People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Abou-bekr Belkaid University – Tlemcen

Faculty of Humanities and Social Sciences
Department of Social Sciences

Courses of

mathematical statistics Supplement

Intended for 3rd year undergraduate students

Field: Demography

Dr. MORTAD Nadjla

Academic year: 2023-2024

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Abou-bekr Belkaid University – Tlemcen

Faculty of Humanities and Social Sciences
Department of Social Sciences

Courses of

mathematical statistics Supplement

Intended for 3rd year undergraduate students

Field: Demography

Dr. MORTAD Nadjla

Academic year: 2023-2024

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة أبي بكر بلقايد

كلية العلوم الإنسانية والعلوم الاجتماعية

المجلس العلمي للكلية

الرقم: 167/م ع ك ع ا/ 2024

التاريخ: 2024/07/04

شهادة المجلس العلمي خاصة بالسند التربوي

إن رئيس المجلس العلمي لكلية العلوم الإنسانية والعلوم الاجتماعية

بناء على محضر المجلس العلمي للكلية بتاريخ: 2024/04/25.

بناء على محضر تعيين خبيرين متخصصين بتاريخ: 2024/04/25،

بناء على تقييم الخبرة النهائية للسند التربوي الخاص بالدكتورة: **مرتاض نجلاء**، تحت عنوان:

"Courses of mathematical statistics supplement" لفائدة طلبة السنة الثالثة ديموغرافيا.

يشهد بأن السند التربوي المذكور أعلاه قابل للنشر والتوزيع، ويمكن اعتماده من الناحية العلمية.

رئيس المجلس العلمي

كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

رئيس المجلس العلمي للكلية

الأستاذ الدكتور: فقيه العيد

عميد الكلية

جامعة أبو بكر بلقايد
كلية العلوم الإنسانية و العلوم الإجتماعية
- 1 -
تلمسان

عميد كلية العلوم الإنسانية والعلوم الإجتماعية - جامعة تلمسان
أ.د. نصر الدين بن داود

Table of contents

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإسلامية
جامعة أبي بكر بلقايد

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

List of Tables

المجلس العلمي
كلية العلوم الإنسانية والعلوم الاجتماعية
جامعة أبي بكر بلقايد

List of Figures

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة باجي مختار - عنابة

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Foreword

This document "Mathematical statistics supplement ", aligns with the taught curriculum and serves as a supplementary course in statistics. It is aimed at third year undergraduate students in demography (LMD system) to enhance their knowledge in the fields of descriptive and inferential statistics.

Descriptive statistics aim to summarize the information contained in data concisely and effectively through graphical representations, measures of central tendency, measures of dispesion, and ultimately to identify the essential characteristics of the studied phenomenon[1].

Inferential statistics, on the other hand aim is to make predictions and decisions based on observations through the study of hypothesis tests by studying hypothesis tests.

This course material is organized into two chapters and ten courses.; the first chapter is devoted to descriptive statistics with definitions of notions, fundamental concepts, the study of statistical distributions in a single dimension, the presentation of data, graphical representations, characteristics of central tendency and dispersion.

The second chapter covers inferential statistics, specifically hypothesis testing; their principle, and types such as conformity tests and homogeneity tests.

Various solved exercises are integrated into the mentioned courses to facilitate student understanding.

[1]- The courses will take into account series with one variable.

The objective of these courses are twofold: firstly, to enable students to organize observed data; group them into tables and graphs, reduce them into parameters and indicators, and explain the obtained results.

Secondary, to provide students with a methodology aimed at establishing a decision rule that, based on sample results, allows for choosing between two statistical hypotheses.

Pedagogical plan for the course:

Subject: Mathematical statistics Supplement

Field: Demography

Level: Third- year demography (Semester5)

Total Hours: 36 hours (lectures + tutorials)

Coefficient:2

Evaluation: Continuous assessment+ final exam

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Chapter I: Descriptive statistics

Course 1: Introduction to descriptive statistics

The word statistics – from the Italian "Statista", statesman – originally referred to the collection and evaluation of data concerning a state. This science of the State was a purely descriptive representation of geographical and social facts such as climate, population, customs, economic organizations, etc.

From ancient times, leaders have carried out surveys of the population: Emperor Yao (around 2200 BC) to find out agricultural production, the Egyptian pharaohs (from 1700 BC- C.), Emperor Augustus in Rome for the number of soldiers, the income of the citizens[1].

The word "Statistics" was created more precisely in the 18th century by the German professor Gotteried Achenwall (1719-1772). But statistics were used long before as we have already seen. Indeed, human and land population counts have been carried out since ancient times for war and taxation.[2].

In the 19th century, there was the appearance of probability calculation which is closely linked to games of chance. This gave rise to a discipline called "Mathematical Statistics". during this period, the Belgian Adolphe Quetelet (A796-1874) transposed the calculation of probability to economics and demography.

Today, this part of mathematics has taken on a large role thanks to new techniques and the power of computers. Geography, medicine, human sciences, economics, biology, politics, no field is spared.

Descriptive statistics methods make it possible to carry out studies using exhaustive data, that is to say concerning all individuals in the population concerned by the study.[3].

---

[1]- Jt, 3. . (2021). Some elements of descriptive statistics. http://www.gymomath.ch/javmath.

[2]- Al Abassi, I., El Marhoum, A (1999). Descriptive statistics course. Collection of the Faculty of Legal, Economic and Social Sciences, Marrakech.

[3]- Bressoud, é., & kahané, C. (2010). Descriptive statistics. 2nd Pearson edition.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الاجتماعية
جامعة زيان عاشور بالجلفة

1-Definitions

Statistics is a method of collecting, presenting, and analyzing observations relating to individuals belonging to the same precisely defined set to highlight certain general properties of this set. Descriptive statistics is the part of statistics whose role is to describe a phenomenon; in other words; to measure it, to evaluate it, to classify the measurements, to present these measurements in the form of tables or graphs, to summarize these measurements using a few indicators to have a quick and simple idea of the phenomenon studied and also to make it possible to make comparisons.

When the data only concerns a sample of the population, as in the case of surveys, we use inferential statistics (inductive statistics), which uses probability theory.

2-Terminology: basic vocabulary

The term statistics is rich in meanings, in the singular it is a set of mathematical techniques for processing digital data. In the plural, statistics are synonymous with numbers, data and digital information, indicating a plurality of phenomena through the numbers attached to the apprehension of these phenomena[4].

-Descriptive and inferential statistics

Descriptive statistics is a set of methods for describing, presenting, and summarizing data. These methods can be numerical (sorting, drawing up tables, calculating averages, etc.) and/or lead to graphic representations[5]. Inferential statistics makes it possible to generalize the results of the sample to a larger group, this group is called universe or population. Inferential statistics solve two fundamental problems: parameter estimation and hypothesis testing which will be the subject of the 2nd chapter.

-Population

The population designates a set of statistical units. Statistical units, also called individuals, are abstract entities that represent people, animals or objects. Statistics is used to describe all the statistical units that make up the population.

---

[4]- Bailly, p. Carrère, Ch (2007). Descriptive statistics - Course. Grenoble University Press. PUG.

[5]- Goldfarb, b., & Pardoux, c. (2011). Introduction to the statistical method, manual, and correct exercises. Paris: 6th edition, Dunod.

المجلس العلمي
العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

The term statistical population predates demography and was originally applied to categories of humans. Later, demography came with the idea of equality of individuals, which led to the idea of the census.[6]

-Statistical population

The statistical population is all the elements to which the study relates. The elements of the population are called statistical individuals or statistical units. The population constitutes the reference universe of the study. If the population includes N individuals, we will note $\Omega = \{\omega 1 \ldots; \omega N\}$, ωi designating for i varying from 1 to N the individuals who compose it. A sample of size n is a subset of n individuals from the population ($n \leq N$).

-Statistical unit

Each element belonging to a population is called a statistical unit.

-Sample

A sample is a subset of the entire population. It must have the fundamental properties of the whole from which it comes.

-Investigation

Survey is a set of operations that aim to collect information relating to a population in an organized manner.

-Frequencies and relative frequencies

The effective (also called absolute frequency) of the modality X_i, is noted n_i, and designates the number of individuals in the population presenting the modality X_i.

The total number of the population is then: $n= n1+n2 +n3+\ldots..+ni$.

$n= \sum ni$ (the sum of ni, for i varying from 1 to r).

The frequency (default relative frequency) of the modality x_i is noted f_i, and is defined by : $f_i = \frac{n_i}{N}$.

The frequency expresses the proportion of individuals presenting a given modality. It can be expressed in the form of a decimal number (generally with a precision of four digits after the decimal point) or the form of a percentage. In statistics, the term frequency is used more often than that of proportion.

[6]- Bressoud, é., & kahané, C. (2010). Descriptive statistics. 2nd Pearson edition.

المجلس العلمي

-Modality

Each character has two or more modalities. These are the different situations where the statistical units can be found about the character considered, for example the "nationality" character can have as modality: Algerian, Tunisian, French… the "number of children per family" character can have as modalities: Zero children, one child, two children, and more than two children…etc.

-Variable or statistical character

A given individual in the population can be studied according to certain properties. These properties are called statistical characters or variables. A statistical variable denoted X, is an application defined on a statistical population and values in a set M, called the set of modalities. The modalities correspond to the possible values of the statistical variable (Figure 1).

If the number of modalities is noted r, the set of modalities of the variable X will be noted: $M = \{x1; x2; \ldots; xr\}$.

-Qualitative variables

A statistical variable is said to be qualitative if the modalities are not measurable.

Gender, occupation, and marital status are some examples of qualitative variables. The modalities of a qualitative variable are words and can be classified on two types of scale: nominal or ordinal. These two types of scale correspond to two types of qualitative variables.

- Nominal qualitative variables

Nominal qualitative variables cannot be measured. However, their terms can be coded. The order and origin of the coding are arbitrary, this coding can be numeric, alphabetic, or alphanumeric. Therefore, a qualitative statistical variable is said to be defined on a nominal scale if its modalities are not naturally ordered.

- Ordinal qualitative variables

An ordinal scale assumes the existence of a total order relationship between the categories, that is to say, that we can classify all the categories, from the smallest to the largest (or, conversely, from the largest to the smallest). The coding can be numeric, alphabetical, or alphanumeric. Therefore, a qualitative statistical variable is

المجلس العلمي
جامعة أبي بكر بلقايد

said to be defined on an ordinal scale, if all of its modalities can be endowed with an order relationship.

-Quantitative variables

Any variable that is not qualitative can only be quantitative. The different modalities of a quantitative variable constitute the set of numerical values that the variable can take.

A statistical variable is said to be quantitative if its modalities are measurable. The modalities of a quantitative variable are numbers linked to the chosen unit, which must always be specified.

There are two types of quantitative variables: discrete variables and continuous variables.

These variables have clearly ordered modalities in common, for which the difference between the values has significance, and on which it is possible to carry out mathematical operations such as calculations of averages.

- Discrete quantitative variables

When the modalities are isolated numerical values, such as the number of children per household, we speak of a discrete variable. Therefore, a quantitative statistical variable is said to be discrete if the set of its modalities is finite or countable. Thus, the set of modalities can be given in the form of a list of numbers, $M=\{x_1, x_2, x_3 \dots \dots x_i \dots\}$ finite or infinite.

Most often, the modalities belong to the set N of natural numbers ($N = \{0; 1; 2; 3 \dots .\}$). However, a discrete variable can take non-integer values[7]

- Continuous quantitative variables

When the variable, for example, the height of an individual, can take all the values of an interval, these values can then be grouped into classes, and in this case, we speak of a continuous variable.

A quantitative statistical variable is said to be continuous if all of its modalities are not countable. Thus, a continuous variable can take all the values of an interval. To study a continuous statistical variable, classes or intervals of possible values are defined. The classes retained constitute the modalities of the variable.

[7]- Bressoud, é., & kahané, C. (2010). Descriptive statistics. 2nd Pearson edition.

المجلس العلمي

We call: class amplitude $[a_i; b_i]$ the real A_i representing the length of the interval and defined by : $A_i = [b_i - a_i]$

The class center of the class $[a_i; b_i[$ is the real number noted xi representing the middle of the interval and given by: xi = (ai + bi) / 2; it is the arithmetic mean of the class limits.

The number of classes, k is calculated by one of two formulas:

Sturge's rule $k=1+3.3\log(n)$

Yule's rule $k=2.5(n)^{1/4}$

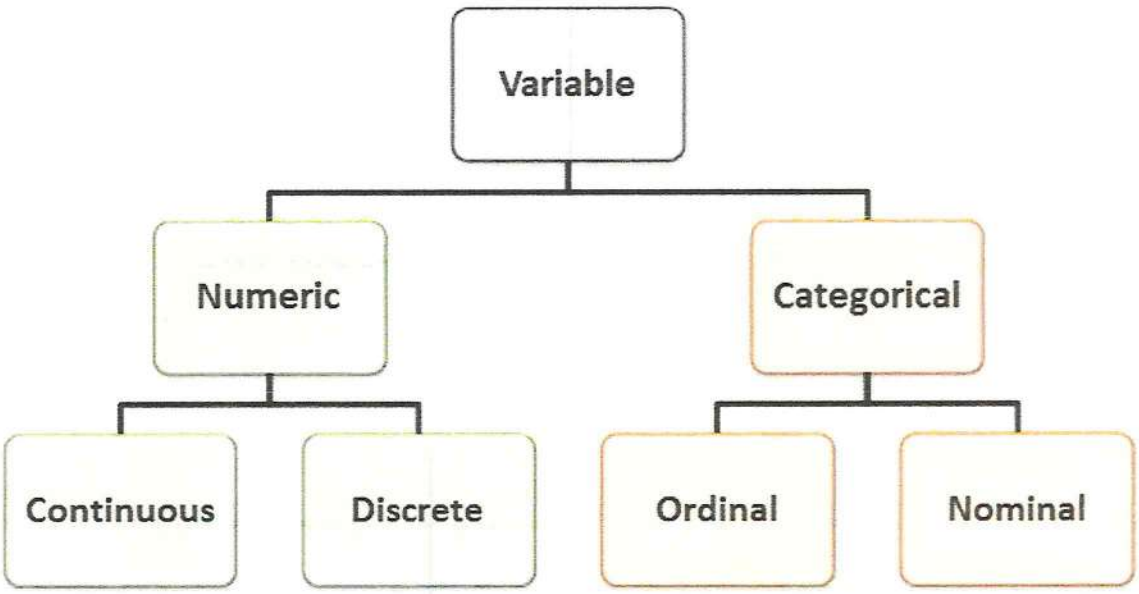


Figure 1: Statistical variables

Course 2: Data presentation

The statistical data comes from raw data presented in the form of statistical tables in which the numbers and/or frequencies are indicated (Figure 2).

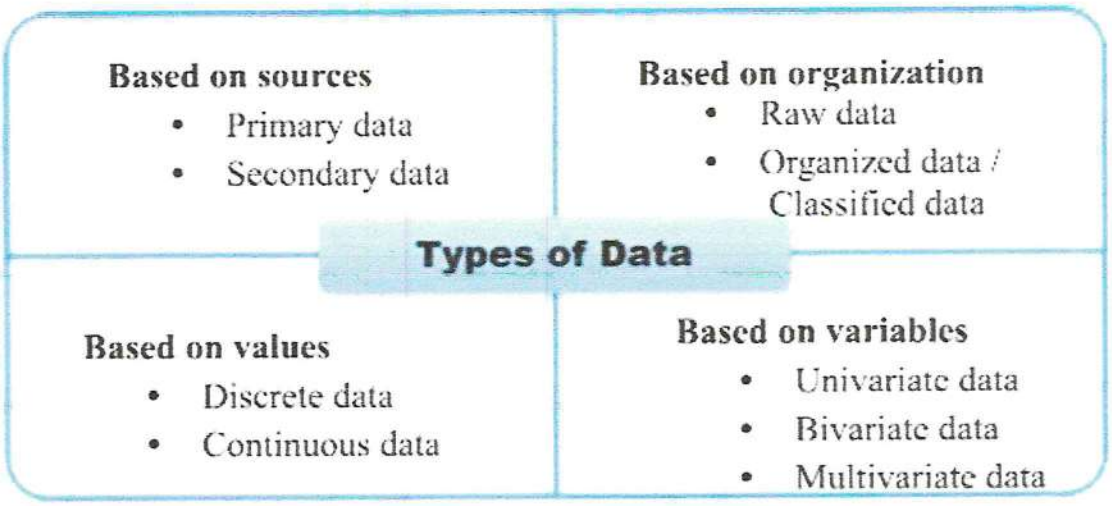


Figure 2: Type of Data[8]

1. Distribution of frequencies

The statistical tables containing the numbers and/or frequencies are the first exploitation of the raw data.

2. Raw data

We call raw data that we have collected without deviating from the notion of order.

3. Statistical table

It is essential to define the population and rigorously specify the variable(s) recorded on each individual in the population or sample representing it. When the observations have been collected, the first task is to present them, as clearly as possible, in the form of a statistical table.

This table reveals the statistical distribution by presenting the type of couples (xi; ni)[9], where the xi are the modalities and the ni their respective effectives. It is

[8] - In these courses, we focus on univariate Data.

[9]- i integer varying from 1 to r, if r designates the number of modalities of the character.

المجلس العلمي

also possible to present the frequency distribution, that is to say, the couples of type (xi; fi).

4. Raw data and flat sorting

Or elementary table the table showing for each statistical unit the modality of the variable studied. Flat sorting is the transformation that allows you to move from the raw data table to the statistical distribution table presenting the categories and numbers, the categories being classified in ascending order.

5. Discretization

In the case of a continuous quantitative statistical variable, it is necessary to define classes to be able to propose a flat sorting. We call discretization the division into classes of a quantitative statistical series.

This division into classes raises many questions: choice of amplitudes, constant or variable amplitudes, number of classes, etc. (see the exercise solved later).

