An introdution to statistics and probability

Doctor : OUAOUA Amar
August 20, 1955 University. Skikda.

August 2023

CONTENTS

0.1 Introduction

The purpose of this handout is to provide an introduction to the general principles of probability and statistics, which may be of interest to anyone interested in this discipline, regardless of their specialization. It may be of particular interest as a reference for second-year students in civil engineering, mechanical engineering, process engineering, control systems, and electronics. Probability is one of the modules taught by numerous researchers, as most sciences rely on probabilities for studies; results are often based on probabilities that are primarily governed by laws, and each phenomenon has a specific law. On the other hand, statistics plays an increasingly important role in almost all aspects of human behavior. Initially developed in public affairs, which explains its name, its influence has now expanded to agriculture, medicine, biology, physics, and other branches of science and technology. At the beginning of each chapter, we provide important definitions and theorems, and then we offer some exercises with solutions. The first part of the handout is devoted to probability, which includes combinatorial analysis, probability calculations, and random variables. The second part focuses on statistics. First, we cover the terminology of statistics, and then we offer some exercises with solutions. The first part of the handout is devoted to probability, which includes combinatorial analysis, probability calculations, and random variables. The second part focuses on statistics. First, we cover the terminology of statistics, and then we have two chapters: statistical series with one variable and statistical series with two variables. Through this content, the main objective is to understand the definitions and theorems, and the student will also be able to study probability and statistical models. .

Recommended Prerequisite Knowledge: The student should have a good understanding of the previous lessons related to Mathematics 1 and 2, which promote a better comprehension of the lessons covered in this module, which are as follows:

1-Concept of sets and subsets: -Concept of set-element of a set.inclusion relation (properties).
- Intersection and union (properties).
- Complement of a set (properties).
- Difference of two sets (properties).
- Partition of a set.

2- Integrals: - Integration of a continuous function (definition and primitive function).
- Integral over a bounded interval (properties).
- Integral over an unbounded interval (properties).
- Integration by parts.
- Integration using a change of variable.

3- Expansions: - Expansions of exponential, logarithmic, and power functions.

4-Derivatives: - Higher-order derivatives.
- Successive derivatives.

Part I

Probability

A brief history

The calculation of probabilities: "A problem related to games of chance proposed to an austere Jansenist by a man of the world was the origin of the calculus of probabilities." Denis Poisson (1781-1840). The Chevalier de Méré presented Blaise Pascal (1623-1662) with problems on games of chance, including the "problem of parts": The prize of a tournament is won by the first participant who wins a fixed number of rounds. If the game is interrupted before completion, how should the prize be fairly distributed among the participants?

Many incorrect solutions had been proposed for this centuries-old problem. Pascal provided a correct solution, which he submitted to Pierre de Fermat (1601-1665) in 1654. He published his solution in his work "Treatise on the Arithmetical Triangle" in 1665. In 1657, Christiaan Huygens (1629-1695) published the book "De ratiociniis in ludo aleae," which presented the fundamental concepts of probability calculus, such as the calculation of the expected value of a random variable taking a finite number of values. In his posthumous work "Ars conjectandi" in 1713, Jacques Bernoulli (1654-1705) further developed the results of Huygens. He also demonstrated, using combinatorial calculations, the law of large numbers (convergence of the empirical mean to the true mean), which was instrumental in the development of probability theory. In 1733, in "The Doctrine of Chances," Abraham de Moivre (1667-1754) provided a precise statement, in a particular case, of the convergence rate of the law of large numbers. This was the first version of the central limit theorem. This result was further extended by Pierre-Simon Laplace (1749-1827). Laplace, using infinitesimal calculus and developing generating functions and characteristic functions in his work "Théorie Analytique des Probabilités," published in 1812, went beyond the scope of combinatorial calculations and gave new impetus to the field of probability calculus.

General results on the law of large numbers and the central limit theorem were established in the 19th century by Denis Poisson (1781-1840), Irénée-Jules Bienaymé (1796-1878), and the St. Petersburg school, including Pafnouti Tchebychev (1821-1894), Andreï Markov (1856-1922), and Alexandre Liapounov (1857-1918). In the 20th century, the theory of measure and integration clarified the fundamental concepts of probability calculus, such as probability measures, random variables, probability distributions, expectations, and conditional probabilities. The monograph "Grundbegriffe der Wahrscheinlichkeitsrechnung" by Andreï Kolmogorov (1903-1987), published in 1933, provided the theoretical framework that still underlies the field of probability calculus today.

In the first half of the 20th century, probability calculus experienced a new surge with the study of stochastic processes and, more importantly, their numerous applications. These applications multiplied in the second half of the century: modeling physical phenomena (especially at the microscopic level for complex fluids or materials and in statistical physics) or biological phenomena (in demography and epidemiology, but also at the cellular or DNA level), in computer science (algorithm analysis, image processing, or network analysis), in economics (insurance or market finance), as well as in engineering (reliability, optimization, risk analysis, management of random environments). Finally, with the increased power of computers, simulations and Monte Carlo methods, developed in the 1940s, have amplified the use of random models and have become an important field within probability theory.

CHAPTER 1

Combinatorial analysis

Principle of multiplication: Allows counting the number of results of experiments that can be decomposed into a sequence of sub-experiments.

Principle: Suppose that an experiment consists of a sequence of m sub-experiments. If the $i-th$ experiment has n_i possible outcomes for $i = 1, ..., m$, then the total number of possible outcomes of the overall experiment is

$$n = n_1 n_2 ... n_m \tag{1.1}$$

1.1 Arrangement

1.1.1 Arrangement with repetition

Definition 1.1 Given a finite set of n objects, an arrangement with repetition of these n objects p by p is defined as any ordered grouping of p objects chosen from the n objects with repetition.

- The number of arrangements of n objects p by p is equal to

$$n^p. \tag{1.2}$$

Example 1.1 The arrangements with repetition of 2 elements taken from $\{1, 2, 3, 4\}$ are

$$\{1,1\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,1\}, \{2,2\}, \{2,3\}, \{2,4\}, \{3,1\}$$
$$\{3,2\}, \{3,3\}, \{3,4\}, \{4,1\}, \{4,2\}, \{4,3\}, \{4,4\}.$$

There are 16 of them.

1.1.2 Arrangement without repetition

Definition 1.2 Given a finite set of n objects, an arrangement without repetition of these n objects taken p at a time refers to an ordered grouping of p objects chosen from the n objects without repetition.

- The number of arrangements of n objects taken p at a time is equal to $n(n-1)...(n-p+1)$ and will be denoted as A_n^p.

$$A_n^p = \frac{n!}{(n-p)!}. \tag{1.3}$$

Example 1.2 The arrangements without repetition of 2 elements taken from $\{1,2,3,4\}$ are

$$\{1,2\},\{1,3\},\{1,4\},\{2,1\},\{2,3\},\{2,4\},\{3,1\}$$
$$\{3,2\},\{3,4\},\{4,1\},\{4,2\},\{4,3\}.$$

There are 12 of them.

1.2 Permutation

1.2.1 Permutation without repetition

Definition 1.3 Given a finite set of n objects, an arrangement without repetition of these n objects refers to any ordered grouping of n objects chosen from the n objects.

- The number of permutations of n objects is equal to

$$n! \tag{1.4}$$

A permutation is an arrangement without repetition of these n objects taken n at a time, thus falling within the scope of the previous paragraph.

It can be noted that $n!$ (n factorial) is the number of ways to arrange n objects.

Remark 1.1

$$P_n = n! = n(n-1)(n-2)\dots 2.1 \tag{1.5}$$

1.2.2 Permutation with repetition

The n objects are not all distinct, and it is assumed that they are divided into K distinct groups with respective sizes n_1, n_2, ...,n_k, such that $\sum_{1}^{k} n_i = n$.

