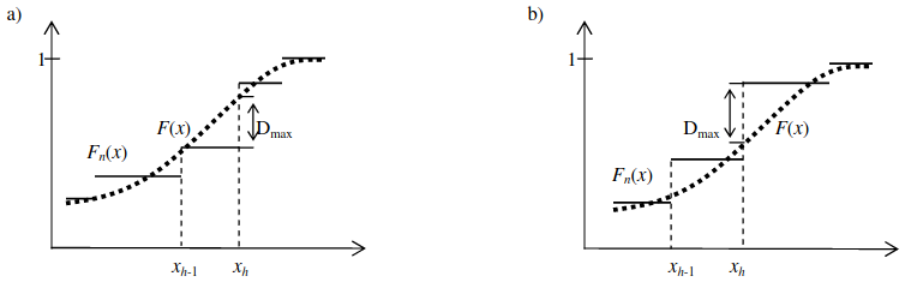

Figure 4.2: The computation of D_n.

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 none of the signs are zero (which implies that none of the values 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* μ_0 *and 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 distribution of the following two samples:*

- *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:

$$H_0 : F_X(x) = F_Y(x)$$
$$H_1 : F_X(x) \neq F_Y(x),$$

where F_X *is the cumulative distribution function of population X and* F_Y *is the cumulative distribution function 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 distributions of the population 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 distribution 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

$$H_0 : F_X(x) = F_Y(x)$$
$$H_1 : F_X(x) \neq F_Y(x),$$

where F_X is the cumulative distribution function of population X and F_Y is the cumulative distribution function 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 distributions 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 are equally distributed. 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 $\alpha = 0.5$. 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:

$H_0 : F_1(x) = F_2(x) = F_3(x)$
H_1 : *There is a difference on the distribution 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 follow the same distribution.