6. Distribution of cumulative frequencies and relative frequencies: quantitative variables

6.1-Increasing cumulative frequencies on discrete variable

If X designates a discrete quantitative variable, we call increasing cumulative frequency, denoted nicc, the number of statistical individuals for which we have :

$$n_1cc = n_1 \text{ et } n_icc = n_1 + n_2 + \ldots + n_i = \sum_{k=1}^{i} n_k .$$

If the series has r modalities, x_r then designates the greatest value of X

$$n_rcc = n_1 + n_2 + \ldots + n_r = \sum_{k=1}^{r} n_k = n$$

6.2-Increasing cumulative relative frequencies on discrete variable

With the same hypotheses, we define the increasing cumulative frequency, denoted ficc, representing the proportion of statistical individuals for which X is less than or equal to xi. We have:

$$f_1cc = f_1 \text{ et } f_icc = f_1 + f_2 + \ldots + f_i = \sum_{k=1}^{i} f_k ,$$

Or :

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * تلمسان
المجلس العلمي

$$f_i cc = \frac{n_i cc}{n}.$$

If the series has r modalities, x_r then designating the largest value of X, we have:

$$f_r cc = f_1 + f_2 + ... + f_r = \sum_1^r f_k = 1$$

Or 100 if relative frequencies are expressed as percentages.

In the case of a continuous quantitative variable, the data are grouped into classes [ai; bi[, and we define, in the same way as for a discrete variable, nicc the number of statistical individuals for which X is less than or equal to bi, and ficc the proportion of statistical individuals for which X is less than or equal to bi.

6.3-Decreasing cumulative frequencies on discrete variable

If X designates a discrete quantitative variable, we call decreasing cumulative number, denote nicd, the number of statistical individuals for which

$$\text{On a : } n_1 cd = n\ ;\ n_i cd = n_i + n_{i+1} + \ldots + n_r = \sum_{k=i}^{r} n_k,$$
$$n_r cd = n_r.$$

r designating the number of modalities.

6.4-Decreasing cumulative relative frequencies on discrete variable

With the same hypotheses, we define the decreasing cumulative frequency, denoted ficd, representing the proportion of statistical individuals for which X is greater than or equal to xi. We have:

$$f_1 cd = 1\ ;\ f_i cd = f_i + f_{i+1} + \ldots + f_r = \sum_{k=i}^{r} f_k$$

And:

$$f_r cd = f_r$$

Or :

$$f_i cd = \frac{n_i cd}{n}.$$

العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي

In the case of a continuous quantitative variable, the data are grouped into classes [ai; bi[,and we define, in the same way as for a discrete variable, nicd the number of statistical individuals for which X is greater than or equal to ai, and ficc the proportion of statistical individuals for which X is greater than or equal to ai.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

Solved exercise

The series represents the blood glucose level (glycemia) determined in 32 subjects is given below in g/l

Ordered series:

0.85	0,95	1,00	1,06	1,11	1,19
0,87	0,97	1,01	1,07	1,13	1,20
0,90	0,97	1,03	1,08	1,14	
0,93	0,98	1,03	1,08	1,14	
0,94	0,98	1,03	1,10	1,15	
0,94	0,99	1,04	1,10	1,17	

1- Give the discretization of this series (the division into classes).

2- How many classes are there?

3- Calculate absolute frequencies and increasing cumulative numbers.

Solution :

We have:

n= 32 and the Yule formula gives:

$= 2.5(32)^{1/4} = 5.94 \approx 6.$

We have 6 classes to divide; the fi and the ncc are mentioned in the table below:

Classe g/l	$c_i g/l$	n_i	f_i	$n_i c$
[0,85 : 0,91[	0,88	3	3/32	3
[0,91 : 0,97[	0,94	4	4/32	7
[0,97 : 1,03[	1,00	7	7/32	14
[1,03 : 1,09[	1,06	8	8/32	22
[1,09 : 1,15[	1,12	6	6/32	28
[1,15 : 1,21]	1,18	4	4/32	32
		$n = \sum n_i = 32$	$\sum f_i = 1$	

Course 3: Graphical representations of univariate data

Although the statistical table contains all the information gathered, it is very important to translate it into a graph. The graphical representation of a frequency distribution makes it possible to visualize and detect its main characteristics.

The appearance of statistical graphics, linked to the use of coordinates, essentially owes its origin to the philosopher and mathematician René Descartes (1596-1650)[10]. These graphs constitute an essential visual synthesis of the information contained in the statistical table.

The graphics used depend on the nature of the variable. The diagrams used to represent the distributions of numbers (or frequencies) are circular diagrams (or sectors), organ pipe diagrams, bar charts, histograms, and the polygon of numbers. For cumulative distributions, we will use the polygons of increasing and decreasing cumulative numbers (or frequencies).

1-Graphs for qualitative variables

Qualitative variables – nominal or ordinal – can be represented using a circular diagram or an organ pipe diagram.

1.1-Circular diagram

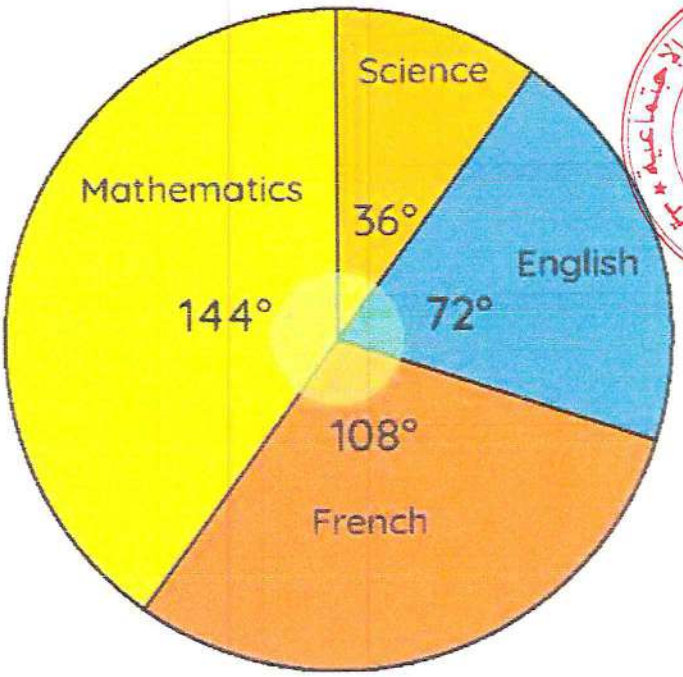
The circular diagram, also called a "pie chart", allows a representation of the distribution of a variable in a circle which represents 100% of the modalities (Figure 3).

A circular diagram is a graph made up of a circle divided into sectors whose central angles are proportional to the numbers (or frequencies). The areas of the sectors are proportional to the numbers. The angle αi of an effective modality ni is given in degrees by:

$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$$

It is also possible to use a semi-circular graph in the shape of a half-circle (180°).

[10]- Bressoud, é., & kahané, C. (2010). Descriptive statistics. 2nd Pearson edition.



كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Figure 3: A pie chart shows the favorite subjects of students in a class

1.2-Organ pipe diagram (in bars)

The organ pipe diagram is a representation of the distribution of a variable according to horizontal or vertical rectangles all having the same base, of arbitrary width (Figure 4).

An organ pipe diagram is a graph that, for each modality of a qualitative variable, associates a rectangle with a constant base whose height is proportional to the number (or frequency). The areas of the sectors are proportional to the numbers. Rectangles are generally disjoint, vertical or horizontal.

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

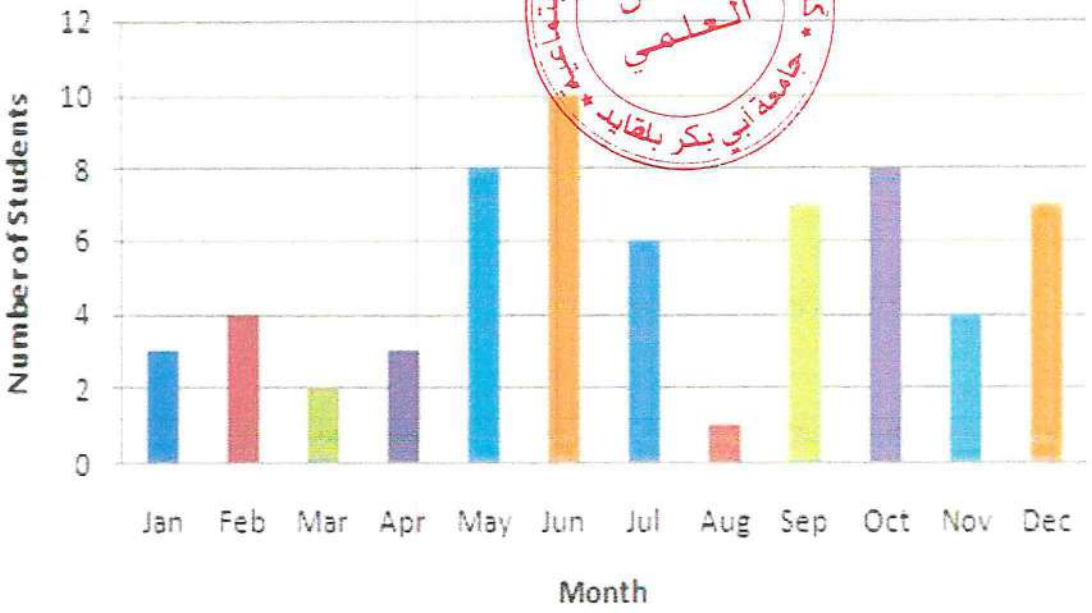


Figure 4: A bar graph: Birthday of students by month

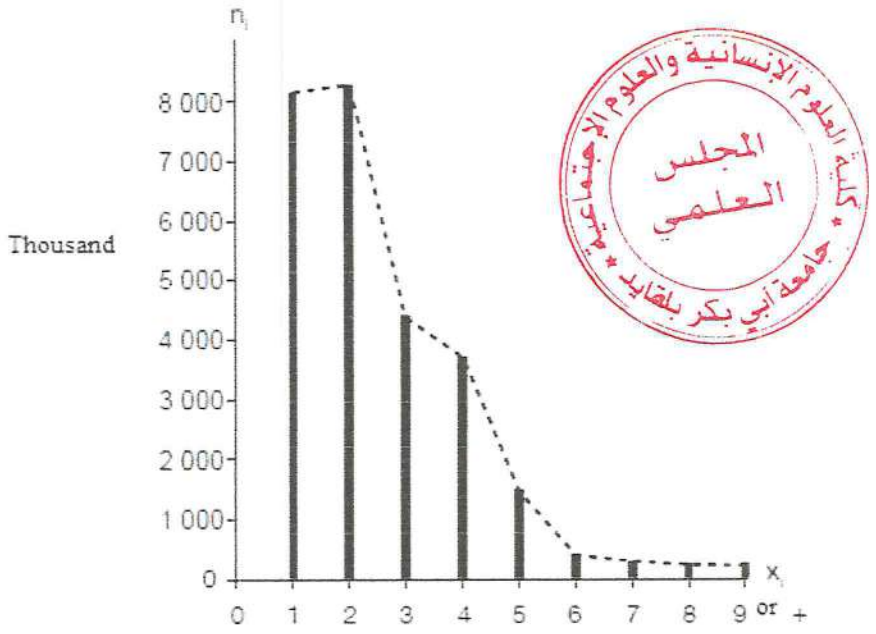
2-Graphs for quantitative variables

The graphic representation of a quantitative variable depends on its nature: discrete or continuous.

2.1-Discrete variables: bar diagram

The distribution of a discrete quantitative variable can be represented by a bar chart.

A bar chart is a graph that associates with each modality of a discrete quantitative variable a segment (stick) whose height is proportional to frequencies (or relative frequencies) (Figure 5).



كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

Figure 5: Bar and polygon diagram of population numbers (number of people per household in a country

2.2- Continuous variables: Histogram

A histogram is a diagram composed of contiguous rectangles whose areas are proportional to the numbers (or frequencies) and whose bases are determined by the class intervals (Figures 6 and 7).

In the case of a continuous quantitative variable, we define the effective density di of a class of effective ni and amplitude Ai by: di = ni / Ai (or, in the case of frequencies, fi / Ai).

When creating a histogram, it is essential to distinguish two cases.

1. If the class amplitudes are equal, the height of the rectangles will correspond to the numbers (or frequencies) of the classes.

2. If the amplitudes are different, to constitute the histogram, it is necessary to:

– calculate, for each class, the amplitude Ai;

– calculate the density di = ni / Ai for a histogram of the numbers, and di = fi / Ai for a histogram of the frequencies;

– assign to each rectangle a height proportional to the density di of the corresponding class.

كلية العلوم الإنسانية والعلوم الإسلامية
المجلس العلمي
جامعة أبي بكر بلقايد

Let min(Ai) be the minimum class amplitude, the height is then called "corrected effective" and noted nic = di × min(Ai); this convention amounts to adopting min (Ai) as the class amplitude unit.

The classes having amplitudes min (Ai) are then represented by rectangles whose height is the number. Likewise, it is possible to retain as height the corrected frequency fic = di min(Ai), with di = fi / Ai in the case of a frequency histogram. The use of min (Ai) is an optional convention; a histogram is correct as long as the corrected numbers (or frequencies) are proportional to the densities.

Solved exercise

The human resources manager of a company noted the following statistical distribution corresponding to the seniority of management personnel in the company, expressed in years:

Classes	Frequency
[6,5; 8[	3
[8; 9,5[	8
[9,5; 11[	12
[11; 12,5[	19
[12,5; 14[	9
[14; 15,5[	5
[15,5; 17[	4
Total	60

The histogram of the workforce is presented with, on the same graph, the polygon of the workforce plotted as a solid curve. This polygon makes it possible to represent the distribution in the form of a curve; when the class amplitudes are equal, it is obtained by joining the midpoints of the upper bases of each rectangle of the histogram by line segments.

We generally add a class of zero sizes, on either side of the histogram, to respect the area compensation rule: the total area of the domain located between the x- axis

المحاضر المعلومي
جامعة أبي بكر بلقايد * تلمسان

and the polygon is equal to the sum of the areas of the rectangles of the histogram. It represents the total frequency.

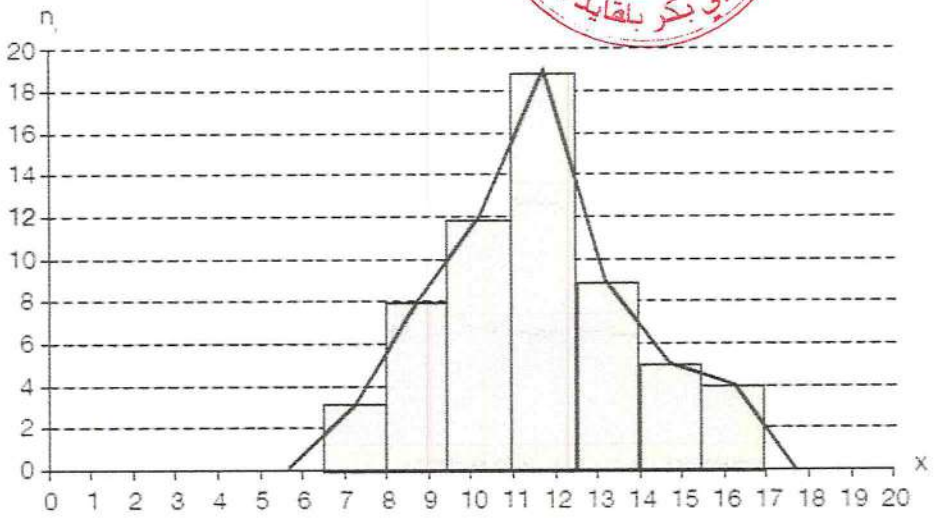


Figure 6: Histogram and frequency polygon, equal intervals (seniority of company staff)

- Let's take another example (the previous one modified slightly):

Classes	Frequency
[6,5; 8[	3
[8; 9,5[	8
[9,5; 11[	12
[11; 12,5[	19
[12,5; 14[	9
[14; 17[	9
Total	60

The classes being of unequal amplitudes, it is necessary to calculate the amplitudes (Ai), the densities (di), and then the corrected numbers (nic) for each class.

Classes	ni	ai	Di	Nic
[6,5; 8[	3	1.50	2	3
[8; 9,5[	6	1.50	5.33	8
[9,5; 11[	12	1.50	6	12
[11; 12,5[	19	1.50	12.67	19
[12,5; 14[	9	1.50	6	9
[14; 17[	9	1.50	3	4.5
Total	60	3		

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

We can then draw the histogram of the figure from the corrected numbers, as well as the polygon of the numbers, in a solid line.

To trace the polygon of the numbers, we carried out an artificial division into pseudo-classes of amplitude 1.5, from which we took the midpoints of the upper bases to respect the rule of compensation of the areas: the areas of the triangles outside the domain delimited by the polygon are equal to those of the triangles which are located under the polygon.

Thus, the total area of the domain located under the polygon of the numbers is equal to the total area of the rectangles of the histogram.

What is done in this example from the numbers can also be done from the frequencies, to draw the histogram and the frequency polygon[11]

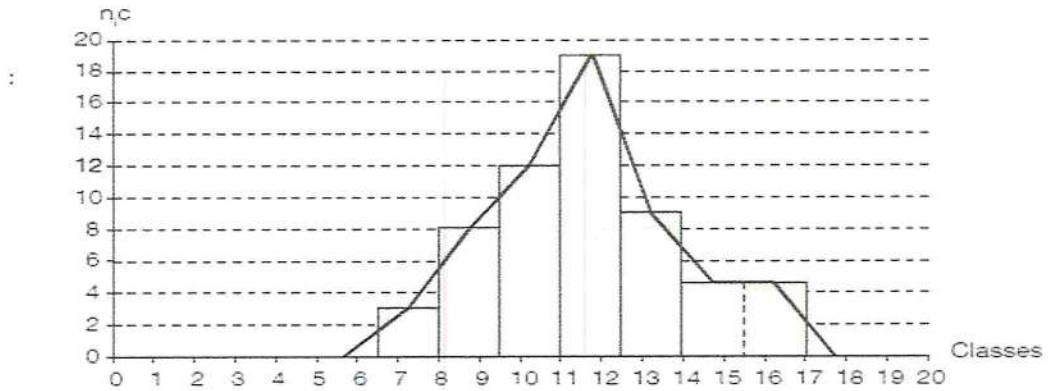


Figure 7: Histogram and frequency polygon, unequal classes

[11]- Bressoud, é., & kahané, C. (2010). Descriptive statistics. 2nd Pearson edition.

المجلس
العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

3-Cumulative Diagrams

The notions of cumulative numbers and frequencies provided the opportunity to introduce the notion of distribution function.