Definition 1.4 Given a finite set of n objects, a permutation with repetition of these n objects refers to any ordered grouping of n objects chosen from the n objects.

- The number of permutations with repetition of n elements divided into K groups with respective sizes n_1, n_2, ...,n_k is given by

$$\overline{P}_n = \frac{n!}{n_1!n_2!\ldots n_k!}. \tag{1.6}$$

Example 1.3 The number of words that can be constructed from the word 'Canada' containing 6 letters is $\overline{P}_6 = \frac{6!}{1!1!1!3!} = 120$.

1.3 Combination

1.3.1 Combination without repetition

Definition 1.5 Given a finite set of n objects, a combination without repetition of these n objects taken p at a time is defined as an unordered grouping of p objects chosen from the n objects without repetition

- The number of combinations of n objects taken p at a time is equal to:

$$C_n^p = \frac{A_n^p}{p!} = \frac{n!}{p!(n-p)!}. \tag{1.7}$$

Example 1.4 The combinations without repetition of 2 elements taken from $\{1, 2, 3, 4\}$ are

$$\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}.$$

There are 6 of them.

1.3.2 Combination with repetition

Definition 1.6 Given a finite set of n objects, a combination with repetition of these n objects taken p at a time is defined as an unordered grouping of p objects chosen from the n objects with repetition.

- The number of combinations with repetition of these n objects taken p at a time is equal to:

$$K_n^p = C_{n+p-1}^p = \frac{(n+p-1)!}{p!(n-1)!}. \tag{1.8}$$

Example 1.5 The combinations with repetition of 2 elements taken from $\{1,\ 2,\ 3\}$ are:

$$\{1,\ 1\}, \{1,\ 2\}, \{1,\ 3\}, \{2,\ 2\}, \{2,\ 3\}, \{3,\ 3\}.\ \ K_3^2 = \frac{(3+2-1)!}{2!(3-1)!} = 6.$$

Proposition 1.1 *We have:*

$$C_n^p = C_n^{n-p} \tag{1.9}$$

$$C_n^p = C_{n-1}^p + C_{n-1}^{p-1} \tag{1.10}$$

Pascal's Triangle The previous formula allows us to construct **Pascal's Triangle**, where we plot p horizontally and n vertically

n \ P	0	1	2	3	4	5	6
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1

Each number in Pascal's Triangle, denoted as C_n^p, is obtained by adding the number just above it $\left(C_{n-1}^p\right)$ and the number above and to the left $\left(C_{n-1}^{p-1}\right)$. As

$$C_3^2 = 2 + 1 = 3.$$

By convention, $0! = 1$, which implies that C_n^0 is equal to 1 for all n.

The Binomial Formula We can find the Binomial Formula, which is the formula that gives the coefficients in the expansion of the Binomial $(a+b)$ raised to any power n

$$\begin{aligned}(a+b)^0 &= 1\\ (a+b)^1 &= C_1^0 a + C_1^1 b\\ (a+b)^2 &= C_2^0 a^2 + C_2^1 ab + C_2^2 b^2\end{aligned}$$

More generally, it is possible to prove that:

$$(a+b)^n = C_n^0 a^n + C_n^1 a^{n-1} b + C_n^2 a^{n-2} b^2 + ... + C_n^{n-1} a^1 b^{n-1} + C_n^n b^n. \tag{1.11}$$

CHAPTER 2

Introduction to Probability, Conditioning, and Independence

2.1 Introduction to Probability

2.1.1 Algebra of events

An experiment is a random experiment ε whose outcome is uncertain. The possible outcomes of an experiment typically involve randomness. The set of possible outcomes (possible results, eventualities) is called the sample space.

We denote by Ω the set of possible outcomes, $\mathcal{F}$ the set of events, and $\mathbb{P}$ the probability.

An outcome of the random experiment ε is called an event.

An event is called an elementary event if it cannot be decomposed using other events from the same random experiment.

A set of events associated with ε, any subset of $\mathcal{F}$ from $\mathcal{P}(\Omega)$, satisfying the following three properties

1) ϕ and $\Omega \in \mathcal{F}$

2) If A$\in \mathcal{F}$, then A$^c \in \mathcal{F}$

3) Let I be a finite or infinite subset of $\mathbb{N}$. If $(A_i)_{i\in I}$ is a sequence of events in $\mathcal{F}$, then $\underset{i\in I}{\cup} A_i \in \mathcal{F}$ and $\underset{i\in I}{\cap} A_i \in \mathcal{F}$.

A set of events satisfying the three properties below is called a $\sigma-$algebra.

A and B are two subsets of a set Ω, then we have

* $A \cup B = \{x \in \Omega;\ x \in A \text{ or } x \in B\ \}$

* $A \cap B = \{x \in \Omega;\ x \in A \text{ and } x \in B\ \}$

* $A - B = \{x \in \Omega;\ x \in A \text{ and } x \notin B\ \}$

Remark 2.1 *$A^c = \Omega - A$; $(A^c)^c = A$; $\Omega^c = \phi$; $\phi^c = \Omega$

$* A \cap A = A$; $A \cup A = A$; $A \cap A^c = \phi$; $A \cap \Omega = A$; $A \cup \Omega = \Omega$; $A - B = A \cap B^c$
$* A \subseteq B \Rightarrow A \cap B = A$; $A \subseteq B \Rightarrow A \cup B = B$
$*(A \cap B)^c = A^c \cup B^c$
$*(A \cup B)^c = A^c \cap B^c$.

With this representation method, logical operations on events such as "and," "or," and "negation" are translated into set operations: intersection, union, and complement. Here is a correspondence table between the two languages.

Notations	Set-based vocabulary	Probabilistic vocabulary
ϕ	empty set	impossible event
Ω	complete set	certain event
ω	element of Ω	elementary event
A	subset of Ω	event
$\omega \in A$	ω belongs to A	the outcomes of ω are possible realizations of A
$A \subset B$	A is included in B	A implies B
$A \cup B$	union of A and B	A or B
$A \cap B$	intersection of A and B	A and B
A^c	complement of A in Ω	complementary event to A
$A \cap B = \phi$	A and B are disjoint	A and B are incompatible

Table 1

2.1.2 Definitions

Definition 2.1 Two incompatible or disjoint events A and B if $A \cap B = \phi$.

Definition 2.2 A probability on $(\Omega,\ \mathcal{F})$ is a mapping $\mathbb{P}$ from $\mathcal{F}$ to $[0,1]$ satisfying the following two properties

1) $\mathbb{P}(\Omega) = 1$

2) For any sequence $(A_i)_{i \in I}$, finite or countably infinite, of mutually exclusive events in $\mathcal{F}$, we have

$$\mathbb{P}\left(\underset{i \in I}{\cup} A_i\right) = \sum_{i \in I} \mathbb{P}(A_i). \tag{2.1}$$

2.1.3 Probability Spaces

Definition 2.3 A finite probability space associated with a random experiment ε is the triplet $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, where the sample space Ω is finite.

2.1.4 General Theorems of Probability

For all events A and B in $\mathcal{P}(\Omega)$, we have

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
2. $\mathbb{P}(\phi) = 0$.
3. $\mathbb{P}(A - B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$
4. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
6. If A and B incompatibles, then

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. More generally, if $A_1, ..., A_n$ are n events pairwise incompatible, then

$$\mathbb{P}\left(\overset{n}{\underset{i=1}{\cup}} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{2.2}$$

Proof. . ■

Equiprobability and uniform probability We will say that there is equiprobability if all elementary events are equal in probability

In this case, $\mathbb{P}$ is the uniform probability on $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$.

Application If event A is the union of k elementary events, then

$$\mathbb{P}(A) = \frac{Card(A)}{Card(\Omega)} = \frac{\text{Number of favorable cases}}{\text{Number of possible cases}}. \tag{2.3}$$

2.2 Conditioning and independence

2.2.1 Conditioning

Definition 2.4 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, A and B be two events. The conditional probability of event A given event B is defined as the quotient

$$\mathbb{P}(A/B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ , with } \mathbb{P}(B) \neq 0 \text{ noted by } \mathbb{P}_B(A). \tag{2.4}$$

$\mathbb{P}_B(.)$ defines a probability on Ω.

Example 2.1 An urn contains 5 white balls numbered from 1 to 5 and two green balls numbered from 1 to 2. We draw a ball.

What is the probability of obtaining a white ball given that it is numbered 1.

Solution. Let event A be the event "the ball is white" and event B be the event "the ball is numbered 1. ■

Then $\mathbb{P}(B) = \frac{2}{7}$, The event $A \cap B$ is "the ball is white and numbered 1", then $\mathbb{P}(A \cap B) = \frac{1}{7}$.

Therefore $\mathbb{P}(A/B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{2}$.

Theorem 2.1 *Let A and B be two events in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A/B)\,\mathbb{P}(B) \tag{2.5}$$

formula of compound probabilities.

If A and B are two independent events and $\mathbb{P}(B) \neq 0$, then

$$\mathbb{P}_B(A) = \mathbb{P}(A/B) = \mathbb{P}(A). \tag{2.6}$$

2.2.2 Independence

Definition 2.5 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, we say that two events A and B are independent if we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B). \tag{2.7}$$

m events A_1, $A_2, ..., A_m$ are independent if

$$\mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)\,...\mathbb{P}(A_m). \tag{2.8}$$

2.3 Bayes' formula

Complete system of events:

Let $(A_i)_{i \in I}$ be a sequence of events. We call the family $(A_i)_{i \in I}$ a complete system of events when

1) $\forall (i,j) \in I^2$, such that $i \neq j$, $A_i \cap A_j = \phi$;

2) $\bigcup_{i \in I} A_i = \Omega$.

Formula of total probabilities Let $(A_i)_{i\in I}$ be a complete system of events with all non-zero probabilities. For any event B, we have:

$$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B \cap A_i) = \sum_{i\in I} \mathbb{P}(B/A_i)\,\mathbb{P}(A_i) \tag{2.9}$$

Bayes' formula: Let $(A_i)_{i\in I}$ be a complete system of events with non-zero probabilities. For any event B with non-zero probability, we have:

$$\mathbb{P}(A_i/B) = \frac{\mathbb{P}(A_i)\,\mathbb{P}(B/A_i)}{\sum_{j\in I} \mathbb{P}(A_j)\,\mathbb{P}(B/A_j)}. \tag{2.10}$$

CHAPTER 3

Random Variables and Standard Probability Distributions

3.1 Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random variable is any function $X : \Omega \to \mathbb{R}$ that satisfies:

$$X(\Omega) = \{a / \exists \omega \in \mathbb{R} / a = X(\omega)\}. \tag{3.1}$$

$\forall a\ \mathbb{R} : X^{-1}(\{a\}) \in \mathcal{F}$, $X^{-1}(\{a\}) = \{\omega : \omega \in \Omega \text{ and } X(\omega) = a\}$.

3.1.1 Various types of random variables

Discrete random variable: A random variable is said to be discrete when $X(\Omega)$ has a finite number or a countably infinite number of elements.

Probability density: A probability function (probability density) f such that

$$f(x_i) = \mathbb{P}(X = x_i) \tag{3.2}$$

is a function that satisfies the following two conditions:

1) $f(x_i) \geq 0$

2) $\sum_X f(x_i) = 1$.

The probability distribution: The probability distribution is generally given by the following table

X_i	X_1	X_2	X_3	.....
$f(X_i)$	$f(X_1)$	$f(X_2)$	$f(X_3)$	.....

Table 2

Distribution function: The cumulative distribution function of the variable X is the probability that $X \leq x$, denoted by $F(x)$.

$$F(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f(x_i). \tag{3.3}$$

Characteristics of random variables:

Mathematical Expectation: The mathematical expectation of a random variable X is given by:

$$E(X) = \sum_{i=1}^{n} f(x_i)x_i = \sum_{i=1}^{n} \mathbb{P}_i x_i. \tag{3.4}$$

and is also referred to as the mean.

Variance and Moments: The mathematical expectation of the random variable X^k is called the non-central moment of order k of variable X, and is given by:

$$m_k = E(X^k) = \sum_{i=1}^{n} \mathbb{P}_i x_i^k. \tag{3.5}$$

Remark 3.1 -The mathematical expectation of the random variable $(X - E(X))^k = (X - m_1)^k$ is called the central moment of order k of variable X and is given by:

$$M_k = E(X - m_1)^k = \sum_{i=1}^{n} \mathbb{P}_i (x_i - m_1)^k. \tag{3.6}$$

Remark 3.2 The central moment of order 2 is called variance

$$Var(X) = E(X - m_1)^2 = E\left(X^2\right) - E(X)^2. \tag{3.7}$$

-The standard deviation σ_X is the square root of the variance

$$\sigma_X = \sqrt{Var(X)} \geq 0. \tag{3.8}$$

Continuous Random Variables: A random variable X is said to be continuous if the set $X(\Omega)$ is an interval or a union of intervals in the real numbers $\mathbb{R}$.

Mathematical Expectation: If the integral

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{3.9}$$

converges, its value will be called the mathematical expectation of the random variable X.

Probability Density: A function f is called the probability density of the variable X if

1) $f(x) \geq 0, \forall x \in \mathbb{R}$
2) $\int_{-\infty}^{+\infty} f(x)dx = 1.$

Distribution Function: The cumulative distribution function of the random variable X is called the function F such that:

$$F(x) = \int_{-\infty}^{x} f(t)dt. \tag{3.10}$$

Variance and Moments: -Non-central moment of order k is

$$m_k = E(X^k) = \int_{-\infty}^{+\infty} x^k f(x)dx. \tag{3.11}$$

-Central moment of order k is

$$M_k = E\left[(X - m_1)\right]^k = \int_{-\infty}^{+\infty} (x - m_1)^k f(x)dx. \tag{3.12}$$

-The variance is given by:

$$Var(X) = \int_{-\infty}^{+\infty} (x - m_1)^2 f(x)dx = E\left(X^2\right) - E(X)^2. \tag{3.13}$$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \mathbb{P}(X < x). \tag{3.14}$$

Consequence:

$$\mathbb{P}(a \leq X < b) = \int_a^b f(x)dx = F(b) - F(a). \tag{3.15}$$

Proposition 3.1 *Let X be a random variable and let $Y = aX + b$. We then*

$$E(Y) = aE(X) + b, \tag{3.16}$$

$$Var(Y) = a^2 Var(X). \tag{3.17}$$

3.2 Common Probability Distributions(Usual Probability Laws)

3.2.1 Common Discrete Probability Distributions

Bernoulli distribution:

Definition 3.1 It is said that the real random variable X follows the Bernoulli distribution with parameter p if it takes only the values 0 and 1, such that

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p = q. \tag{3.18}$$

It is denoted as: $X \sim \mathcal{B}(p)$.