3.1-Distribution function of a discrete variable

The distribution function of a discrete quantitative variable is a step function, that is to say, constant by interval. In addition, it is increasing from 0 to 1 and defined by:

- if $x < x1$, $F(x) = 0$;
- if $x = xi$, $F(x) = ficc$;
- if $xi \leq x < xi + 1$, $F(x) = ficc$;
- if $x \geq xr$, $F(x) = 1$.

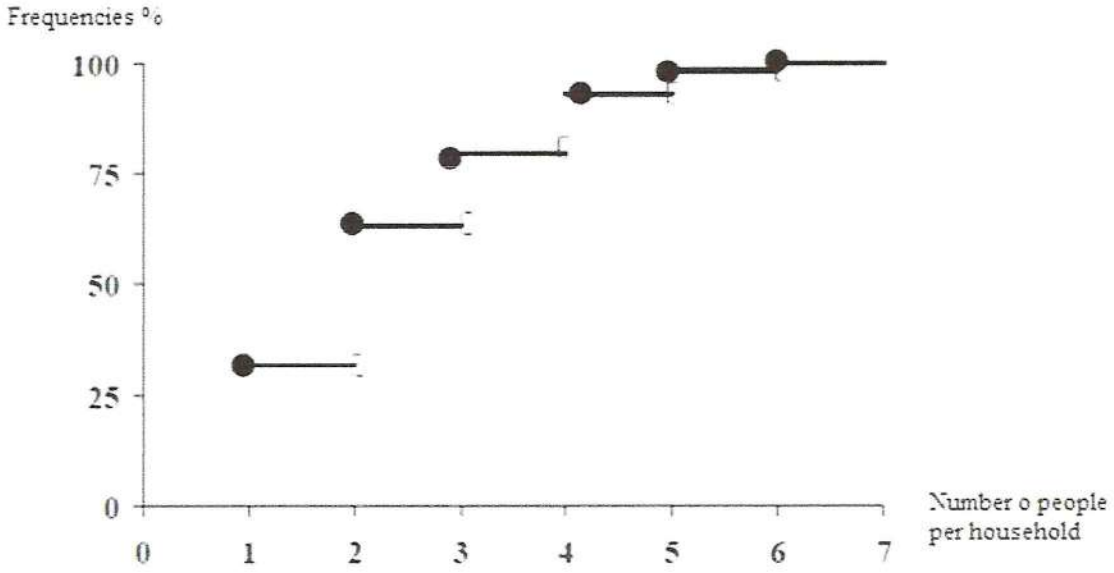


Figure 8: Cumulative diagram (discrete variable)

3.1-Distribution function of a continuous variable

A priori, the distribution function of a continuous variable is only known for the ends of classes. However, if we accept the hypothesis of uniform distribution of observations within each class, we can estimate the values of F(x) by linear interpolation. This amounts to approximating the graphic representation by a piecewise affine function: concretely, we trace the curve by joining two known consecutive points by a line segment (this curve is also called Galton's ogive)[12].

[12]- Bardin, B.M. (2016). Descriptive statistics course. Cango-Brazzavill: DEUG HAL.

3.1.1-Polygons of increasing and decreasing cumulative frequencies

In the case of a continuous variable, we define the polygons of the cumulative increasing and decreasing frequencies (or numbers), they will be used in particular to determine the median of the series.

Example :

Classes	Fi	Terminal Class	fcc	fcd
		0	0	100.00
[0; 15[	21.40	15	21.40	76.60
[15; 25[	16.60	25	38.00	62.00
[25,60[	51.10	60	89.10	10.90
60 and over	10.90	95	100.00	0.00

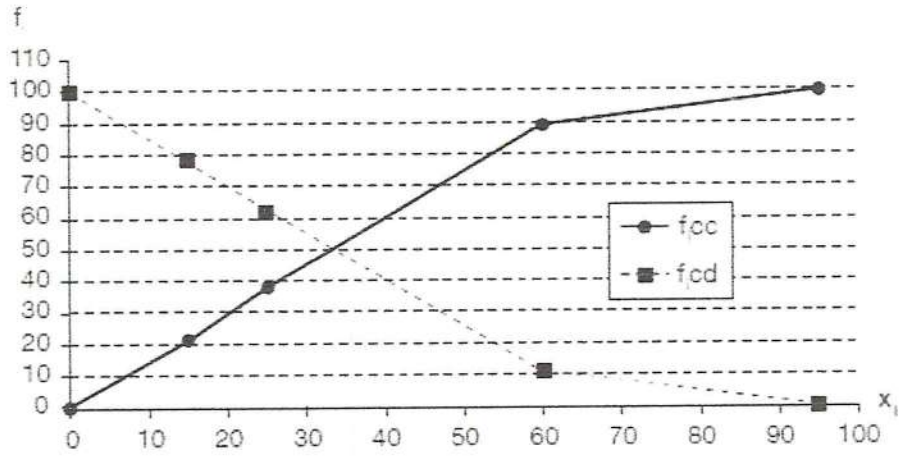


Figure 9: Polygon of increasing and decreasing cumulative frequencies (continuous variable)

-How to make choice out of various graphs

- A bar graph is used to indicate the comparison among different categories where independent variable is non-numerical.
- A pie graph is used for comparing parts of a whole, they do not show any changes over time.

- A histogram is used for representing the data intervals.
- Frequency Polygon is used to understand the shapes of distribution.
- The histogram and frequency polygon are two different ways to represent the same data set.

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Course 4: The characteristics of central tendency

Quantitative variables can be summarized by so-called "central tendency" characteristics.

As a result, man is led to determine the central characteristics (mean, median, etc.), to construct graphs (histograms, pie charts, etc.), to calculate dispersion characteristics (standard deviation, variation ratio, interquartile range…, etc.)[13]

The parameters of central tendency are quantities likely to best present a set of data.

These central values are the means, median, and mode.

I-The Mean

The mean constitutes one of the fundamental parameters of central tendency but is not sufficient to characterize a distribution. Complementary to the mode and especially to the median.

There are several types of averages; each adapted to specific situations.

A- The arithmetic mean

1- The simple arithmetic mean : The mean for ungrouped data

Example: either the series of numbers. The arithmetic mean of this series of figures is calculated as follows: $\{8, 5, 9, 13, 25\}$

$$\bar{X} = \frac{8+5+9+13+}{5} = 12$$

In general ;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

[13]Bardin, B.M. (2016). Descriptive statistics course. Cango-Brazzavill: DEUG HAL.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر

2- The mean for grouped data

Example1:

Let the series of numbers {8, 13, 5,8, 5, 9, 13, 25, 13, 9 }. Some numbers like 8 and 9 or 13 are repeated. The presentation can be simplified by grouping the data by values (Table 1).

ni	5	8	9	13	25
ni	2	2	2	3	1
ni.xi	10	16	18	39	25

Table 1: Calculation of the mean for grouped data

The mean for grouped data is then calculated by taking the weighted sum, that is to say, the sum of ni xi and dividing by n. it is equal to

$$\bar{X}=\frac{(5.2)+(8.2)+(9.2)+(13.2)+(25.1)}{10}=\frac{108}{10}=10.8$$

In general :

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Example2

Consider the following Table:

Class Interval	Ni	ci	ni.ci
[5 – 13[	6	9	54
[13 – 28[	3	7.5	22.5
[28 – 54[	5	41	205

Table 2: Calculation of the mean for values that are grouped by classes

We apply the previous formula but replacing xi with this:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}(n_i\cdot c_i)$$

In the example; we therefore:

$$\bar{x} = \frac{(6\times 9)-(3\times 7{,}5)-(5\times 41)}{14} = \frac{54-22{,}5-205}{14} = \frac{281{,}5}{14} \simeq 20{,}11$$

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

3- The Trimmed mean

Example: the series of grades of a student during the year:$\{11, 12, 13, 2, 14\}$

If we calculate the simple arithmetic mean,we obtain:

$$\bar{x} = \frac{12-13-11-14-2}{5} = \frac{52}{5} = 10{,}4$$

But if we remove the number 2 and recalculate the average trimmed from 4 notes, we will have:

$$\bar{x} = \frac{12-13-11-14}{4} = \frac{50}{4} = 12{,}5$$

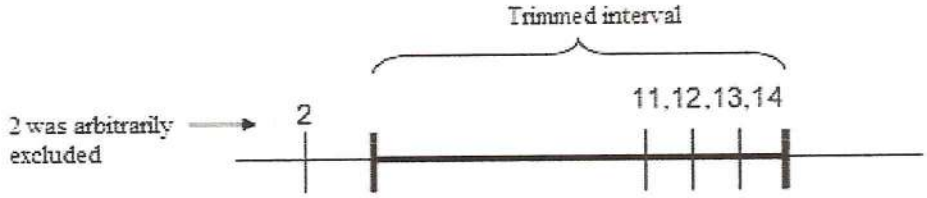


Figure 10: The Trimmed mean

B- The quadratic mean

1- The simple quadratic mean

Example: either the series of numbers. If we calculate the simple arithmetic mean we obtain Zero. $\{-4, -2, 0, 2, 4\}$

The quadratic mean is an average that finds applications when dealing with phenomena having a sinusoidal character with an alternation of positive values and negative values. It is widely used in electricity[14].

[14]Bardin, B.M. (2016). Descriptive statistics course. Cango-Brazzavill: DEUG HAL.

كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد
المجال العلمي

We calculate the simple square mean by adding the square of all the values in the series and taking the square root of the total.

Let's take the previous example:

$$Q = \sqrt{\frac{(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2}{5}} = \sqrt{\frac{16 + 4 + 0 + 4 + 16}{5}} = \sqrt{\frac{40}{5}} = \sqrt{8} \simeq 2{,}83$$

In general :

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

2- The quadratic mean for grouped data

Example :

Consider the following Table 3:

Xi	Ni	x_i^2	$ni.x_i^2$
25	10	625	6250
8	16	64	1024
4	25	16	400
12	20	144	2880

Table 3: Calculation of the quadratic mean for grouped data

The weighted root mean square formula:

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(n_i \cdot x_i^2)}$$

Applying this formula in another example:

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(n_i \cdot x_i^2)} = \sqrt{\frac{10554}{71}} \simeq 12{,}1921$$

NB: when the values are grouped into classes, it is necessary to calculate the class centers and apply the formula above, replacing xi with here.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

C- The geometric mean

1- The simple geometric mean

The geometric mean of n positive values xi is the nth root of the product of these values.

The geometric average is an instrument allowing the calculation of average rates, particularly annual average rates. Its use only makes sense if the values have a multiplicative nature.[15].

$$G = \left[\prod_{i=1}^{n} x_i \right]^{\frac{1}{n}}$$

Example

Or the series of numbers. The geometric mean of this series is equal to: $\{8, 5, 9, 13, 25\}$

$$G = \left[8 \times 5 \times 9 \times 13 \times 25\right]^{\frac{1}{5}} = \sqrt[5]{117000} \simeq 10{,}32$$

2- The geometric mean for grouped data

The general formula for the weighted geometric mean is:

$$G = \left[\prod_{i=1}^{n} x_i^{n_i} \right]^{\frac{1}{N}}$$

Where: a series of numbers and the corresponding numbers. $\{X_1, X_2, X_3 \ldots\ldots X_h\}\{n_1, n_2, n_3 \ldots\ldots n_h\}$

Using the data from Table 3:

We calculate the geometric mean for grouped data as follows:

[15]Bardin, B.M. (2016). Descriptive statistics course. Cango-Brazzavill: DEUG HAL.

$$G = \left[\prod_{i=1}^{n} x_i^{n_i}\right]^{\frac{1}{n}} = \left[25^{10} \times 8^{16} \times 4^{25} \times 12^{20}\right]^{\frac{1}{71}}$$

$$\ln G = \frac{1}{71}[10\ln 25 + 16\ln 8 + 25\ln 4 + 20\ln 12]$$

$$\ln G = \frac{1}{71}[32,1888 + 32,2711 + 34,6574 + 49,6981]$$

$$\ln G = \frac{149,815}{71} \simeq 2,1100704$$

$$G = e^{2,1100704} = 8,2488$$

D- The harmonic mean

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the values. The harmonic mean can be used when it is possible to assign a real meaning to the reciprocals of the data in particular for exchange rates, equipment rates, and speeds. It is particularly used in the calculation of indices[16]

1- The simple harmonic mean

Or a series of numbers. The formula for the simple harmonic mean is:$\{X_1, X_2, X_3 \dots \dots \dots X_n\}$

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example: either the series of numbers: .$\{8, 5, 9, 13, 25\}$

The harmonic mean of this series:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{5}{\frac{1}{8} + \frac{1}{5} + \frac{1}{9} + \frac{1}{13} + \frac{1}{25}} = \frac{5}{0.5530342} \simeq 9,04$$

2- The harmonic mean for grouped data

Or a series of numbers and the corresponding numbers. $\{X_1, X_2, X_3 \dots \dots \dots X_h\}\{n_1, n_2, n_3 \dots \dots \dots n_h\}$

The harmonic mean formula:

[16]- Goldfarb, b., & Pardoux, c. (2013). Introduction to statistical method, statistics and probability. Paris: 7th edition, Dunod.

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Using the example from Table 3, the harmonic mean is:

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{71}{\frac{10}{25} + \frac{16}{8} + \frac{25}{4} + \frac{20}{12}} = \frac{71}{0,4 + 2 + 6,25 + 1,66667} = \frac{71}{10,3167} = 6,882$$

Solved exercise

Consider the following series:{2, 5, 11, 18}

1-Calculate the arithmetic mean, the quadratic mean, the geometric mean and the harmonic mean.

2- What do you notice?

$\bar{X} = 9$

G= $\sqrt[4]{2x5x11x18} = 6.67$

H = $\frac{4}{\frac{1}{2}+\frac{1}{5}+\frac{1}{11}+\frac{1}{18}} = 4.72$

Q = $\sqrt{\frac{1}{4}(2^2 + 5^2 + 11^2 + 18^2)} = 10.88$

We note that :

Xmin≤ H≤ G≤ ≤Q ≤Xmax $\bar{X}$

II-The median

The median of a series is the value that divides this series, previously classified, into two series with equal numbers. In the first series, we find the values lower than the median. In the second series, we find the values greater than the median.

The median is only calculated for quantitative data and its method of calculation depends on the type of data. We will distinguish four cases: ungrouped series whose number is odd and where no value is repeated,

– Series grouped by values,

– Series grouped by value classes.

-Ungrouped series whose number is even and where no value is repeated.

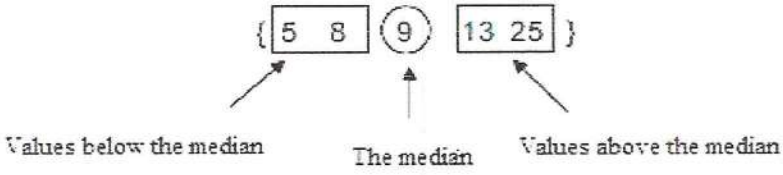
A – Calculation of the median: odd number and no value is repeated

Consider the series of following numbers: {8, 5, 9, 13, 25}

To find the median, we must:

a) Classify the series in ascending order of values

b) Locate the value which divides the total number into two equal sub numbers by applying the formula (n+1)/2, that is to say here (5+1)/2=3. The third value in the series is 9[17].



The total effective is divided into equal parts.

B – Calculation of the median: even number and no value is repeated

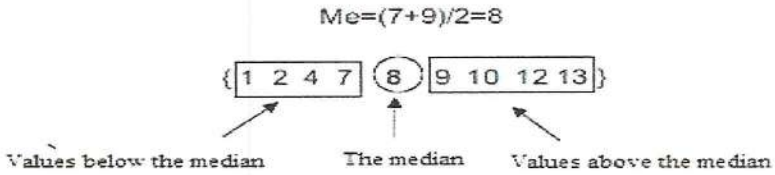
When the number is even, it is necessary to calculate the median of the series.

Example: either the following series:{13, 1, 9, 10, 2, 4, 12, 7}

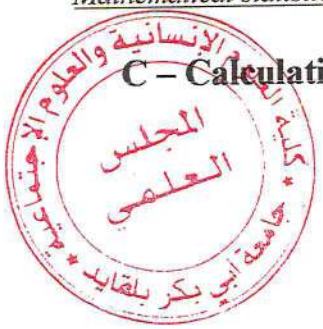
To find the median, you must:

a) Classify the series in ascending order of values

b) Apply the formula (n+1)/2, that is to say here (8+1)/2=4.5. This tells us that the middle interval is made up of the 4th and 5th values. The median is therefore equal to the simple arithmetic mean of these two values:



[17]Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.



C – Calculation of the median: numbers grouped by values

x_i	n_i	f_i	$F(x)$	$N(x)$
2	2	0,066	0,066	2
8	3	0,1	0,167	5
9	4	0,133	0,3	9
10	4	0,133	0,433	13
11 ◄	5 ◄	0,167 ◄	0,6 ◄	18 ◄
12	3	0,1	0,7	21
13	6	0,2	0,9	27
15	1	0,033	0,933	28
18	2	0,067	1	30

Median =11

0,5

n/2=15

Table 4: Calculation of the median when data is grouped by values

We propose here the method described by Bernard PY[18]:

To determine the median, we locate 0.5 in the column of cumulative frequencies F(x) or n/2 in the column of cumulative numbers N(x).

- We then choose the value F(x) equal to or immediately greater than 0.5 (or the value N(x) equal to or immediately greater than n/2) and we follow the direction of the arrows as indicated in Table 4. In our example, there is no value F(x) equal to 0.5, the value immediately greater than 0.5 is 0.6 (and the value immediately greater than n/2=30/2=15 is 18).
- So, by following the arrows, we go back to the value which corresponds to the median, i.e. 11. We then notice that the median does not separate the number into two equal parts. Indeed, 13 values are less than 11 (i.e. 43.3% of the frequencies) and 12 values are greater than 11 (i.e. 40% of the frequencies).

[18]- In his work: Descriptive Statistics, Éditions Economica, page 76.

المجلس العلمي

Graphical determination of the median

Figure 11 illustrates the determination of the median from N(xi), the cumulative curve of the numbers. This "stepped" curve has as its ordinate the numbers whose value is strictly less than xi.