Characteristics:

$$E(X) = p, \ Var(X) = pq, \quad \sigma_X = \sqrt{pq}. \tag{3.19}$$

Proof. We have:

$$E(X) = \sum_{i=1}^{2} x_i \mathbb{P}_i = 1 \times \mathbb{P}(X = 1) + 0 \times \mathbb{P}(X = 0) = p.$$

And

$$E\left[(X - E(X))^k\right] = E\left[(X - p)^k\right] = (-p)^k q + (1 - p)^k p.$$

Let

$$E\left[(X - E(X))^k\right] = (-p)^k q + (1 - p)^k p.$$

In particular

$$\begin{aligned} V(X) &= E\left[(X - E(X))^2\right] = (-p)^2 q + (1 - p)^2 p \\ &= p^2 q + (1 - p)^2 p = pq(p + q) = pq. \end{aligned}$$

■

Binomial distribution:

Definition 3.2 A random variable X follows a **Binomial distribution** with parameters n and p if it has the following probability density:

$$\forall\ k = 0, 1, 2, ..., n;\ \ \mathbb{P}(X = k) = C_n^k p^k q^{n-k}, \quad q = 1 - p \tag{3.20}$$

It is denoted as: $X \sim \mathcal{B}(n, p)$.
A Bernoulli variable is a specific case of a binomial variable.

$$X \sim \mathcal{B}(1, p). \tag{3.21}$$

Characteristics:

$$E(X) = np,\ Var(X) = np(1-p) = npq, \quad \sigma_X = \sqrt{npq}. \tag{3.22}$$

Proof. We have

$$\sum_{k=0}^{n} f(x) = \sum_{k=0}^{n} \mathbb{P}(X = k) = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k}. \tag{3.23}$$

This results in, using the binomial formula.

$$\sum_{k=0}^{n} f(x) = (p + (1-p))^n = 1. \tag{3.24}$$

Thus, if we take, for any integer k less than or equal to n,

$$\mathbb{P}(X = k) = f(x), \tag{3.25}$$

we define a discrete random variable and its probability distribution.
The definition of $E(X)$ yields

$$E(X) = \sum_{k=0}^{n} k C_n^k p^k (1-p)^{n-k} = \sum_{k=1}^{n} k C_n^k p^k (1-p)^{n-k} \tag{3.26}$$

Since, we have $C_n^k = \frac{n}{k} C_{n-1}^{k-1}$, it follows

$$E(X) = np \sum_{k=1}^{n} C_{n-1}^{k-1} p^{k-1} (1-p)^{n-k} \tag{3.27}$$

Therefore, we can write again using the binomial formula

$$E(X) = np\,[p + (1-p)]^{n-1} = np. \tag{3.28}$$

Similarly, we have

$$E\left[X\left(X-1\right)\right]=\sum_{k=0}^{n}k\left(k-1\right)C_{n}^{k}p^{k}\left(1-p\right)^{n-k} \tag{3.29}$$

Since, we have

$$C_{n}^{k}=\frac{n\left(n-1\right)}{k\left(k-1\right)}C_{n-2}^{k-2} \tag{3.30}$$

It follows

$$\begin{aligned} E\left[X\left(X-1\right)\right] &= n\left(n-1\right)p^{2}\sum_{k=0}^{n}C_{n-2}^{k-2}p^{k-2}\left(1-p\right)^{n-k} \\ &= n\left(n-1\right)p^{2}\left[p+\left(1-p\right)\right]^{n-2}=n\left(n-1\right)p^{2} \end{aligned} \tag{3.31}$$

Hence,

$$\begin{aligned} E\left(X^{2}\right) &= E\left[X\left(X-1\right)\right]+E\left[X\right] \\ &= n\left(n-1\right)p^{2}+np \end{aligned}$$

Next,

$$\begin{aligned} Var\left(X\right) &= E\left(X^{2}\right)-\left[E\left(X\right)\right]^{2} \\ &= np\left(1-p\right)=npq. \end{aligned} \tag{3.32}$$

■

Example 3.1 We toss a fair money 10 times. What is the probability of getting a total of 8 heads? Let X be a random variable that assigns the number of heads in these 10 money tosses.

We have: $X\sim\mathcal{B}\left(10,\frac{1}{2}\right)$
$\mathbb{P}\left(X=8\right)=C_{10}^{8}\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{10-8}=0.0439$

Poisson distribution:

Definition 3.3 A random variable X follows a **Poisson distribution** with parameter $\lambda\left(\lambda>0\right)$ if it has the following probability density:

$$\mathcal{P}\left(k,\lambda\right)=\ \mathbb{P}\left(X=k\right)=\frac{e^{-\lambda}\lambda^{k}}{k!},\quad k=0,\ 1,\ 2,... \tag{3.33}$$

Characteristics:

$$E(X) = \lambda, \quad Var(X) = \lambda, \quad \sigma_X = \sqrt{\lambda}. \tag{3.34}$$

Proof. We have

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1 \tag{3.35}$$

Then the relation $\mathbb{P}(X = k) = f(x)$ defines a discrete random variable X and its probability distribution.

The definition of $E(X)$ yields

$$E(X) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \tag{3.36}$$

As

$$\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{\lambda} \tag{3.37}$$

it follows

$$E(X) = \lambda.$$

Similarly, we have

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} \\ &= \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \end{aligned} \tag{3.38}$$

Since, $\sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}$, we obtain

$$E[X(X-1)] = \lambda^2$$

Therefore

$$\begin{aligned} E\left(X^2\right) &= E[X(X-1)] + E[X] \\ &= \lambda^2 + \lambda \end{aligned}$$

This results in

$$\begin{aligned} Var(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= \lambda. \end{aligned} \tag{3.39}$$

■

Hypergeometric distribution

Among the N individuals in a population, Np have a certain characteristic A. We simultaneously draw n individuals. Let X be the random variable representing characteristic A. The probability that we have drawn x individuals possessing the studied characteristic is:

$$f(x) = \mathbb{P}(X = x) = \frac{C_{Np}^{x} C_{N(1-p)}^{n-x}}{C_{N}^{n}} \tag{3.40}$$

Characteristics: We have

$$E(X) = np; \quad Var(X) = \frac{N-n}{N-1} np(1-p). \tag{3.41}$$

Proof.

$$\begin{aligned} E(X) &= \sum_{x=0}^{x=n} x.f(x) = np. \sum_{x-1=0}^{x-1=n-1} \frac{C_{Np-1}^{x-1} C_{N(1-p)}^{(n-1)-(x-1)}}{C_{N-1}^{n-1}} \\ &= np \end{aligned}$$

The expectation of X is independent of N and identical to that of the $\mathcal{B}(n,p)$ distribution.

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{x=n} x(x-1) \frac{C_{Np}^{x} C_{N(1-p)}^{n-x}}{C_{N}^{n}} \\ &= np \frac{(Np-1)(n-1)}{N-1} \sum_{x-2=0}^{x-2=n-2} \frac{C_{Np-2}^{x-2} C_{N(1-p)}^{(n-2)-(x-2)}}{C_{N-2}^{n-2}} \end{aligned}$$

Hence,

$$E[X(X-1)] = np \frac{(Np-1)(n-1)}{N-1}$$

So,

$$Var(X) = E[X(X-1)] + E(X) - [E(X)]^2.$$

Then,

$$Var(X) = \frac{N-n}{N-1} np(1-p).$$

■

Geometric distribution

Definition 3.4 Consider a sequence of Bernoulli trials with parameter p. The probability that the first success occurs on the k^{th} trial is:

$$\mathbb{P}(X = k) = q^{k-1} p \text{ with } q = 1 - p \tag{3.42}$$

Characteristics: We have

$$E(X) = \frac{1}{p}; \quad Var(X) = \frac{1-p}{p^2}. \tag{3.43}$$

Proof. We have

$$\begin{aligned} E(X) &= \sum_{i=1}^{\infty} ipq^{i-1} = p\sum_{i=1}^{\infty} iq^{i-1} \\ &= \frac{p}{q}\sum_{i=1}^{\infty} iq^{i} = \frac{p}{q}S \end{aligned} \tag{3.44}$$

The sum

$$S = q + 2q^2 + 3q^3 + ...$$

Therefore,

$$Sq = q^2 + 2q^3 + 3q^4 + ...$$

Then,

$$\begin{aligned} S - Sq &= q + q^2 + q^3 + ... \\ &= \frac{q}{1-q} \end{aligned}$$

consequently,

$$S = \frac{q}{(1-q)^2} = \frac{q}{p^2}$$

Finally,

$$E(X) = \frac{p}{q}S = \frac{p}{q}\frac{q}{p^2} = \frac{1}{p}.$$

■

For the variance, the same procedure.