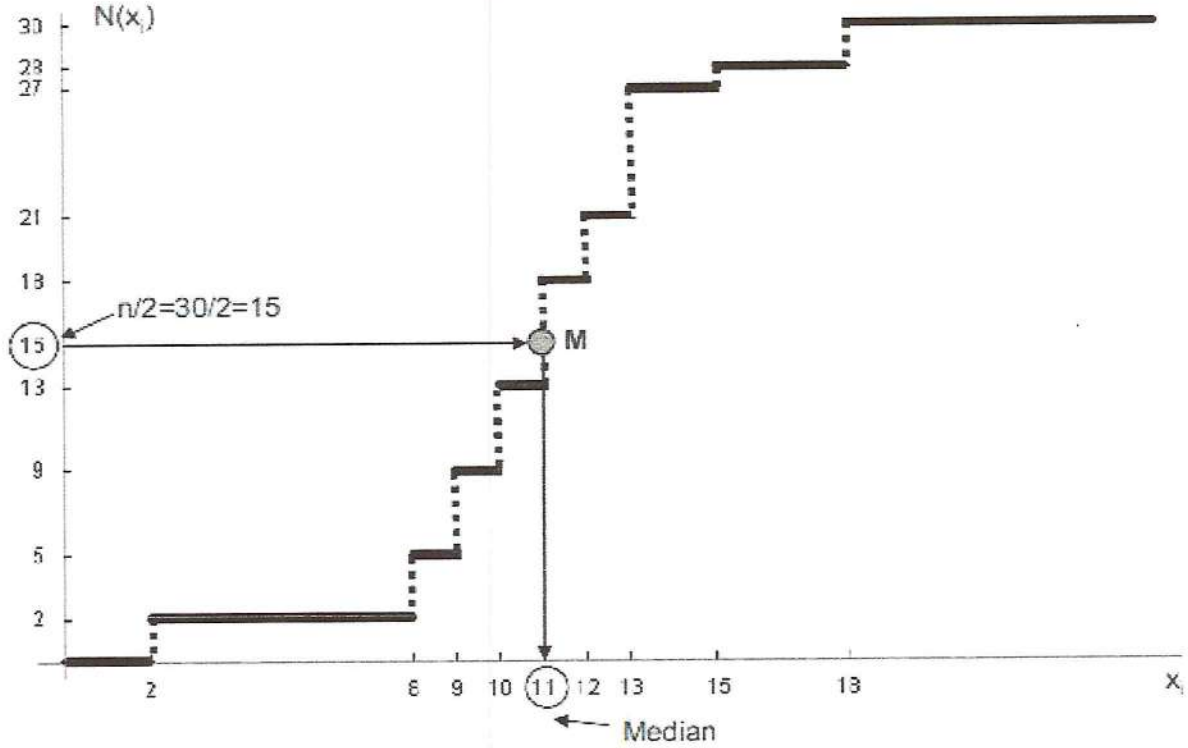


Figure 11: Graphical determination of the median from the cumulative curve

To find the median, we must locate n/2=30/2=15 on the y-axis, then draw a horizontal arrow to point M. Once at point M, you must draw a vertical arrow in the direction of the abscissa. We then read the value of the median which, in our example, is equal to 11[19]

D – Calculation of the median: numbers grouped by value classes

In this case, we apply the following formula:

$$M_e = x^1 + a_i \left[\frac{\frac{n}{2} - N(x_{i-1})}{n_i} \right]$$

[19]Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

x_{i1} – Lower limit of the median class

$N(x_{i-1})$ – Strictly lower cumulative effective than xi

x_i – Median class a_i – Amplitude of median class

Example: In Table 5, the values of the variable x are grouped by classes of values of equal amplitudes.

x_i	n_i	$N(x_i)$
[0-5[	2	2
[5-10[	7	9
[10-15[	18	27
[15-20[	3	30

Table 5: Values grouped by classes of equal amplitudes[20]

$$M_e = x_{i1} + a_i \times \left[\frac{\frac{n}{2} - N(x_{i-1})}{n_i} \right] = 10 + 5 \times \left[\frac{15-9}{18} \right] = 11.666$$

III-The mode

The mode of a series is the most frequent value in that series. A series can have several modes. The calculation depends on the data type.

A – Mode calculation: simple series, no value is repeated

Consider the series of following numbers: {8, 5, 9, 13, 25}

There is no mode because each value is only repeated once (the frequency of each value is 1).

B – Calculation of mode: numbers grouped by values

Example: Consider the series of numbers {8, 8, 8, 7, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6}

The most frequent value in this series is: 4. The bar chart (Figure 9) clearly shows the modal value found.

[20]- The procedure is the same if the classes are of unequal amplitudes.

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

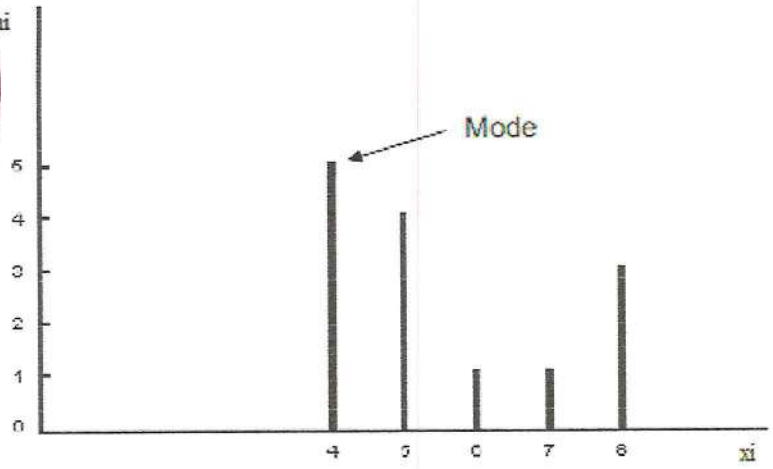


Figure 12: Determination of mode

C – Calculation of the mode: numbers grouped by classes of equal amplitudes

Example: we take the example from Table 5

x_i	n_i	$N(x_i)$
[0-5[	2	2
[5-10[	7	9
[10-15[	18	27
[15-20[	3	30

The data are grouped by classes of equal amplitudes.

The applied mode formula:

$$\text{Mode} = x_i^{inf} + a\frac{d_1}{d_1 + d_2}$$

x_i^{inf} = Borne inférieure de la classe modale a = Amplitude de classe

$d_1 = n_i - n_{i-1}$ et $d_2 = n_i - n_{i-1}$

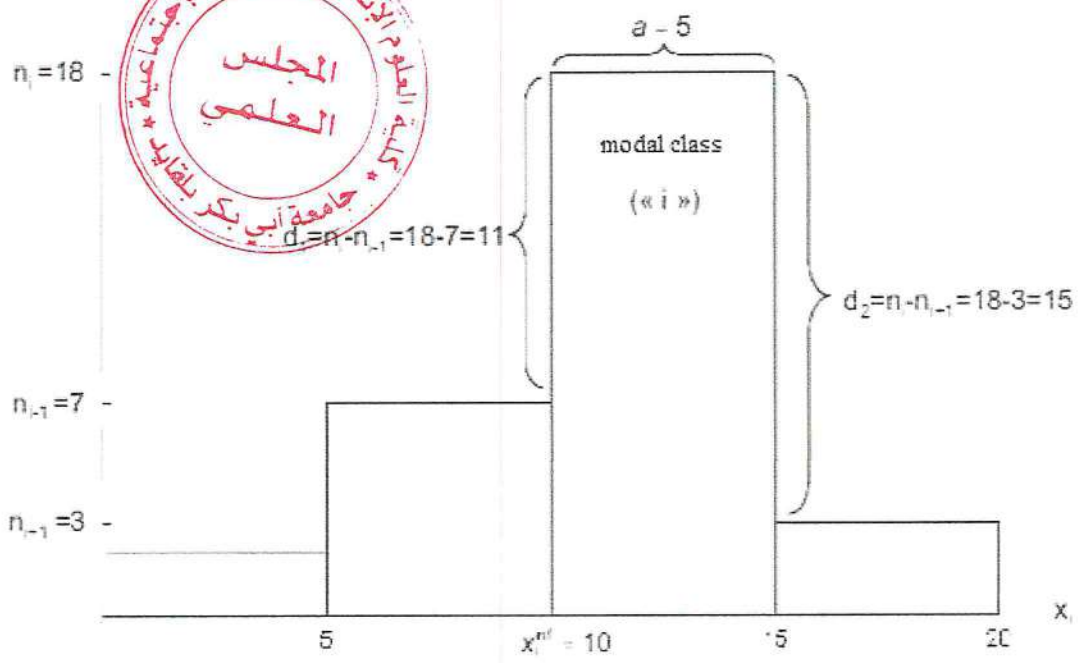


Figure 13: Calculation of the mode (classes of equal amplitude)

$$\text{Mode} = x_i^1 + a\frac{d_1}{d_1 + d_2} = 10 + 5 \times \left(\frac{11}{11+15}\right) = 12.115$$

D – Calculation of the mode: numbers grouped by classes of unequal amplitudes

Example: consider the following table presenting data by unequal classes.

x_i	n_i	a_i	$h_i = \frac{n_i}{a_i}$
[0-10[	9	10	0,9
[10-12[	9	2	4,5
[12-20[	12	8	1,5

Table 6: Values grouped by class of unequal amplitudes

In this case, to calculate the mode, it is necessary to apply the previous formula, but the definition of d1 and d2 changes, because it is necessary to replace the effective ni with the corrected amplitudes hi=ni /ai. We therefore have, the following about Figure 14 that represents the histogram corresponding to Table 6 (on the ordinate we have the ni /ai and on the abscissa we have the classes of values of unequal amplitudes).

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

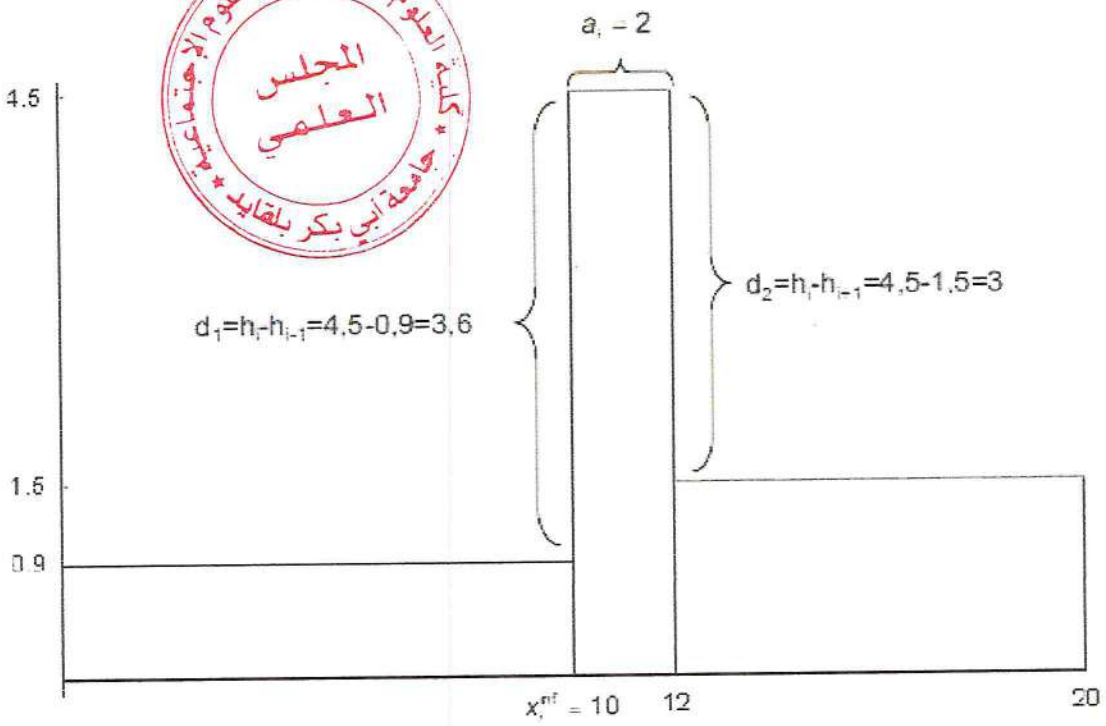


Figure 14: Calculation of the mode (classes of unequal amplitudes)

$h_{i-1}=n_{i-1}/a_{i-1}=9/10= 0,9$ $h_i=n_i/a_i=9/2=4,5$ $h_{i-1}=n_{i-1}/a_{i-1}=12/8=1,5$

$d_1=h_i-h_{i-1} = 4,5-0,9=3,6$ $d_2=h_i-h_{i-1} =4,5-1,5=3$

et $$\text{Mode} = x_1^1 + a_i \frac{d_1}{d_1+d_2} = 10 + 2 \times \left(\frac{3,6}{3,6+3} \right) = 11,09$$

IV-Characterizing the shape of a distribution using the arithmetic mean, median, and mode

A- Perfectly symmetrical distribution

Consider the following table:

x_i	1	2	3	4	5
n_i	2	4	5	4	2

Table 7: Perfectly symmetrical distribution

The calculation of the three indices reveals that:

$\overline{X}$= Me= Mo= 3

The distribution is perfectly symmetrical as illustrated by the bar chart in Figure 15.

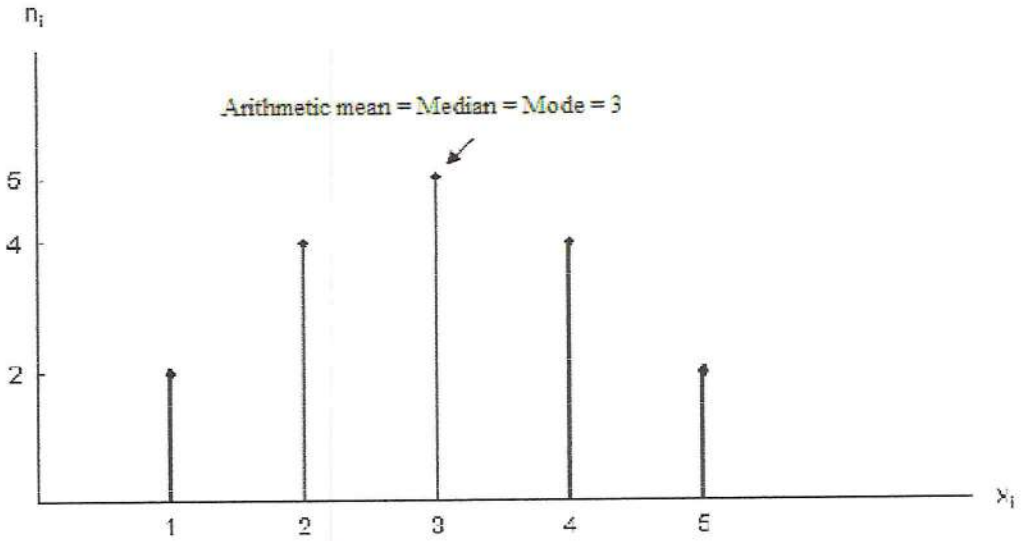


Figure 15: Perfectly symmetrical distribution[21]

[21]- Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.

المجلس العلمي

B - Distribution spread to the right

Consider the following table:

x_i	1	2	3	4	5
n_i	10	8	6	4	2

Table 8: Spread distribution on the right

We note that :$\bar{X}$= 2.33> Me= 2> Mo= 1

The distribution is spread to the right as shown in the bar chart in Figure 16.

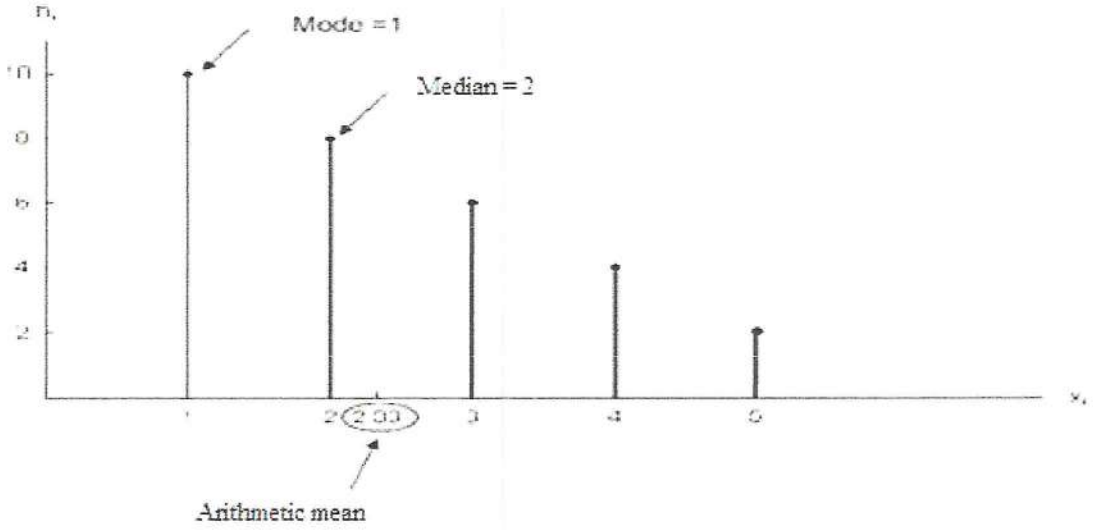


Figure 16: Spread distribution on the right

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

C - Distribution spread to the left

Consider the following table:

x_i	1	2	3	4	5
n_i	2	4	6	8	10

Table 9: Spread distribution on the left

The calculation of the three indices reveals that: $\overline{X} = 3.7 < Me < MO = 5$

The distribution is spread out to the left, as shown in the bar chart (Figure 17).