3.2.2 Common continuous probability distributions (Usuel Continuous Probability Laws)

Uniform distribution:

Definition 3.5 A continuous random variable X follows a uniform distribution over $[a, b]$ if its probability density function is defined by the function f:

$$f(x) = \begin{cases} 0, & \text{if } x \notin [a,b] \\ \frac{1}{b-a}, & \text{if } x \in [a,b] \end{cases} \tag{3.45}$$

It is denoted as: $X \sim \mathcal{U}_{[a,b]}$.

Characteristics:

$$E(X)=\frac{a+b}{2},\ Var(X)=\frac{(b-a)^2}{12},\quad \sigma_X=\sqrt{\frac{(b-a)^2}{12}}. \tag{3.46}$$

Proof. By definition of $E(X)$, we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} xf(x)\,dx=\int_a^b \frac{x}{b-a}dx \\ &= \frac{b+a}{2}. \end{aligned} \tag{3.47}$$

Similarly,

$$\begin{aligned} E\left(X^2\right) &= \int_{-\infty}^{+\infty} x^2 f(x)\,dx=\int_a^b \frac{x^2}{b-a}dx \\ &= \frac{b^2+ab+a^2}{3}. \end{aligned} \tag{3.48}$$

Since,

$$\begin{aligned} Var(X) &= E\left(X^2\right)-[E(X)]^2 \\ &= \frac{(a-b)^2}{12}, \end{aligned}$$

it results

$$Var(X)=\frac{(a-b)^2}{12}. \tag{3.49}$$

■

Exponential distribution:

Definition 3.6 A continuous random variable X follows an exponential distribution with parameter λ if its probability density function is defined by the function f:

$$f(x)=\begin{cases} \lambda\, e^{-\lambda x}, & \text{if } x\geq 0, \\ 0, & \text{if } x<0. \end{cases} \tag{3.50}$$

Charactcristics:

$$E(X)=\frac{1}{\lambda},\ Var(X)=\frac{1}{\lambda^2},\quad \sigma_X=\frac{1}{\lambda}. \tag{3.51}$$

Proof. By definition of $E(X)$

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x f(x)\, dx = \int_{0}^{+\infty} \lambda x e^{-\lambda x} dx \\ &= \lim_{x \to +\infty} \int_{0}^{x} \lambda t e^{-\lambda t} dt, \end{aligned} \tag{3.52}$$

integration by parts yields

$$\int_{0}^{x} \lambda t e^{-\lambda t} dt = -x e^{-\lambda x} + \frac{1}{\lambda}\left(1 - e^{-\lambda x}\right)$$

Therefore,

$$E(X) = \frac{1}{\lambda}.$$

Similarly,

$$\begin{aligned} E\left(X^k\right) &= \int_{-\infty}^{+\infty} x^k f(x)\, dx = \int_{0}^{+\infty} \lambda x^k e^{-\lambda x} dx \\ &= \lim_{x \to +\infty} \int_{0}^{x} \lambda t^k e^{-\lambda t} dt, \end{aligned} \tag{3.53}$$

By integrating by parts, we obtain

$$\int_{0}^{x} \lambda t^k e^{-\lambda t} dt = -x^k e^{-\lambda x} + \frac{k}{\lambda} \int_{0}^{x} \lambda t^{k-1} e^{-\lambda t} dt$$

Consequently,

$$\begin{aligned} E\left(X^k\right) &= \frac{k}{\lambda}.\frac{k-1}{\lambda}....\frac{2}{\lambda} E(X) \\ &= \frac{k!}{\lambda^k} \end{aligned}$$

Finally,

$$\begin{aligned} Var(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= \frac{1}{\lambda^2}. \end{aligned} \tag{3.54}$$

■

Normal distribution (Standard normal distribution or Gaussian distribution):

Definition 3.7 A continuous random variable X follows a standard normal distribution denoted by $\mathcal{N}(0,1)$, if its probability density function is defined on the real numbers $\mathbb{R}$ by the function f:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.55}$$

f being a density, the area between the x-axis and the curve $y = f(x)$ has a finite area equal to 1.

The distribution function of X: $x \mapsto \phi(x) = \int_{-\infty}^{x} f(t)\, dt$.

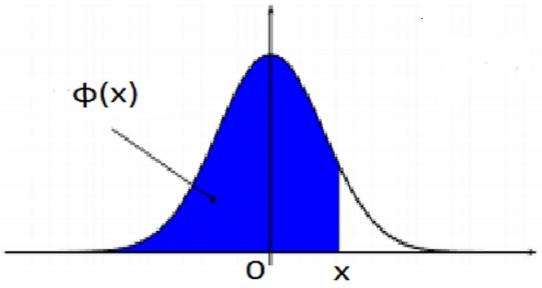

Figure 1

$$\phi(-x) = 1 - \phi(x). \tag{3.56}$$

Characteristics:

$$E(X) = 0,\ Var(X) = 1, \quad \sigma_X = 1. \tag{3.57}$$

Proof. According to the definition of

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} dx \tag{3.58}$$

The function $x \longmapsto x e^{-\frac{x^2}{2}}$ is odd, so we have

$$E(X) = 0 \tag{3.59}$$

Similarly,

$$Var(X) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx \tag{3.60}$$

The function $x \longmapsto x^2 e^{-\frac{x^2}{2}}$ is even, so we have,

$$\begin{aligned} Var(X) &= 2\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx \\ &= \frac{2}{\sqrt{\pi}} \int_0^{+\infty} x^2 e^{-x^2} dx \end{aligned}$$

Let's make the change of variables using polar coordinates, which easily gives us

$$Var(X) = \frac{2}{\sqrt{\pi}} \times \frac{\sqrt{\pi}}{2} = 1. \tag{3.61}$$

■

Remark 3.3 if we have the random variable $Y = \sigma X + \mu$, then

$$\begin{aligned} E(Y) &= E(\sigma X + \mu) = \sigma E(X) + \mu \\ &= \mu. \end{aligned} \tag{3.62}$$

And

$$Var(Y) = \sigma^2 Var(X) = \sigma^2. \tag{3.63}$$

Cauchy Distribution

Definition 3.8 A continuous random variable X follows a Cauchy distribution if its probability density function is:

$$f(x) = \frac{k}{1+x^2}, \quad k > 0, \ x \in \mathbb{R}. \tag{3.64}$$

Remark 3.4 $E\left(X^k\right)$ does not exist for $k \geq 1$.

Gamma Distribution

Definition 3.9 A positive random variable X follows a Gamma distribution with parameter k if its density is:

$$f(x) = \frac{1}{\Gamma(k)} e^{-x} x^{k-1} \tag{3.65}$$

where $\Gamma(k) = \int_0^{+\infty} t^{k-1} e^{-t} dt$.

Characteristics:

$$E(X) = k,\ Var(X) = k. \tag{3.66}$$

We have, when k is a natural integer

$$\Gamma(k) = (k-1)! \tag{3.67}$$

Proof. By definition of $E(X)$, we have

$$\begin{aligned} E(X) &= \int_0^{+\infty} xf(x) = \frac{1}{\Gamma(k)} \int_0^{+\infty} x^k e^{-x} dx \\ &= \frac{\Gamma(k+1)}{\Gamma(k)} = \frac{k\Gamma(k)}{\Gamma(k)} = k. \end{aligned} \tag{3.68}$$

Similarly,

$$\begin{aligned} E\left(X^2\right) &= \int_0^{+\infty} x^2 f(x) = \frac{1}{\Gamma(k)} \int_0^{+\infty} x^{k+1} e^{-x} dx \\ &= \frac{\Gamma(k+2)}{\Gamma(k)} = \frac{k(k+1)\Gamma(k)}{\Gamma(k)} = k(k+1) \end{aligned} \tag{3.69}$$

Therefore,

$$\begin{aligned} Var(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= k(k+1) - k^2 = k. \end{aligned} \tag{3.70}$$

■

Part II

Statistics

The statistis The word 'statistics' comes from the German 'Statistik,' which in the mid-17th century referred to the analysis of data relevant to the state. The processing of a large amount of numerical data that is sorted, classified, or summarized corresponds to what is now called 'statistics' in the plural. They are distinct from 'statistics' in the singular, which corresponds to the modeling of this data, seen as the results of experiments in the presence of randomness, and the study of this randomness. The emergence of statistics can be dated to the early 19th century, with the study of data from astronomy on the positions and trajectories of planets. In particular, in 1805, Adrien-Marie Legendre (1752-1832) introduced the method of least squares to estimate coefficients from data, and in 1809, Carl Friedrich Gauss (1777-1855), using a model of errors based on the normal distribution, found that by maximizing the density of the normal distribution of errors (i.e., the likelihood or posterior distribution), he could derive the least squares estimate. These works influenced Pierre-Simon Laplace (1749-1827), who in 1810 showed that the normal distribution naturally emerges as the distribution of errors thanks to the central limit theorem. In his book on the 'average man' in 1835, Adolphe Quetelet (1796-1874) used Laplace's results to analyze social data using the normal distribution and demonstrated the stability of this data over several years.

It wasn't until the end of the 19th century for a new breakthrough in the field of statistics. In 1885, Francis Galton (1822-1911) presented a study on the height of boys based on the average height of their parents. He observed both a phenomenon of dependence, which would be translated into a correlation effect, and a return to the mean or regression. Karl Pearson (1857-1936) and Udny Yule (1871-1951), building upon the work of Francis Edgeworth (1845-1926) on multidimensional normal distributions, extended linear regression to a broader framework. It's also important to highlight the χ^2 goodness-of-fit tests introduced by Pearson in biometrics in the late 19th century.

At the beginning of the 20th century, the 1920s were marked by the foundational work of Ronald Fisher (1890-1962), which was motivated by agronomy problems. Fisher introduced concepts such as statistical modeling, sufficiency, and maximum likelihood estimation. The use of statistical models allowed for the analysis of small datasets. Similarly on the same theme, we note the contributions of William Gosset (1876-1937) for Gaussian samples. Working for the Guinness brewery, he used the pseudonym 'Student' to publish his work.

Motivated by the study of the effects of different treatments in agriculture, Jerzy Neumann (1894-1981) introduced interval estimation and hypothesis testing in 1934, and further developed these concepts with Egon Pearson (1895-1980). The term "null hy-

pothesis" comes from the hypothesis corresponding to the absence of an effect from the treatment being considered. In 1940, Abraham Wald (1902-1950) proposed a unified view of the theory of estimation and hypothesis testing.

Starting from the 1950s, statistics experienced exponential growth with applications in all fields: engineering, experimental sciences, social sciences, medicine and life sciences, economics, etc. It has become an indispensable tool for the analysis and understanding of data.

CHAPTER 4

Statistical Series with One Variable

4.1 Basic Definitions

Statistics is a scientific method that involves gathering numerical data about large sets, and then analyzing, interpreting, and critiquing this data.

Population : The entire set under observation that will undergo statistical analysis. Each element within this set is an individual or statistical unit.

Sample: It's a subset of the considered population.

- The number of individuals in the sample is referred to as the sample size.

Characteristic: It's the specific property or aspect that one intends to observe within the population or sample. A characteristic that is the subject of a study is also referred to as a statistical variable.

Absolute frequency: The absolute frequency of a class is the number of elements from the population observed within that class. It's denoted as: n_i

Frequency: Frequency is the ratio of this frequency to the total absolute frequency of the population. It is denoted by f_i:

$$f_i = \frac{n_i}{n}. \tag{4.1}$$

n is the total number of the absolute frequencies.

Example 4.1 Consider the following statistical series:

12; 13; 4; 13; 12; 14; 15; 7; 15; 13; 12; 15; 7

x_i	4	7	12	13	14	15
n_i	1	2	3	3	1	3
f_i	$\frac{1}{13}$	$\frac{2}{13}$	$\frac{3}{13}$	$\frac{3}{13}$	$\frac{1}{3}$	$\frac{3}{13}$

Table 3

$$\sum_{i=1} n_i = n, \quad \sum_{i=1} f_i = 1. \tag{4.2}$$

Accumulated increasing absolute frequency: The accumulated increasing frequency is the number $N_i \uparrow$ such that

$$N_k \uparrow = n_1 + n_2 + ... + n_k. \tag{4.3}$$

Accumulated Decreasing absolute frequency: The decremental accumulated frequency is the number $N_i \downarrow$ such that

$$N_k \downarrow = n - (n_1 + n_2 + ... + n_{k-1}). \tag{4.4}$$

Example 4.2 Consider the following statistical series:

12; 13; 4; 13; 12; 14; 15; 7; 15; 13; 12; 15; 7

x_i	n_i	$N_k \uparrow$	$N_k \downarrow$
4	1	1	13
7	2	3	12
12	3	6	10
13	3	9	7
14	1	10	4
15	3	13	3

Table 4

Increasing cumulative frequency: The increasing cumulative frequency is the number $F_i \uparrow$ such that

$$F_k \uparrow = \frac{N_k \uparrow}{\sum_{i=1} n_i}. \tag{4.5}$$

Decreasing Cumulative Frequency: The decreasing cumulative frequency is the number $F_i \downarrow$ such that

$$F_k \downarrow = \frac{N_k \downarrow}{\sum_{i=1} n_i}. \tag{4.6}$$

Example 4.3 Let's consider the following statistical series:

12; 13; 4; 13; 12; 14; 15; 7; 15; 13; 12; 15; 7

x_i	n_i	$F_k \uparrow = \frac{N_k \uparrow}{\sum_{i=1} n_i}$	$F_k \downarrow = \frac{N_k \downarrow}{\sum_{i=1} n_i}$
4	1	$\frac{1}{13}$	1
7	2	$\frac{3}{13}$	$\frac{12}{13}$
12	3	$\frac{6}{13}$	$\frac{10}{13}$
13	3	$\frac{9}{13}$	$\frac{7}{13}$
14	1	$\frac{10}{13}$	$\frac{4}{13}$
15	3	1	$\frac{3}{13}$

Table 5

Different Types of Statistical Variables: * When the variable does not lend itself to numerical values, it is called qualitative (example: political opinions, eye colors...). It can be ordered or not, dichotomous or not.

* When the variable can be expressed numerically, it is called quantitative (or measurable). In this case, it can be discrete or continuous.

■ It is discrete if it takes only isolated values from each other. A discrete variable that takes only integer values is called discrete (example: number of children in a family).

■ It is called continuous when it can take all values within a finite or infinite interval (example: diameter of parts, salaries...).

4.2 Measures of Central Tendency

Central tendency parameters, or 'measures of central tendency,' are values that aim to best represent a set of data. The term 'central tendency' comes from the fact that these parameters provide an idea of what is happening at the center of a distribution, of a dataset.