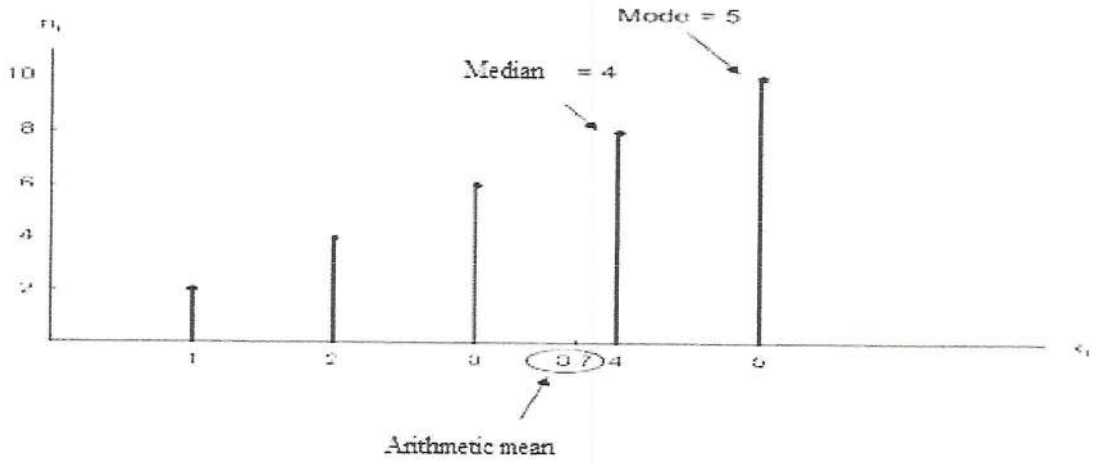


Figure 17: Spread distribution on the left

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

Course 5: Dispersion

The central tendency characteristics allow us to have an order of magnitude of the series but do not tell us about the internal structure of the series.

1- The variation interval

The interval, or "spread" or range is the difference between the largest value and the smallest value of the variable.

Student A: {8, 9, 10, 11, 12} Student B: {2, 4, 16, 18}

The range of A grades is: 12- 8= 4

The range of B grades is: 18-2= 16

We notice that the B notes are more dispersed than A.

This characteristic is the simplest but also the least significant. Its meaning is clear and its calculation is extremely fast. These advantages make it frequently used in industrial manufacturing control rather than carrying out complex calculations in the workshop.

2- The interquartile range

The interquartile range is a measure of variation that is not influenced by extreme values, unlike the variation interval. The interquartile range measures the extent of the middle 50% of values in a classified data series.

Quartiles: the list of N data is arranged in ascending order.

The first quartile: is the smallest Q1 data in the list such that at least a quarter of the data in the list are $\leq$ Q1.

We obtain it by dividing n/4.

The third quartile: is the smallest Q3 data in the list such that at least three-quarters of the data in the list are $\leq$ Q3.

We obtain it by calculating $n\frac{3}{4}$.

Example :

Consider the following series which represents the students' grades:

{3; 3; 4; 10; 12; 13; 14; 18; 19; 20}

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

The first quartile Q1 = 10/4 = 2.5≅3

We take the third value. 3 students had grades ≤ 4.

Q1=4

The third quartile Q3= 18. $Q3=\frac{10x3}{4}= = 7.5.\frac{30}{4}$

Q3 = 7.5≅8. We take the eighth value. 8 students had grades ≤ 18.

The second quartile corresponds to the median

NB: if the data is repeated, the ncc (increasing cumulative numbers) is calculated.

$Q1= \frac{ncc}{4}$ $Q2= \frac{ncc}{2}$ $Q3= \frac{3ncc}{4}$

2.1-The boxplot

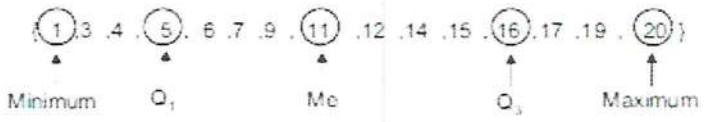
The box plot, from the English "Box and Whiskers", sometimes also called "box plot", is a graph that summarizes the dispersion of a series from 5 values: the minimum value and the maximum value (these are the "whiskers"), the interquartile range (designated by its two values Q1 and Q3) and the median (these last three values constituting the "box").

Example: or the next series of numbers, where no value is repeated. The number of digits is odd.

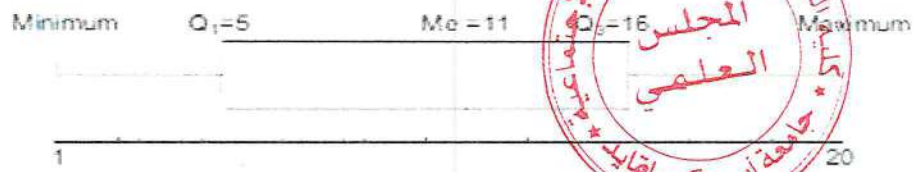
{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6, 5}

We know that Me = 11, Q = 5 and Q3 = 16 having calculated them in example 1 of section 2 of this chapter. As for the minimum and maximum values, they are respectively equal to 4 and 20. Let's classify the series in ascending order to better show the different values involved in the boxplot.

{1,3,4,5,6,7,9,11,12,14,15,16,17,19,20}



The corresponding "box plot" graph is therefore: (Figure 18)



كلية العلوم الإنسانية والعلوم الإسلامية
المجلس العلمي
جامعة أبي بكر بلقايد

Figure 18: Boxplot

3-Variance, standard deviation and coefficient of variation

Variance, standard deviation and coefficient of variation are the most frequently used indicators to measure the dispersion of a series. These indicators provide information on the dispersion of the data around the average[22]

The more the data is concentrated around the average, the lower the values of these three indicators are. Conversely, the more the data is dispersed around the average, the higher these three indicators are.

3.1-The variance

Consider a series of values of a variable X:$\{x_1, x_2, x_3 \ldots\ldots x_k\}$

Or the associated effective. The variance of this series is written:$\{n_1, n_2, n_3 \ldots\ldots n_k\}$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 \qquad \text{population}$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 \qquad \text{sample}$$

The variance of the series will be written:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2$$

Noticed

When the data is grouped by class, it is the class center ci which replaces xi in the previous formula.

[22]Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.

Example[23]

Consider the following series: {2, 5, 7, 1, 9, 13, 6, 15, 8, 16}

Calculate the variance of this series.

كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

x_i	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert^2$
2	-6.2	38.44
5	-3.2	10.24
7	-1.2	1.44
1	-7.2	51.84
9	0.8	0.64
13	4.8	23.04
6	-2.2	4.84
15	6.8	46.24
8	-0.2	0.04
16	7.8	60.84

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 8.2$ $\qquad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{237.6}{10} = 23.76$

- If the numbers are repeated by values:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$$

Example: Consider the following table:

x_i	2	6	9	11	15
n_i	5	9	4	3	5

Calculate the variance.

x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
2	5	10	4	20
6	9	54	36	324
9	4	36	81	324
11	3	33	121	363
15	5	75	225	1125
	26	208		2156

Totals

$$\bar{x} = \frac{1}{26}\sum_{i=1}^{5} n_i x_i = \frac{208}{26} = 8$$

$$\sigma^2 = \frac{1}{26}\sum_{i=1}^{5} n_i x_i^2 - \bar{x}^2$$

$$\sigma^2 = \frac{1}{26}2156 - (8)^2$$

$$\sigma^2 = 82.9231 - 64 = 18.9231$$

[23] Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.

3.2-The standard deviation

The standard deviation is equal to the square root of the variance:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2}$$

Example :

Consider the following series: {2, 5, 7, 1, 9, 13, 6, 15, 8, 16}

The variance of this series is equal to:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{237.6}{10} = 23.76$$

The standard deviation is equal to:

$$\sigma = \sqrt{23.76} = 4.87$$

3.3-The coefficient of variation

$$CV = \left(\frac{\sigma}{\bar{x}}\right) \times 100$$

Example[24]: The number of employees increased from 200 in 1994 to 280 in 2004. We want to know if the dispersion of salaries has increased. To do this, we will calculate the coefficient of variation in 1994 and 2004.

Salary	Effective 1994	Effective 2004
1000-2000	40	56
2000-3000	70	118
3000-4000	80	92
4000-5000	5	10
5000-10000	5	4

[24]Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.

Salary	1994 n_i	c_i	$n_i c_i$	c_i^2	$n_i c_i^2$
1000-2000	40	1500	6000	2250000	90000000
2000-3000	70	2500	175000	6250000	437500000
3000-4000	80	3500	280000	12250000	980000000
4000-5000	5	4500	22500	20250000	101250000
5000-10000	5	7500	37500	56250000	281250000
	200		**575000**		**1890000000**

Totals

كلية العلوم الإنسانية والعلوم الإسلامية * جامعة أبي بكر بلقايد * المجلس العلمي

To calculate the coefficient of variation, you must first calculate the arithmetic mean and the standard deviation:

$$\overline{x} = \frac{1}{200}\sum_{i=1}^{5} n_i c_i = \frac{575000}{200} = 2875$$

$$\sigma = \sqrt{\frac{1890000000}{200} - (2875)^2} = 1088.29$$

The coefficient of variation of wages for the year 1994 is equal to:

$$CV_{1994} = \left(\frac{\sigma}{\overline{x}}\right) \times 100 = \frac{1088.29}{2875} \times 100 = 37.8536$$

المجلس العلمي
جامعة أبي بكر بلقايد

Chapter II: Inferential statistics

Course 6: Hypothesis testing (General)

1-Principle of a hypothesis test

Hypothesis testing is another important aspect of statistical inference.

A hypothesis test is an inference process making it possible to control (accept or reject) from the study of one or more random samples, and the validity of hypotheses relating to one or more populations.[25].

Statistical inference methods allow us to determine, with a given probability, whether the differences observed in the samples can be attributable to chance or whether they are large enough to mean that the samples come from presumably different populations.

Statistical testing is a way of making an informed comparison based on the observation of a sample. Everything being based on probability theory, it will not be possible to know whether the decision taken following a hypothesis test is the right one, but it is however possible to control some of the errors and keep them within limits. acceptable rates in the context[26].

To decide whether the formulated hypothesis is supported or not by the observations, we need a method that allows us to conclude whether the difference observed between the value of the statistic obtained from the sample and that of the parameter specified in the hypothesis is too significant. to be solely attributable to random sampling.

2-Vocabulary

- Statistical hypothesis: a statistical hypothesis is a statement (an assertion) concerning the characteristics (parameter values, shape of the distribution of observations) of a population.

---

[25]- Ruch, J.-J. (2012-2013). Statistics: Hypothesis testing.

[26]Houde, L. (2014). Hypothesis Testing-Quantitative Analysis of Management Problems. University of Quebec at Trois-Rivières, Department of Mathematics and Computer Science.

المجلس العلمي جامعة أبي بكر بلقايد

-Hypothesis testing: or statistical test is an approach that aims to provide a decision rule allowing, based on sample results, to choose between two statistical hypotheses. The statistical hypotheses that are considered a priori are called the null hypothesis and the alternative hypothesis.

-Null hypothesis (H0): this is the hypothesis according to which a parameter of the population is fixed a priori at a particular value is called the null hypothesis and is denoted H0.

-The counter-hypothesis (alternative hypothesis) (H1 or Ha): this is the hypothesis that we would like to demonstrate. The more research that supports the alternative hypothesis, the more reason there is to believe that it is true. It is said in science that the more a theory stands up to the test of facts, the stronger it is.

-Simple hypotheses: comparison with a single value ($\mu=50$).

-Compound hypotheses: comparison with several values ($\mu < 50$).

-Error of the first kind: the decision to reject H0 knowing that H0 is true.

-Significance threshold (α): probability of committing type 1 error. It is the risk agreed in advance to wrongly reject the null hypothesis H0 when it is true.

-Error of the second type: accept H0 knowing that H0 is false. We will use an estimator that minimizes the second type of error, but this error is more difficult to quantify.

-Power of a test: probability of not committing the type 2 error, $1-\beta$ where $\beta=P$ (accepting H0 | H0 is false).

- Critical region: set of statistical values calculated on the samples for which H0 is rejected.

-Critical value: junction point of the acceptance region and the critical region.

3-Types of tests

Depending on the hypothesis tested, several types of tests can be carried out:

-Tests intended to check whether a sample can be considered as extracted from a given population, about a parameter such as the mean or the observed frequency (conformity tests), or about its observed distribution (tests adjustment). In this case, the theoretical law of the parameter is known at the population level.

-Tests intended to compare several populations using an equivalent number of samples (equality or homogeneity tests) are the most commonly used. In this case the theoretical law of the parameter is unknown at the population level. We can add to this category the independence test which seeks to test the independence between two characteristics, generally qualitative[27].

4- Steps of a hypothesis test

1- Formulate the null hypothesis H0 and the alternative hypothesis H1.

2- Set in advance (before carrying out the survey) the significance threshold α that is to say, specify the risk of wrongly rejecting a true hypothesis H0.

3- Specify the conditions for applying the test, specifying or not the shape of the sampled population, indicating whether we are in the presence of a large sample, whether the population variance is known or unknown, etc.

4- Specify the appropriate statistic for the test and set the reduced deviation.

5- Adopt a decision rule that will lead to the rejection or non-rejection of H0 at the chosen threshold α. This decision rule is defined based on the critical values of the reduced deviation.

6- Calculate the numerical value of the reduced deviation, a value deduced from the sample results.

7- Decision and conclusion. Compare the numerical value obtained for the reduced deviation with the decision rule adopted in Step 5. Decide between the two hypotheses formulated in Step 1 and conclude.

5- Formulation of hypotheses H0 and H1 and type of test

To guide the discussion, suppose we assert that the averageμ for example (a parameter) of a population is equal to a particular value μ0.

5.1- One Tailed or two-tailed test

The nature of H0 determines the way of formulating H1 and consequently the unilateral or bilateral nature of the test.

[27]- Mouchiroud, D. (2003). Mathematics: Tools for Biology. Deug SV – UCBL.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

5.2-Two Tailed-test

When we are interested in the change in the mean μ for example in one of the other directions (μ>μ0 or μ<μ0), we opt for a bilateral test.

The hypotheses H0 and H1 are then:

H0: μ= μ0

H1: μ≠μ0

We can schematize the regions of rejection and non-rejection of H0 as follows:

$$H_0 : \mu_1 = \mu_0$$
$$H_1 : \mu_1 \neq \mu_0$$

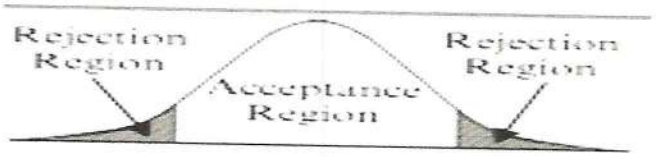


Figure 19: Two-Tailed test

If, following the results of the sample, the value of the statistic Clies in the interval $\bar{x}c1 \leq \bar{x} \leq \bar{x}c2$, we cannot reject H0 at the chosen significance threshold.

If $\bar{x} > \bar{x}c2$ or $\bar{x} < \bar{x}c1$, we reject H0 and favor H1.

5.3-One-Tailed-test

When we are interested in the change in the mean μ for example in a single direction, we opt for a one-sided test. The hypotheses are as follows if we are interested in a change on the left side:

H0: μ= μ0

H1: μ<μ0

Left tail. We will favor H1 and $\bar{x} < \bar{x}c$ (rejection of H0).

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

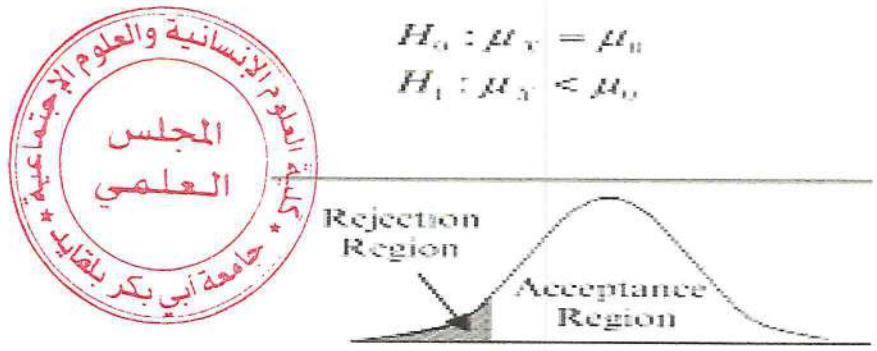


Figure 20: One tailed-test (Left tail)

The hypotheses can also be stated as follows if we are interested in a change in the other direction (right side):

H0: $\mu = \mu 0$

H1: $\mu > \mu 0$

Right-tail, We reject H0 if we then consider H1 to be probable. $\bar{x} > \bar{x}c$.

Figure 21: One-Tailed test (Right tail)

Remarks :

- A one-sided test always has only one critical value.
- Regardless of the type of test, the null hypothesis always has the strictly equal sign and specifies the particular value of the parameter.
- Hypothesis H1 is formulated by choosing one or the other of the three forms mentioned. We will choose the one most relevant to the practical situation analyzed.
- In most hypothesis tests, the sign in the hypothesis H1 denotes which direction the critical region or rejection region of H0 is located.

المجلس العلمي

5.4- Choice of a statistical test

The choice depends on the nature of the data, the type of hypothesis that we wish to control, the statements that we can make concerning the nature of the populations studied (normality, equality of variances) and other criteria that we will specify. .

A statistical test or a statistic is a function of random variables representing the sample whose numerical value obtained for the sample considered makes it possible to distinguish between true H0 and false H0.

The conclusion that will be deduced from the results of the sample will have a probabilistic character: we can only decide by being aware that there is a certain risk that it will be erroneous. This risk is given to us by the significance threshold of the test.

5.5- Risk of error of the first type α

The risk of errorαis the probability that the experimental or calculated value of the S statistic belongs to the critical region if H0 is true. In this case H0 is rejected and H1 is considered true.

The risk α of the first kind is that of rejecting H0 when it is true

$$\alpha = P(\text{ reject H0 / H0 true})$$

or accept H1 even though it is false

$$\alpha = P(\text{ accept H1 / H1 false})$$

5.6- Risk of error of the second type β

The risk of error β is the probability that the experimental or calculated value of the statistic does not belong to the critical region if H1 is true. In this case, H0 is accepted and H1 is considered false.

The risk β of the second kind is that of accepting H0 even though it is false

$\beta = P(\text{accept H0 / H0 false})$ or $P(\text{accept H0 / H1 true})$

or reject H1 when it is true

$$\beta = P(\text{ reject H1 / H1 true})$$

5.7- The power of a test (1 -β)

The tests are not done to "demonstrate" H0 but to "reject" H0. The ability of a test to reject H0 when it is false constitutes the power of the test.

The power of a test is: 1 -β= P(reject H0 / H0 false) = P(accept H1/H1 true).

كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

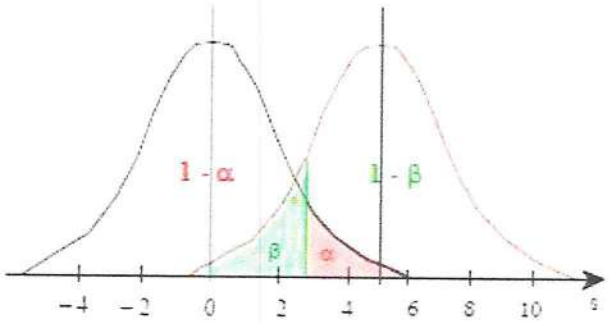


Figure 21: The relationship between the two risks of errors

Noticed :

The power of a test depends on the nature of H1; a one-sided test is more powerful than a two-sided test.

The power of a test increases with sample size N studied at the value of α constant. The power of a test decreases when α decreases[28].

The different situations that can be encountered in the context of hypothesis testing are summarized in Figure 23:

[28]Mouchiroud, D. (2003). Mathematics: Tools for Biology. Deug SV – UCBL.

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

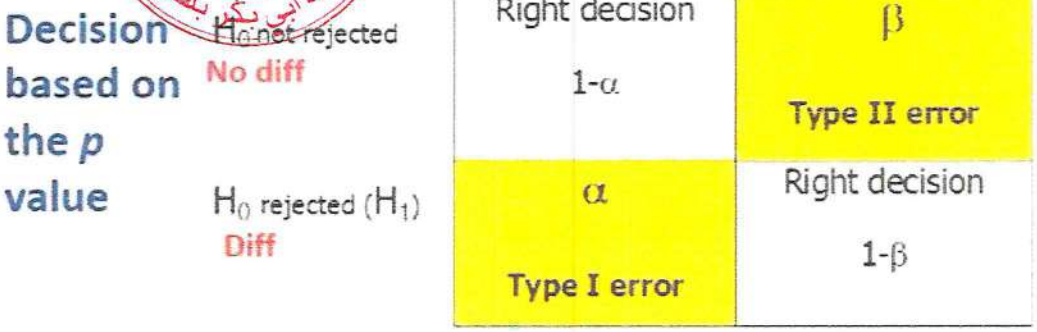


Figure 23: Type of errors

المجلس العلمي
كلية العلوم الإجتماعية
جامعة أبي بكر بلقايد

Course 7: Conformity tests

1- Conformity tests

Conformity tests are intended to check whether a sample can be considered as extracted from a given population or representative of this population, concerning a parameter such as the mean, variance, or observed frequency. This implies that the theoretical law of the parameter is known at the population level.[29].

1.1-One mean test

Test For	*Null Hypothesis (H_0)*	*Test Statistic*	*Distribution*	*Use When*
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n > 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown

Figure 24: One mean test

The aim is to know if a sample of $\overline{x}$ mean, "μ estimated", belongs to a known reference population of hope μ_0 (H0 true) and does not differ from μ_0 that by sampling fluctuations or belongs to another unknown population of expectation μ (H1 true).

To test this hypothesis, there are two statistics: the variance σ^2 of the reference population is known (testε) or this variance is unknown and must be estimated (T test).

[29]Mouchiroud, D. (2003). Mathematics: Tools for Biology. Deug SV – UCBL.

كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

1.1.1- Known population variance (known standard deviation)

1.1.1.1 -Test statistics

Either $\bar{x}$ the sampling distribution of the mean in the unknown population follows a normal law such that:

$$\bar{X} \rightarrow N(\mu, \sqrt{\frac{\sigma^2}{n}}).$$

The statistic studied is the difference: S =μ0 whose probability distribution is as follows:

$S \rightarrow N(0, \sqrt{\frac{\sigma^2}{n}})$ with under H0, E(S) = 0 and

$$V(S) = \frac{\sigma^2}{n}$$

We can establish thanks to the central limit theorem the reduced centered variable Z such that:

$$Z = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

Sous $H_0 : \mu = \mu_0$ avec σ^2 connue

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$ follows a reduced centered normal law N(0,1)

1.1.1.2 –Application and decision

The hypothesis tested is as follows:

H0: μ=μ0 against H1:μ ≠ μ0

A value z of the random variable Z is calculated:

$$z = \frac{|\bar{x} - \mu_0|}{\sqrt{\frac{\sigma^2}{n}}}$$ also noted ε_{obs}

Z calculated (εobs) is compared with the value ε threshold read from the table of the reduced centered normal law for a fixed error risk α (Decision rule 1).

-if εobs > εthreshold the hypothesis H0 is rejected at the risk of error α: the sample belongs to a population of expectation μ and is not representative of the reference population of expectation μ0.

- if εobs ≤ εthreshold the hypothesis H0 is accepted: the sample is representative of the reference population with expectation μ0.

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

Solved exercises

Exercise 1 (Two-tailed test for one mean)

A company sells bags of sugar. The filling is adjusted so that the average weight is 1 kg per bag. To test the quality of the filling process, 10 bags of sugar were taken. It turned out that the average weight of these bags was 996g with a standard deviation of 6g. it is specified that the weight of the sugar is normally distributed.

At the 95% confidence level, can we consider that there is a difference between the actual weight and the indicated weight?

Solution :

This is a Two-tailed test for one mean.

1st method:

1st step: Formulation of hypotheses

H0:μ= 1000g

H1:μ ≠1000g

2nd step: Calculation of the decision variable

$$Z=\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Z= -2.10

3rd step: The decision rule:

$$Z= -2.10 \notin \left[-Z\alpha_{/2};\ +Z\alpha_{/2}\right]$$

Therefore, we reject H0.

2nd method:

1st step: Formulation of hypotheses

H0:μ= 1000g

H1:μ ≠1000g

2nd step:

Calculation of the two limits of the confidence interval: (the two critical values)

$$\bar{x}_{c1}= \mu_0 - Z\alpha_{/2}\ \frac{\sigma}{\sqrt{n}}$$

$$\bar{x}_{c1} = 996.28$$

$\bar{x}_{c2} = \mu_0 + Z\alpha/2 \frac{\sigma}{\sqrt{n}}$

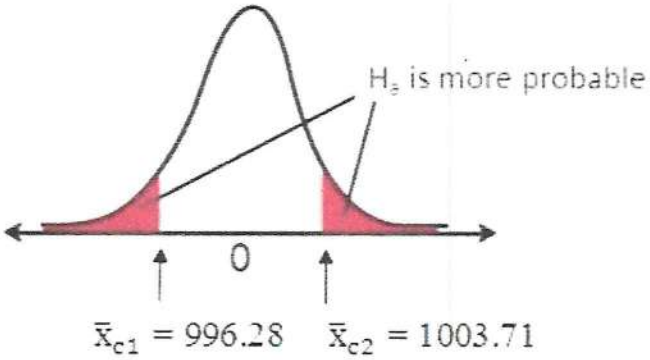
$\bar{x}_{c2} = 1003.71$

3rd step: The decision rule

We note that

$\bar{x} = 996g \notin [996,28; 1003,71]$

We reject H0.



Exercise 2 (Right-tailed test for one mean)

A military recruitment center knows from experience that the weight of its recruits is normally distributed with an average of 80 kg and a standard deviation of 10 kg. The center wants to test at a significance level of 1%, if the average weight of recruits in the current army exceeds 80kg, then it takes a random sample of 25 recruits and finds that their average weight is 85kg. Will the center accept or reject this hypothesis?

Solution :

This is a right-tailed test for one mean.

1st method:

1st step: Formulation of hypotheses

H0:$\mu = \mu 0 = 80$

H1:$\mu > 80$

2nd step: Calculation of the decision variable

$$Z = VDR = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{85 - 80}{\frac{10}{\sqrt{25}}}$$

$Z = 2.5$

3rd step: The decision rule:

$Z\alpha = 2.33$

$Z = 2.5 > 2.33$

So, we reject H0 and accept H1:$\mu > 80$, the weight of the recruits exceeds 80kg.

2nd method:

1st step: Formulation of hypotheses

H0: $\mu = \mu 0 = 80$

H1: $\mu > 80$

2nd step:

Calculation of the upper bound of the confidence interval:

$\bar{x}_c = \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} = 80 + 2.33.$

$$\bar{x}_c \frac{10}{\sqrt{25}}$$

$$\bar{x}_c = 84{,}66$$

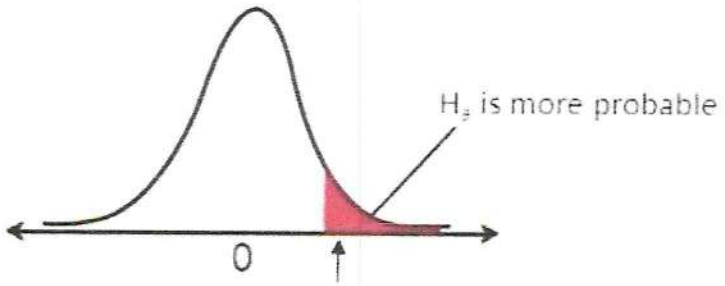
3rd step: The decision rule

We have :

$\bar{x} > \bar{x}_c$

$85 > 84.66$

So, we reject H0 and accept H1:$\mu > 80$, the weight of the recruits exceeds 80kg.



$\bar{x}_c = 84{,}6$

المجلس العلمي

1.1.2- Unknown population variance (unknown standard deviation)

1.1.2.1 -Test statistics

The procedure is the same as for the testε but the population variance is not known, it is estimated by:

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$

This is the point estimate of the variance.

We can establish, thanks to the central limit theorem, the reduced centered variable T such that:

$$T = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

Under H0: μ=μ0 with σ^2 unknown

$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$ follows a Student's law with n-1 degrees of freedom.

1.1.2.2 –Application and decision

The hypothesis tested is as follows:

H0: μ=μ0 against H1:μ ≠ μ0

A value t of the random variable T is calculated:

$$t = \frac{|\bar{x} - \mu_0|}{\sqrt{\frac{\hat{\sigma}^2}{n}}} = \frac{|\bar{x} - \mu_0|}{\sqrt{\frac{s^2}{n-1}}}$$

t calculated (tobs) is compared with the t threshold value read in Student's table for risk of erroαfixed and (n - 1) degrees of freedom.

• if tobs > t threshold value the hypothesis H0 is rejected at the risk of error α: the sample belongs to a population of expectation μ and is not representative of the reference population of expectation μ0.

• if tobs ≤ threshold value the hypothesis H0 is accepted: the sample is representative of the reference population with expectation μ0.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإسلامية
جامعة أبي بكر بلقايد

Solved exercises

Exercise 1 (Left-tailed test for one mean)

The amount of a landline telephone bill in the offices of a large company is a random variable with mean μ and variance σ^2 unknown. The company's management and accounting department decided to eliminate 30% of the telephone lines if μ is less than 20000 DA. Before making a decision, we randomly chose a sample of 49 invoices which revealed an average amount of 22,000 DA, with a standard deviation of 8,000 DA.

At the level $\alpha = 0.05$, can we conclude that there is sufficient evidence of 30% telephone lines?

Solution :

This is a Left-tailed test for one mean.

1st method:

1st step: Formulation of hypotheses

H0:μ=μ0= 20000

H1:μ <20000

2nd step: Calculation of the decision variable

$Z = VDR = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

$Z = \frac{22000 - 20000}{\frac{8000}{\sqrt{49}}}$

Z=1.75

3rd step: The decision rule:

$-Z\alpha = -1.645$

$Z > -Z\alpha$

1.75>1.645

So, accept H0 and reject H1:μ <20000.

2nd method:

1st step: Formulation of hypotheses

H0:μ=μ0= 20000

H1:μ <20000

كلية العلوم الإنسانية والعلوم الإجتماعية * المجلس العلمي * جامعة أبي بكر بلقايد

2nd step:

Calculation of the lower limit of the confidence interval:

$\bar{x}_c = = 20000\text{-}1.645.\mu_0 - Z_\alpha \frac{\sigma}{\sqrt{n}} \bar{x}_c \frac{8000}{\sqrt{49}}$

$$\bar{x}_c = 18120$$

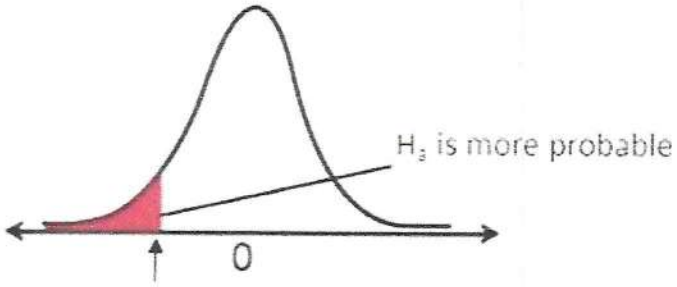
3rd step: The decision rule

We have :

$\bar{x}_c < \bar{x}$

$18120 < 22000$

So, we accept H0 and reject H1:$\mu < 20000$



$\bar{x}_c = 18120$

Exercise 2 (Two-sided test of an average)

We have a sample of size n=16 from a Gaussian population with mean μ and unknown standard deviation. Given the following information:

$\bar{x}$= 49 σ sample = 9 and under α= 0.05

Test the null hypothesis: H0:μ = μ0= 50

And the alternative hypothesis H1: $\mu \neq \mu 0$

Solution :

This is a two-tailed test. We are less than 30 and σ unknown, so the decision variable follows a student law.

1st step :

Formulation of hypotheses

H0:μ= 50

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد

H1:μ ≠50

2nd step

Calculation of the two limits of the confidence interval: (the two critical values)

we replaceσby his estimator

$S=\sqrt{\frac{n}{n-1}}\sigma e \quad S=.\sqrt{\frac{16}{15}}9 \quad S=9.295$

According to the student table (see appendix); α= 0.05 and df= n-1= 15

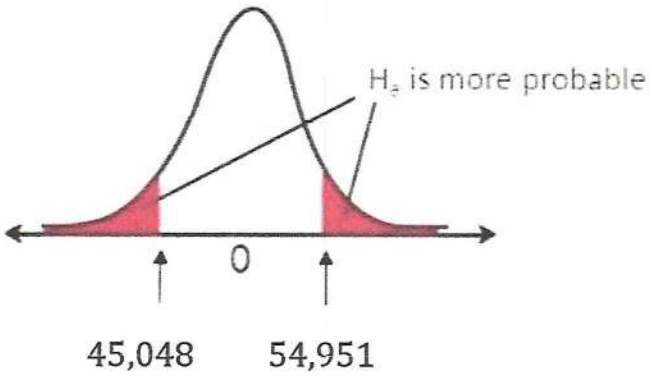
T = 2.131

$Ic=\left[50-2{,}131.\frac{9{,}295}{4};50+2{,}131.\frac{9{,}295}{4}\right]$

Ic =[45,048; 54,951]

$\bar{x}=49\in[45{,}048;54{,}951]$

Therefore, H0 is accepted.



1.2-Comparison of an observed frequency and a theoretical frequency

1.2.1- Principle of the test

Let X be a qualitative variable taking two modalities (success X=1, failure X=0) observed on a population and a sample extracted from this population.

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution	Use When
Population proportion (p)	$p = p_0$	$\dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	Z	$n\hat{p},\ n(1-\hat{p}) \geq 10$

Figure 25: One proportion test

The aim is to test whether or not the proportion p of elements in the population presenting a certain qualitative character can be considered equal to a hypothetical value P0.

The appropriate statistic for this test is the proportion P (estimator of P) whose value is calculated on a sample of size n.

1.2.1.1 -Test statistics

The sampling distribution of the success frequency in the unknown population, $\frac{K}{n}$ follows a normal law such that: $\frac{K}{n}$ follows

$$\mathcal{N}\left(p,\ \sqrt{\frac{p_0 q_0}{n}}\right)$$

The variances are assumed to be equal in the reference population and the population from which the sample is taken. The statistic studied is the difference: $S=\frac{K}{n}$ - p0 whose probability distribution is as follows:

$$S \rightarrow \mathcal{N}\left(0,\ \sqrt{\frac{p_0 q_0}{n}}\right)$$ with under H0 E(S) = 0 and $V(S) = \frac{p_0 q_0}{n}$

كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد
المجال العلمي

We can establish thanks to the central limit theorem the reduced centered variable Z such that:

$$Z = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{\frac{K}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$ but only if np0 and nq0 ≥10

Under H0: p = p0

$$Z = \frac{\frac{K}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$ follows a reduced centered normal law N(0,1)

1.2.1.2 –Application and decision

The hypothesis tested is as follows:

H0: p = p0 versus H1: p≠*p*0

A value z of the random variable Z is calculated:

$$z = \frac{\left|\frac{k}{n} - p_0\right|}{\sqrt{\frac{p_0 q_0}{n}}}$$ also notedε_{obs}.

ε calculated (εobs) is compared with the εthreshold value read from the table of the reduced centered normal law for a fixed error risk α (Decision rules 1).

• if εobs > εthreshold the hypothesis H0 is rejected at the risk of error α: the sample belongs to a population of frequency p and is not representative of the reference population of frequency p0.

• if εobs ≤ εthreshold the hypothesis H0 is accepted: the sample is representative of the reference population with frequency p0.

Solved exercises

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Exercise 1 (Two-tailed test for a proportion)

In a large series company, it is indicated that the percentage of defective parts is 2%. Given the quality of the parts produced, an inspection was carried out on 300 parts; it was indicated that 46 parts were defective at the threshold α=5%. Can we consider that the percentage is significant?

Solution :

1st step :

Formulation of hypotheses

H0:p= p0= 0.02

H1:p ≠p0≠0.02

This is a two-sided test

2nd step

Calculation of the two limits of the confidence interval: (the two critical values)

$$p_{c1}=p0 - Z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}$$

$$p_{c1} = 0.02\text{-}1.96\sqrt{\frac{0{,}02.0{,}98}{300}}$$

$$p_{c1} = 0{,}0041$$

$$p_{c2} = 0.02+ 1.96\sqrt{\frac{0{,}02.0{,}98}{300}}$$

$$p_{c2} = 0{,}0358$$

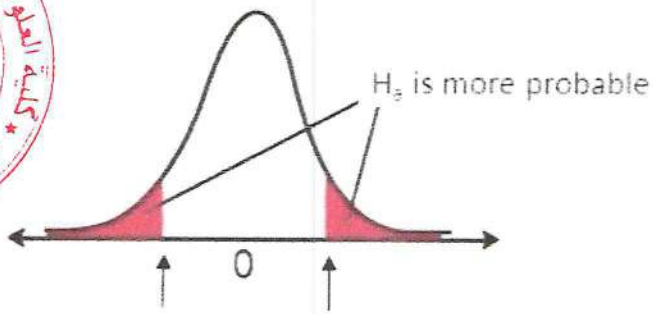
3rd step: The decision rule

$$\hat{p}= = 0.15\frac{46}{300}$$

$\hat{p}$= 0.15∉ [0,0041; 0,0358]

Therefore, we reject H0 and the percentage considered is not significant.

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد



$p_{c1} = 0{,}0041$ $p_{c2} = 0{,}0358$

Exercise 2: (Right-tailed test for a proportion)

Same previous exercise with the following question:

Can we consider this percentage to be abnormally high?

Solution :

1st step :

Formulation of hypotheses

H0:p= p0= 0.02

H1:p >p0>0.02

This is a right-tailed test for a proportion

2nd step

Calculation of the upper limit of the confidence interval: $p_c = p0 + Z_\alpha \sqrt{\frac{pq}{n}}$

$$p_c = 0.02 + 1.645\sqrt{\frac{0{,}02.0{,}98}{300}}$$

$$p_c = 0{,}033$$

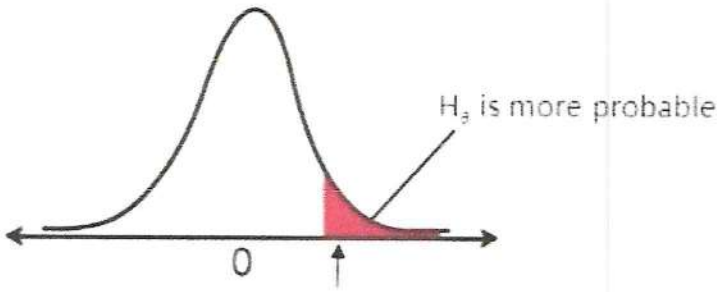
3rd step: The decision rule

$\hat{p} = \frac{46}{300} = 0.15$

$\hat{p} > pc$

0.15>0.033

So, we reject H0 and the percentage is abnormally high.



كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

$p_c = 0{,}033$

Exercise 3: (Left-tailed test for a proportion)

Same previous exercise by adding……..the percentage of defective parts is less than 2%.

…..Can we consider this percentage to be significant?

Solution :

1st step :

Formulation of hypotheses

H0:p= p0= 0.02

H1:p <p0<0.02

This is a left-tailed test for a proportion

2nd step

Calculation of the lower limit of the confidence interval: $p_c = p0 - Z_\alpha \sqrt{\frac{pq}{n}}$

$$p_c = 0.02 + 1.645\sqrt{\frac{0{,}02.0{,}98}{300}}$$

$$p_c = 0{,}0067$$

3rd step: The decision rule

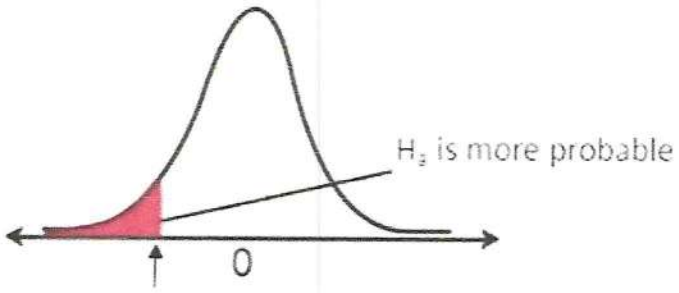
$\hat{p} = \frac{46}{300} = 0.15$

$\hat{p} > pc$

0.15> 0,0067

Therefore, we accept H0 and this test is not significantly<2%

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد



$P_c = 0{,}0067$

Course 8: Tests of homogeneity: Comparison of two means

There are numerous applications that consist, for example, of comparing two groups of individuals with about a particular quantitative characteristic (weight, height, academic performance intelligence quotient, etc.), or comparing two manufacturing processes according to a quantitative characterization particular (breaking resistance, weight, diameter, length, etc.), or even compare the proportions of appearance of a qualitative character of two populations (proportion of defectives, proportion of people favoring a political party, etc.).

1-Comparison of two means

1.1- Principle of the test

Let X be a continuous quantitative character observed on 2 populations following a normal law and two independent samples extracted from these two populations.

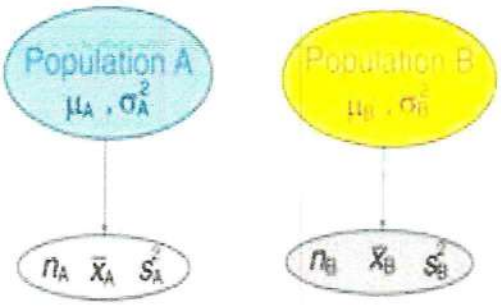


Figure 26: Comparison of two means from two samples

We make the hypothesis that the two samples come from 2 populations whose expectations are equal.

There are several statistics associated with comparing two averages depending on the nature of the data.

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution	Use When
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of n_1-1 and n_2-1	$n_1, n_2 < 30$; and/or σ_1, σ_2 unknown

Figure 27: Statistical tests associated with the comparison of two means

1.2-The population variances are known

1.2.1- Test statistics

Either$\bar{x}_1$the sampling distribution of the mean in population 1 follows a normal law such that:

$$\bar{X}_1 \rightarrow \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n_1}) \qquad \text{and the same for} \qquad \bar{X}_2 \rightarrow \mathcal{N}(\mu_2, \frac{\sigma_2^2}{n_2})$$

$\bar{x}_1$And$\bar{x}_2$ being two independent random variables, we can establish the probability law of the random variable to be studied$\bar{x}_1$-2$\bar{x}$

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) - V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Knowing that$\bar{x}_1$-2$\bar{x}$ follows a normal law

$$\mathcal{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

We can establish thanks to the central limit theorem the reduced centered variable Z such that:

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (E(\bar{X}_1 - \bar{X}_2)}{\sqrt{V(\bar{X}_1 - \bar{X}_2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under H0:μ1 =μ2

avec σ_1^2 et σ_2^2 connues

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ follows a reduced centered normal law N(0,1)

The Table 10 summarizes the case above:

	one-tailed test		two-tailed test
hypothesis	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
test statistic (normal distribution)	$z = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$		
deg. of freedom	—		
rejection	reject H_0 if $z < -z_\alpha$	reject H_0 if $z > z_\alpha$	reject H_0 if $\lvert z \rvert > z_{\alpha/2}$

Table 10: Comparison of two means (variances of known populations)

1.2.2- Application and decision

The hypothesis tested is as follows:

H0:μ1 =μ2 against H1:μ1≠μ2

A value z of the random variable Z is calculated:

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ notée aussi ε_{obs}

ε calculated (εobs) is compared with the εthreshold value read from the reduced centered normal law table for a fixed error risk α.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

• if εobs ≥ εthreshold the hypothesis H0 is rejected at the risk of error α: the two samples are extracted from two populations having expectations μ1 and μ2 respectively.

• if εobs ≤ ε seuil hypothesis H0 is accepted: the two samples are extracted from two populations having the same expectation μ.

Note: For the application of this test, it is imperative that $X \rightarrow N(\mu,\sigma)$ for samples of size < 30 and that the two samples are independent.

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Solved exercises

Exercise 1 (Two-sided test for two means)

Consider the following table:

	Company A	Company B
Number of pieces	75	82
Average amount	38500	42500
Standard deviation	3080	3375

Can we consider at the threshold ofα= 5% that the difference observed between the average amounts is significant?

Solution :

This is a two-tailed test.

Step 1 :

Formulation of hypotheses:

H0:$\mu 1 = \mu 2$ "the difference is not significant"

H1:$\mu 1 \neq \mu 2$ "the difference is significant"

Step 2: calculation of the reduced decision value VDR:

VDR= Z=

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under H0:$\mu 1 = \mu 2$ therefore,$\mu 1 - \mu 2 = 0$ therefore,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{38500 - 4250}{\sqrt{\frac{3080^2}{75} + \frac{3375^2}{82}}}$$

Z=-7.376

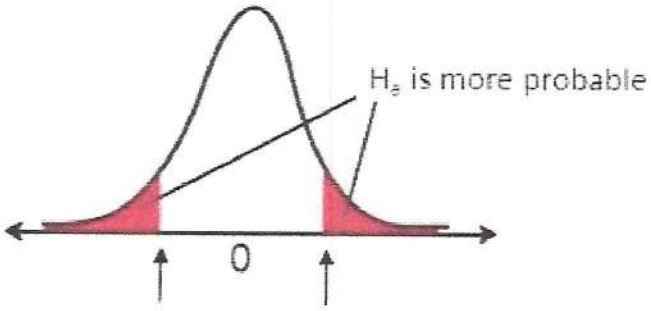
Step 3: decision rule

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

α= 5% therefore:$Z\alpha_{/2}$= 1.96

Z= -7.37∉ [−1,96; +1,96]

We reject H0 and accept H1,the difference observed between the average amounts is significant.



$Z\alpha_{/2}$=- 1.96 $Z\alpha_{/2}$= 1.96

Exercise 2 (Right-tailed test for two means)

Let's take the same previous example.

Can we consider at the threshold ofα= 5% that company A is more efficient than company B?

Solution :

This is a right-tailed test.

Step 1 :

Formulation of hypotheses:

H0:μ1 =μ2

H1:μ1>μ2

Step 2: calculation of the reduced decision value VDR:

VDR= Z= $$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

Under H0:μ1 =μ2 therefore,μ1 -μ2 = 0 therefore,

$$Z=\frac{(\bar{X}_1-\bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

$$Z=\frac{38500-4250}{\sqrt{\frac{3080^2}{75}+\frac{3375^2}{82}}}$$

Z=-7.376

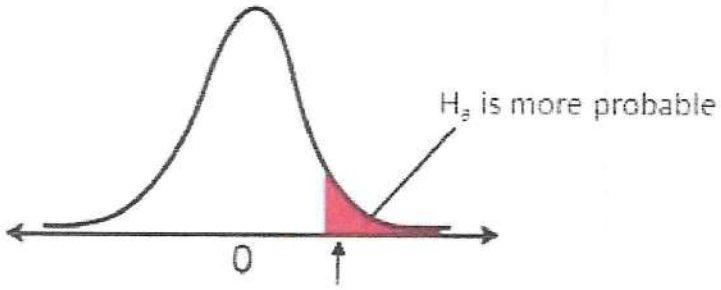
Step 3: decision rule

α= 5% therefore:Z_α= 1.645

Z= -7.37

$Z<Z_\alpha$

We accept H0 and reject H1,Company A is not more efficient than Company B.



Z_α= 1.645

Exercise 3 (Left-tailed test for two means)

The same statement, except the question changes: can we consider at the 5% threshold that company A is less efficient than company B.

Solution :

This is a left-tailed test

Step 1 :

Formulation of hypotheses:

H0:μ1 =μ2

H1:μ1<μ2

Step 2: calculation of the reduced decision value VDR:

$$\text{VDR= Z=}\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Under H0:μ1 =μ2 therefore, μ1 -μ2 = 0 therefore,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{38500 - 42500}{\sqrt{\frac{3080^2}{75} + \frac{3375^2}{82}}}$$

Z=-7.376

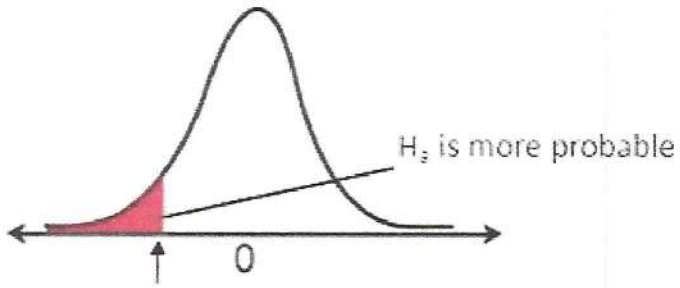
Step 3: decision rule

α= 5% therefore :-Z_α=- 1,645

Z= -7.37

$Z < Z_\alpha$

We reject H0 and accept H1,Company B is more efficient than Company A.



$-Z_\alpha$=- 1,645

المجلس العلمي

1.3-The population variances are unknown and equal

1.3.1- Test statistics

As the population variances are not known, we assume that the two populations present the same variance. H0:$\sigma_1^2 = \sigma_2^2 = \sigma$

•The equality of the variances of the two populations or homoscedasticity then makes it possible to establish the probability law of $1\bar{x}-\bar{x}_2$with :

$$\bar{X}_1 \rightarrow N(\mu_1, \frac{\sigma^2}{n_1}) \text{ et } \bar{X}_2 \rightarrow N(\mu_1, \frac{\sigma^2}{n_2})$$

Knowing that $1\bar{x}-\bar{x}_2$follows a normal law

$$N(\mu_1 - \mu_2, \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

We can establish thanks to the central limit theorem the variable T such that

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (E(\bar{X}_1 - \bar{X}_2)}{\sqrt{V(\bar{X}_1 - \bar{X}_2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Under H0:μ1 =μ2 with$\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$ follows a Student's distribution with (n1 + n2 -2) degrees of freedom

1.3.2- Application and decision

The hypothesis tested is as follows:

H0:μ1 =μ2 against H1:μ1≠μ2

As the population variances are not known, the equality of the variances must be checked

H0:$\sigma_1^2 = \sigma_2^2 = \sigma^2$

against H1:$\sigma_1^2 \neq \sigma_2^2$

A value t of the random variable T is calculated:

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

With

$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$ variance estimation σ^2 common

t calculated (tobs) is compared with the threshold value read in the Student table for a risk of errorαfixed and (n1 + n2 – 2) degrees of freedom.

• if tobs > t threshold hypothesis H0 is rejected at the risk of error α: the two samples are extracted from two populations having expectations μ1 and μ2 respectively.

• if tobs ≤ t threshold hypothesis H0 is accepted: the two samples are extracted from two populations having the same expectation μ.

Noticed : For the application of this test, it is imperative that $X \rightarrow N(\mu,\sigma)$ for samples of size < 30, that the two samples are independent and that the two estimated variances are equal.

المجلس العلمي

1.3-Population variances are unknown and unequal

If the population variances are not known and if their estimates from the samples are significantly different (variance comparison test), two scenarios must be considered depending on the size of the samples compared:

Large samples with n1 and n2 greater than 30.

Small samples with n1 and/or n2 less than 30.

1.3.1-Case where n1 and n2 > 30

The statistic used is the same as for the case where the variances are known.

Under H0:μ1 =μ2

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

follows a reduced centered normal law N(0,1)

As the variances are unknown and significantly different $\sigma_1^2 \neq \sigma_2^2$ we replace the variances of the populations with their point estimates calculated from the samples,

$$\hat{\sigma}_1^2 = \frac{n_1}{n_1 - 1} s_1^2 \quad \text{et} \quad \hat{\sigma}_2^2 = \frac{n_2}{n_2 - 1} s_2^2$$

The hypothesis tested is as follows:

H0:μ1 =μ2 against H1:μ1≠μ2

A value z of the random variable Z is calculated:

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} = \varepsilon_{obs.}$$

ε calculated (εobs) is compared with the εthreshold value read from the reduced centered normal law table for a fixed error risk α.

• if εobs> εthreshold the hypothesis H0 is rejected at the risk of error α: the two samples are extracted from two populations having expectations μ1 and μ2 respectively.

• if εobs≤ εthreshold hypothesis H0 is accepted: the two samples are extracted from

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

two populations having the same expectation μ.

Noticed : For the application of this test, it is imperative that $X \longrightarrow N(\mu,\sigma)$ and that the two samples are independent.

1.3.2-Case where n1 and/or n2 < 30

When the variances are unequal and the sample sizes are small, the probability law followed by $\bar{x}_1$-2$\bar{x}$ is not known. We then resort to non-parametric tests.

العلوم الإنسانية والعلوم الإسلامية
المجلس العلمي
جامعة أبي بكر بلقايد

Course 9: Tests of homogeneity: Comparison of two frequencies

1-Comparison of two frequencies

There are many applications where we must decide whether the difference observed between two proportions is significant or whether it is attributable to sampling chance.

As in the case of comparing two means, we must know the sampling distribution of the difference of two proportions to conduct a test on the equality of two proportions (or to estimate, by confidence interval, the difference).

1.1-Principle of the test

Let X be a qualitative variable taking two modalities (success X=1, failure X=0) observed on 2 populations and two independent samples extracted from these two populations. We make the hypothesis that the two samples come from 2 populations whose probabilities of success are identical.

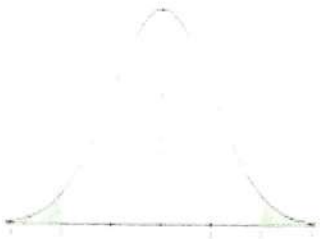
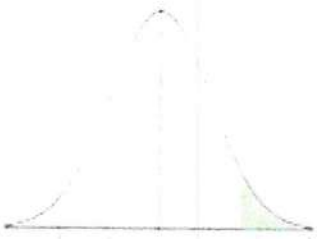
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$	$H_0: p_1 = p_2$ $H_1: p_1 > p_2$	$H_0: p_1 = p_2$ $H_1: p_1 < p_2$
		
$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 \neq 0$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 > 0$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 < 0$

Figure 29: Comparison of two frequencies of two samples

The problem is whether the difference between the two observed frequencies is real or explainable by sampling fluctuations. To resolve this problem, two frequency comparison tests are possible:

كلية العلوم الإنسانية والعلوم الاجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

Testεor reduced centered variable test and chi-square testχ2

1.2-Test statisticsε

• The sampling distribution of the frequency of success in population 1, Follows a normal law such as:$\frac{K_1}{n_1}$

$$\frac{K_1}{n_1} \text{ suit } \mathcal{N}(p_1, \sqrt{\frac{p_1 q_1}{n_1}}) \text{ et de même pour } \frac{K_2}{n_2} \text{ suit } \mathcal{N}(p_2, \sqrt{\frac{p_2 q_2}{n_2}})$$

if and only if n1p1, n1q1, n2p2, n2q2≥5

• $\frac{K_1}{n_1}$ et $\frac{K_2}{n_2}$

being two independent random variables, we can establish the probability law of the random variable to be studied

$$\frac{K_1}{n_1} - \frac{K_2}{n_2}$$

Sachant que $\frac{K_1}{n_1} - \frac{K_2}{n_2}$ suit une loi normale $\mathcal{N}(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}})$.