Three measures of central tendency are distinguished:

4.2.1 The Mean

The mean is one of the fundamental parameters of central tendency, but it's not sufficient alone to characterize a distribution. Complementary to the mode and especially the median, the mean is undoubtedly the most calculated and commonly used measure when describing statistical data sets. There are several types of means, each suited to specific situations: arithmetic mean, geometric mean, harmonic mean, and quadratic mean.

1) Arithmetic Mean

- Simple Arithmetic Mean This is the simplest and most commonly used mean, although not always used appropriately. It is often denoted by $\bar{X}$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{4.7}$$

-Weighted arithmetic mean. The weighted arithmetic mean, let's say it right away, gives, in its classical usage (meaning when all individuals have the same weight), the same result as the simple arithmetic mean. Its formula is, however, different as it introduces the notion of weight through an additional term that can prove useful in certain situations, especially when the individuals composing a population do not have the same weight or coefficient: some individuals, for various reasons, have more influence in the said population than others. This can be the case, for example, when dealing with a series of grades with different coefficients. This average is written as follows:

$$\bar{X} = \frac{\sum_{i=1}^{m} n_i x_i}{\sum_{i=1}^{m} n_i} \tag{4.8}$$

Example 4.4 Consider the following statistical series:

12; 13; 4; 13; 12; 14; 15; 7; 15; 13; 12; 15; 7

$$\bar{X} = \frac{\sum_{i=1}^{m} n_i x_i}{\sum_{i=1}^{m} n_i} = \frac{3 \times 12 + 2 \times 13 + 1 \times 4 + 1 \times 14 + 3 \times 15 + 1 \times 7}{13} =$$

Example 4.5 Consider the following statistical series:

Weight	$[50, 55[$	$[55, 60[$	$[60, 65[$	$[65, 70[$
n_i	2	4	8	6

Table 6

We replace the x_i with the class midpoints c_i.

$$\bar{X} = \frac{\sum_{i=1} p_i c_i}{\sum_{i=1} p_i} = \frac{2 \times 52.5 + 5 \times 75.5 + 8 \times 62.5 + 6 \times 67.5}{20} =$$

2) The geometric mean

- Simple geometric mean Let $\{x_1,\ x_2, ..., x_n\}$ be a statistical series. The formula for the simple geometric mean of this series is given by:

$$\overline{G} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{4.9}$$

- The weighted geometric mean Let $\{x_1,\ x_2, ..., x_m\}$ be a statistical series and $\{n_1,\ n_2, ..., n_m\}$ be the corresponding frequencies. The formula for the **weighted geometric mean** of this series is given by:

$$\overline{G} = \sqrt[n]{\prod_{i=1}^{m} x_i^{n_i}} \tag{4.10}$$

The geometric mean is a tool used to calculate average rates, especially average annual rates.

3) The harmonic mean

- Simple harmonic mean The harmonic mean is used when one wants to determine an average ratio in areas where there are inverse proportional relationships.

Let $\{x_1,\ x_2, ..., x_n\}$ be a statistical series. The formula for the **simple harmonic mean** of this series is given by:

$$\overline{H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{4.11}$$

- For a given distance, the travel time is shorter the higher the speed.
- Rent in the private housing market is higher the smaller the size or surface area of the accommodation.

- The weighted harmonic mean Let $\{x_1,\ x_2, ..., x_m\}$ be a statistical series and $\{n_1,\ n_2, ..., n_m\}$ be the corresponding frequencies. The formula for the **weighted harmonic mean** of this series is given by:

$$\overline{H} = \frac{n}{\sum_{i=1}^{m} \frac{n_i}{x_i}} \tag{4.12}$$

4) The root mean square

- The simple root mean square An average that finds applications when dealing with phenomena that exhibit a sinusoidal nature with alternation between positive and negative values. It is, therefore, widely used in electricity. It is used to calculate the magnitude of a set of numbers. For your information.

Let $\{x_1,\ x_2, ..., x_n\}$ be a statistical series. The formula for the **simple root mean square** of this series is given by:

$$\overline{Q} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}. \tag{4.13}$$

-The weighted root mean square Let $\{x_1,\ x_2, ..., x_m\}$ be a statistical series and $\{n_1,\ n_2, ..., n_m\}$ be the corresponding frequencies. The formula for the weighted root mean square of this series is given by:

$$\overline{Q} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} n_i x_i^2}. \tag{4.14}$$

4.2.2 The mode

The mode (Mod) of a statistical series is the value of the most frequent or dominant characteristic in the sample. In other words, it's the value with the highest (absolute or relative) frequency. When the distribution has more than one mode, it's referred to as a 'multimodal' distribution (bimodal, trimodal, etc.). However, if we're dealing with grouped data in classes, the mode will relate to the class containing the largest number of individuals: this is then referred to as the modal class. Nevertheless, there might be instances where we're interested in obtaining the approximate or exact value of this mode. Therefore, it's recommended to follow the following approach:"

- To obtain an approximate value of the mode, calculate the mean of the class with the highest frequency.

- To obtain an exact value, the mode is calculated as follows:

$$Mod = x_m + \frac{\triangle\, i}{\triangle\, s + \triangle\, i} \times i, \tag{4.15}$$

where

x_m : Lower limit of the modal class;

i : Width of the modal class;

$\triangle\ i$: Absolute frequency difference between the modal class and the nearest lower class;

$\triangle\ s$: Absolute frequency difference between the modal class and the nearest upper class.

Example 4.6 Consider the following statistical series:

12;13;4;13;12;14;15;7;15;12;15;7

$Mod_1 = 12$; $Mod_2 = 15$.

Consider the following statistical series:

12;13;4;11;12;14;11;7;11;12;11;7

$Mod = 11$

Example 4.7 Consider the following statistical series:

Weights	[50,55[	[55,60[	[60,65[	[65,70[	[70,75[	[75,80[	[80,85[
ni	2	5	12	16	14	8	3

Table 7

The modal class is:$[65, 70[$

x_m =65

i =70-65=5

$\triangle\ i$=16-12=4

$\triangle\ s$ = 16-14=2

Therefore,

$$Mod = x_m + \frac{\triangle\ i}{\triangle\ s + \triangle\ i} \times i = 65 + \frac{4}{2+4} \times 5 \simeq 68.33.$$

4.2.3 The Median

In the calculation of the median (Med), two cases are distinguished:

1- If the variable is discrete:

Let n be the number of observations

- If n is even: the median is then equal to the average of the values that surround the middle of the series.

- If n is odd: it is possible to simply identify the value that divides the population into two equal frequencies. The central rank being equal to $[(n+1)/2]$.

2- If the variable is continuous and grouped into classes, we search for the class containing the $\frac{n}{2}$ individual of the sample. This class is called the median class. Assuming that all individuals in this class are uniformly distributed within it, the median is calculated using the following linear interpolation method:

$$Med = x_m + \frac{\frac{n}{2} - N_i \uparrow}{n_i} \times i \tag{4.16}$$

x_m : Lower limit of the median class;

i : Width of the median class;

n_i : Frequency of the median class;

$N_i \uparrow$: Cumulative frequency below x_m;

n : Sample size.

Example 4.8 Consider the following statistical series:

12; 13; 4; 13; 17; 14; 15; 7; 15; 12; 13; 7; 18

The frequency count $n = 13$ is even, so the median is the value at the $(\frac{n+1}{2})$th position, which is $\frac{13+1}{2} = 7$.

The order of values: 4;7;7;12;12;13;13;13;14;15;15;17;18. The value at the $7th$ position is $Med = 13$.