We can establish thanks to the central limit theorem the reduced centered variable Z such that:

$$Z = \frac{\left(\frac{K_1}{n_1} - \frac{K_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Under H0: p1 = p2 with

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$Z = \frac{\left(\frac{K_1}{n_1} - \frac{K_2}{n_2}\right)}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$$ follows a reduced centered normal law N(0,1)

المجلس العلمي

1.3-Application and decision

The p value, the probability of success common to the two populations, is in reality not known. It is estimated from the results observed on the two samples:

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$$

where k1 and k2 represent the number of successes observed for sample 1 and sample 2 respectively.

The hypothesis tested is as follows:

H0: p1 = p2 versus H1: p1≠*p*2

A value z of the random variable Z is calculated:

$$z = \frac{\left|\frac{k_1}{n_1} - \frac{k_2}{n_2}\right|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{avec} \quad \hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$$

z orεcalculated (εobs) is compared with the value εthreshold read on the table of the reduced centered normal law for a risk of error α fixed.

•ifε_{obs}>εthreshold the hypothesis H0 is rejected at risk of errorα: both samples are extracted from two populations having success probabilities p1 and p2 respectively.

•if$\varepsilon_{obs} \leq$ εthreshold the hypothesis H0 is accepted: the two samples are extracted from two populations having the same probability of success p.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة محمد بوضياف المسيلة

Solved exercises

Exercise 1: (Two-tailed test for two proportions)

A factory produces metal parts for an automobile manufacturer. Production is carried out by two machines: machine 1 and machine 2. Among the pieces produced, some are defective:

	Machine 1	**Machine 2**
Number of parts checked	200	150
Number of defective parts	25	20

Can we conclude at the significance level?α= 5% that the difference between the proportions of defective parts produced by the two machines is significant?

Solution :

This is a two-tailed test.

Step 1: Formulation of hypotheses

H0: P1= P2

H1: P1≠P2

Step 2: calculation of the decision variable Z

You must check the conditions:

n1p1≥5 and n1q1≥5

n2p2≥5 and n2q2≥5

$\hat{p}_1 = = 0.125, 200.\ 0.125 = 25 \frac{25}{200} \geq 5$

$\hat{p}_2 = = 0.133, 150.\ 0.133 = 19.5 \frac{20}{150} \geq 5$

$\hat{p} = \frac{n1.p1+n2p2}{n1+n2} = -0.128$

$$z = \frac{\left|\frac{k_1}{n_1} - \frac{k_2}{n_2}\right|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{avec} \quad \hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$$

Z= - 0.221

Step 3: decision rule

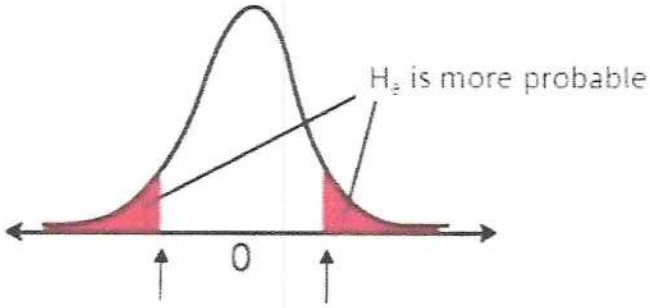
المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

$Z_{0.05} = 1.96$

$Z = -0.221 \in [-1,96;\ +1,96]$

Therefore, we accept the null hypothesis H0 and reject H1.

The difference between the proportions is not significant.



$-1,96;\ +1,96$

Exercise 2: (Right-tailed test for two proportions)

Same exercise with the following question: can we conclude at the 5% significance level that machine 1 is more efficient than machine 2?

Solution :

This is a right-tailed test

Step 1: Formulation of hypotheses

H0: P1= P2

H1: P1>P2

Step 2: calculation of the decision variable Z

You must check the conditions:

n1p1≥5 and n1q1≥5

n2p2≥5 and n2q2≥5

$\hat{p}_1 = = 0.125,\ 200.\ 0.125 = 25\frac{25}{200} \geq 5$

$\hat{p}_2 = = 0.133,\ 150.\ 0.133 = 19.5\frac{20}{150} \geq 5$

$\hat{p} = \frac{n1.p1+n2}{n1+n2} = -0.128$

$$z = \frac{\left|\frac{k_1}{n_1} - \frac{k_2}{n_2}\right|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{avec} \quad \hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$$

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Z= - 0.221

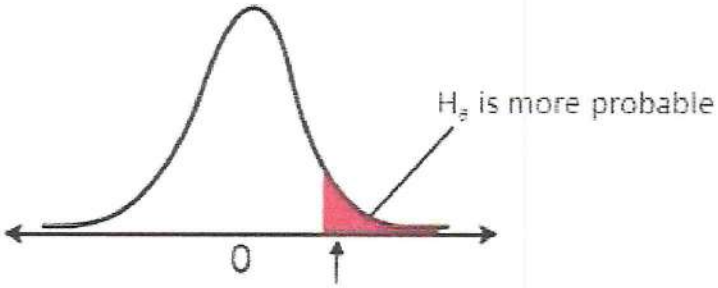
Step 3: decision rule

Zα= 1.645

Z<Zα

Therefore, we accept the null hypothesis H0 and reject H1.

That is to say, machine 1 is not more efficient than machine 2



Z= - 0.221 Zα= 1.645

Exercise 3: (Left-tailed test for two proportions)

Same exercise with the following question: can we conclude at the 5% significance level that machine 2 is more efficient than machine 1?

Solution :

This is a left-tailed test

Step 1: Formulation of hypotheses

H0: P1= P2

H1: P1<P2

Step 2: calculation of the decision variable Z

You must check the conditions:

n1p1≥5 and n1q1≥5

n2p2≥5 and n2q2≥5

$\hat{p}_1 = = 0.125, 200.\ 0.125 = 25 \frac{25}{200} \geq 5$

كلية العلوم الإنسانية والعلوم الإجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

$\hat{p}_2 = = 0.133,\ 150.\ 0.133 = 19.5 \frac{20}{150} \geq 5$

$\hat{p} = \frac{n1.p1+n2}{n1+n} = -0.128$

$$z = \frac{\left|\frac{k_1}{n_1} - \frac{k_2}{n_2}\right|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{avec} \quad \hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$$

Z= - 0.221

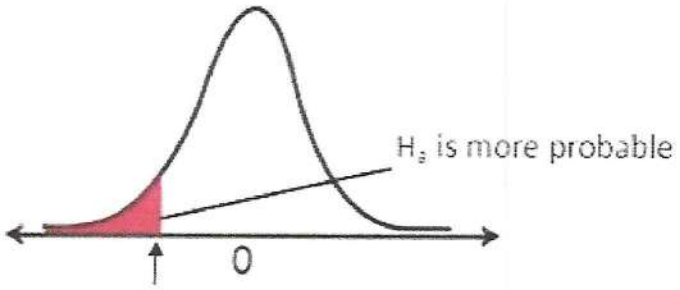
Step 3: decision rule

-Zα= -1.645

Z>Zα

Therefore, we accept the null hypothesis H0 and reject H1.

That is to say, machine 2 is not more efficient than machine 1.



-Zα= -1.645

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

Course 10: Tests of homogeneity: Comparison of two variances

1-Comparison of two variances

The comparison of two normal populations can relate not only to their central value, their average but also to their dispersion.

The most used dispersion characteristic is variance.

Remember that one of the conditions for applying the Student test in the case of comparing two means is that the samples come from two normal populations with identical variances: $\sigma_1^2 = \sigma_2^2$

This hypothesis can be verified using a test for the equality of two variances called the F test or Fisher Snedecor test.

1.1--Principle of the test

Let X be a random variable observed on 2 populations following a normal distribution and two independent samples extracted from these two populations.

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$
		
$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$

Figure 30: Comparison of two variances of two samples

We make the hypothesis that the two samples come from 2 populations whose variances are equal.

The variance comparison test is necessary when comparing two means when the population variances σ_1^2 and σ_2^2 are not known. It is also the statistic with the analysis of variance.

المجلس العلمي
كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد

1.2-Test statistics

The statistic associated with the test for comparing two variances corresponds to the ratio of the two estimated variances.

Under Ho:$\sigma_1^2=\sigma_2^2$

$$F_{obs.} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\frac{n_1}{n_1-1}s_1^2}{\frac{n_2}{n_2-1}s_2^2}$$

follows a Fisher-Snedecor law with (n1-1, n2-1) degrees of freedom with$\sigma_1^2 > \sigma_2^2$ because the ratio of variances must be greater than 1.

1.3-Application and decision

The value of the calculated F statistic (Fobs) is compared with the Fthreshold value read in the Fisher-Snedecor law table for risk of errorαfixed and (n1-1, n2 -1) degrees of freedom.

•if Fobs≥ F threshold hypothesis H0 is rejected at risk of error α: the two samples are extracted from two populations having statistically different variancesσ_1^2And .σ_2^2

• if Fobs ≤ Fseuil hypothesis H0 is accepted: the two samples are extracted from two populations with the same variance.σ^2

Noticed: For the application of this test, it is imperative that $x\rightarrow\mathcal{N}(\mu,\sigma)$and that the two samples are independent.

المجلس العلمي

	one-tailed test		two-tailed test
hypothesis	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$
test statistic (F distribution)	$F = \frac{s_2^2}{s_1^2}$	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$
deg. of freedom	$df_1 = n_1 - 1$		$df_2 = n_2 - 1$
rejection	reject H_0 if $F > F_\alpha$		reject H_0 if $F > F_{\alpha/2}$

Table 11: Test of homogeneity between two variances

Solved exercise

Data was collected during a study on calcium supplements and their effects on blood pressure. A placebo group and a calcium group began the study with a blood pressure measurement. The following results were obtained:

	participants	Standard deviation
Placebo	n=13	S1=9,46
calcium	n=15	S2=8,469

At a significance level of 0.05, test the assertion that the two samples come from populations with the same standard deviation.

We check if the conditions are met;

The two samples are independent.

The samples come from normal populations.

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

كلية العلوم الإنسانية والعلوم الإجتماعية * جامعة أبي بكر بلقايد * المجلس العلمي

Test statistic

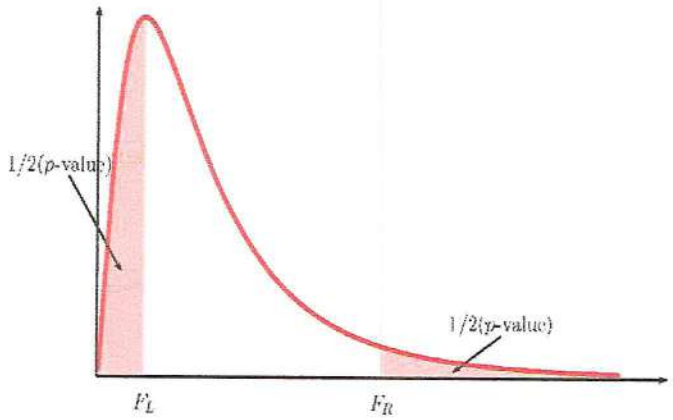
$$F = \frac{9.468^2}{8.469^2} = 1.248$$

Critical values this is a bilateral test with an area of 0.025 (0.05/2), we compares F to the critical value on the right which corresponds to 3.0502

(Fisher table with $\alpha = 0.025$, ddl1=12, ddl2=14, F(12, 14))

Conclusion: $F < 3.0502$, F=1.248 is not located in the critical region. So we cannot reject H0.

Interpretation: There is not enough evidence to reject the hypothesis null of equality of variances.



كلية العلوم الإنسانية والعلوم الإسلامية
المجلس العلمي
جامعة أبي بكر بلقايد

MCQs on hypothesis testing

1-Regarding the power of a statistical test, which assertions are true?

A- It is the probability that the calculated value of the statistic belongs to the distribution under H0.

B-It is the probability that the calculated value of the statistic belongs to the distribution under H1.

C-It is the complement to 1 of the risk of the first kind.

D-It is the complement to 1 of the risk of the second kind.

E-It is most often chosen equal to 5%.

2-Which of the following statement (s) is (are) true?

A-The first type error consists of accepting the null hypothesis when it is false.

B-The first type error consists of rejecting the null hypothesis even though it is true.

C-The second type error consists of accepting the null hypothesis when it is false.

D-The second type error consists of rejecting the null hypothesis when it is false.

E-The error of the first type is most often chosen to be equal to 5%.

3-A consumer association wishes to demonstrate that the average sugar intake of a portion of product A, denoted γ, is greater than 25g. To do this, it will analyze a sample of such portions and carry out a statistical test. Determine the most appropriate null and alternative hypotheses.

$H_0 : \gamma = 25 \text{ ، } H_a : \gamma \neq 25$

$H_0 : \gamma \geq 25 \text{ ، } H_a : \gamma < 25$

$H_0 : \gamma < 25 \text{ ، } H_a : \gamma \geq 25$

$H_0 : \gamma \leq 25 \text{ ، } H_a : \gamma > 25$

كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

4-Continuation of the previous question. What type of test is this?

Two-tailed test Right tailed test Left tailed test

5-We consider two populations, with unknown standard deviations. We give the results from two independent random samples, from the two populations: 1- size n1 = 200; sample mean $\bar{x}1 = 9.5$; sample standard deviation s1 = 4.1.
2- size n2 = 300; sample mean $\bar{x}2 = 8.5$; sample standard deviation s2 = 3.5. What is the point estimate of the difference between the means μ1 and μ2 of the two populations?

1.0 [0.4; 1.6] 18 0.6

6-We consider the same situation as in the previous question. We consider the following hypothesis test: H0: $\mu 1 - \mu 2 \geq 0$; Ha: $\mu 1 - \mu 2 < 0$. What formula should be used to calculate the test statistic $\bar{z}$?

$z = \sqrt{n}\frac{\bar{x}_1 - \bar{x}_2}{s}$ $z = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n_1 + n_2}}}$ $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

7-We consider the same test as in the two previous questions. Determine the rejection zone, for a test threshold $\alpha = 0.05$.

$]-\infty; -1.65[$ $]-\infty; -1.65[\cup]1.65; +\infty[$
$]-\infty; -1.96[\cup]1.96; +\infty[$ $]-\infty; -1.96[$

8- Which of the following statements regarding the null hypothesis H0 of a hypothesis test is (are) correct?

A - The decision rule is constructed assuming that H0 is true

B - We formulate H0 after examining the data

C - The probability of rejecting H0 when it is true is the risk of the first kind

D - We affirm that H0 is true if H0 is not rejected

E - In a link test, H0 assumes that the 2 variables are independent (we admit that the application conditions are verified)

Answers to the MCQs

كلية العلوم الإنسانية والعلوم الإجتماعية
جامعة أبي بكر بلقايد
المجلس العلمي

1- Answers: B,D

2- Answers: B,C,E

3- Answer :

$H_0 : \gamma \leq 25$, $H_a : \gamma > 25$

4- Answer :

Right tailed test

5- Answer

1.0

6- Answer :

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

7- Answer :

$]-\infty; -1.65[$

8- Answers: A,C,E

Bibliographic references

المجلس العلمي
كلية العلوم الإنسانية والعلوم الاجتماعية
جامعة أبي بكر بلقايد

Al Abassi, I., El Marhoum, A (1999). Descriptive statistics course. Collection of the Faculty of Legal, Economic and Social Sciences, Marrakech.

Bardin, B.M. (2016). Descriptive statistics course. Cango-Brazzavill: DEUG HAL.

Bailly, p. Carrère, Ch (2007). Descriptive statistics - Course. Grenoble University Press. PUG.

Bressoud, é., & kahané, C. (2010). Descriptive statistics. 2nd Pearson edition.

Dancy, c., & Reidy, j. (2023). Statistics without math for psychologists. De boeck, 3rd edition.

Elmeddah, y. (2013). The statistical method: Tests relating to variances and means.

Goldfarb, b., & Pardoux, c. (2011). Introduction to statistical method, manual and corrected exercises. Paris: 6th edition, Dunod.

Goldfarb, b., & Pardoux, c. (2013). Introduction to statistical method, statistics and probability. Paris: 7th edition, Dunod.

Houde, L. (2014). Hypothesis Testing-Quantitative Analysis of Management Problems. University of Quebec at Trois-Rivières, Department of Mathematics and Computer Science.

Jt, 3. . (2021). Some elements of descriptive statistics. http://www.gymomath.ch/javmath.

Mazerolle, f. (2005). Descriptive statistics, LMD memo, statistical series with one and two variables - time series. Gualino indices publisher EJA.

Mouchiroud, D. (2003). Mathematics: Tools for Biology. Deug SV – UCBL.

Ruch, J.-J. (2012-2013). Statistics: Hypothesis testing.

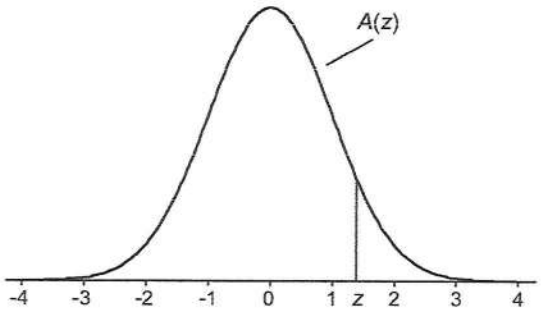
كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

Appendix

المجلس العلمي
جامعة أبي بكر بلقايد

TABLE A.1

Cumulative Standardized Normal Distribution



$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

المجلس العلمي
جامعة أبي بكر بلقايد * كلية العلوم الإنسانية والعلوم الاجتماعية *

TABLE A.2

t Distribution: Critical Values of *t*

Degrees of freedom	*Two-tailed test:*	*Significance level* 10%	5%	2%	1%	0.2%	0.1%
	One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

كلية العلوم الإنسانية والعلوم الاجتماعية
المجلس العلمي
جامعة أبي بكر بلقايد

TABLE A.3

F Distribution: Critical Values of *F* (5% significance level)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62
250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59
600	3.86	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.90	1.85	1.77	1.71	1.66	1.62	1.59
750	3.85	3.01	2.62	2.38	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.70	1.66	1.62	1.58
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58