Example 4.9 Consider the following statistical series:

12; 13; 4; 17; 14; 15; 7; 15; 12; 20; 7; 18

The frequency count $n = 12$ is odd, so the median is the arithmetic mean of the values at the $\frac{n}{2}$th position, which is $\frac{12}{2} = 6$ and at the $(\frac{n}{2} + 1)$th position,which is$\frac{12}{2} + 1 = 7$.

The order of values: 4; 7; 7; 12; 12; 13; 14; 15; 15; 17; 18; 20

The value at the $6th$ position is 13

The value at the $7th$ position is 14, Then $Med = \frac{13+14}{2} = 13.5$

Example 4.10 Consider the following statistical series:

Weights	[50,55[	[55,60[	[60,65[	[65,70[	[70,75[	[75,80[	[80,85[
ni	2	5	12	16	14	8	3
Ni ↑	2	7	19	35	49	57	60

Table 8

We have $\frac{n}{2} = \frac{60}{2} = 30$ The median class is the first class where the cumulative frequency becomes greater than or equal to 30; Therefore, the median class is: $[65, 70[$

x_m =65

i =70 − 65 = 5

n_i =16

$N_i \uparrow$ =19

$n = 60$

Then, $Med = x_m + \frac{\frac{n}{2} - N_i\uparrow}{n_i} \times i = 65 + \frac{\frac{60}{2} - 19}{16} \times 5 \simeq 68.44$

4.3 Graphical Representations

4.3.1 Case of discrete variables

Diagram of bands A diagram of bands is a graphical representation of statistical data using segments. The values of the studied characteristic are represented on the horizontal axis, and the absolute frequencies (or frequencies) are represented on the vertical axis. Each value corresponds to a band. The heights of the bars are proportional to the absolute frequencies (or frequencies) being represented.

Example 4.11 Consider the following statistical series:

xi	1	2	3	4	5	6	7	8	9 or more
ni	8000	8100	4500	3500	1500	500	300	200	300

Table 9

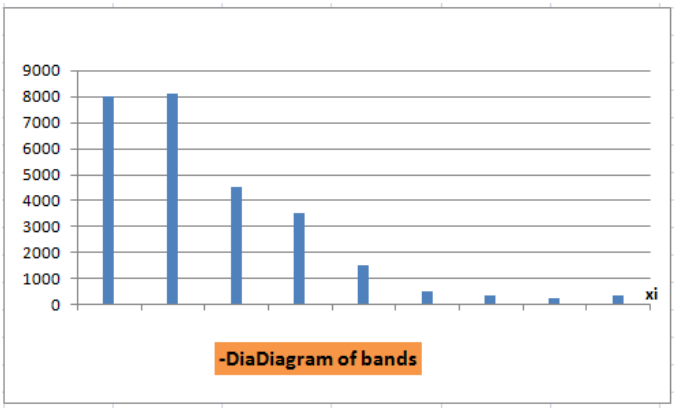

Figure 2

Cumulative increasing and cumulative decreasing curves:

Definition 4.1 The cumulative curve is a graphical representation of the distribution of cumulative absolute frequencies or cumulative frequencies. It differs depending on whether one is working with an observed distribution or with a grouped distribution (class)

Example 4.12 Consider the following statistical series:

xi	ni	Ni ↑	Ni ↓
4	3	3	13
7	2	5	10
10	3	8	8
12	5	13	5

Table 10

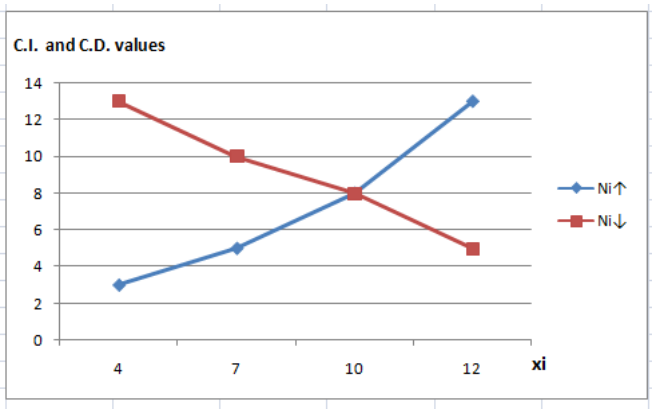

Figure 3

Example 4.13 Consider the following statistical series:

Classes	ni	ci	Ni ↑	Ni ↓
[0,5[	4	2.5	4	28
[5,10[	6	7.5	10	24
[10,15[	3	12.5	13	18
[15,20[	7	17.5	20	15
[20,25[	8	22.5	28	8

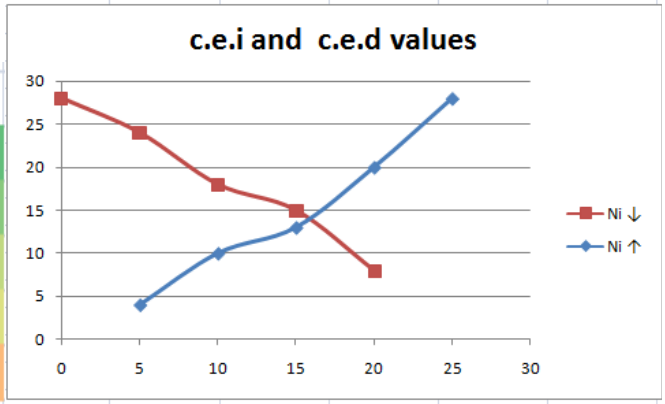

Table 11 *Figure* 4

4.3.2 Case of continuous variables

Histogram: An histogram is a diagram composed of adjacent rectangles whose areas are proportional to the frequencies (or relative frequencies) and whose bases are determined by class intervals.

Polygon of frequencies: Polygon of frequencies (or relative frequencies) formed by connecting the midpoints of the upper bases of the rectangles with a polygon.

Frequency curve: Frequency curve (or relative frequency curve) formed by connecting the midpoints of the upper bases of the rectangles with a curve.

Example 4.14 The human resources manager of a company has recorded the following statistical distribution corresponding to the length of service of managerial personnel in the company, expressed in years:

Classes	ni
[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

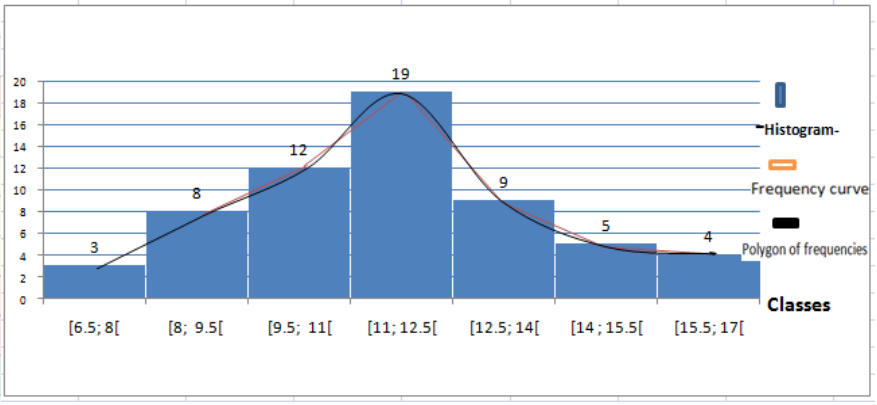

Table 12 *Figure* 5

4.3.3 Case of qualitative variables:

Circular diagram: A circular diagram is a graph consisting of a circle divided into sectors, where the central angles are proportional to the absolute frequencies (or the frequencies). Consequently, the dissected areas are proportional to the frequencies. The angle α_i of a category with a frequency of n_i is given in degrees by:

$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360 \tag{4.17}$$

	ni	fi	αi
Poor	200	0.3	120
middle	100	0.2	60
Rich	300	0.5	180
		1	360

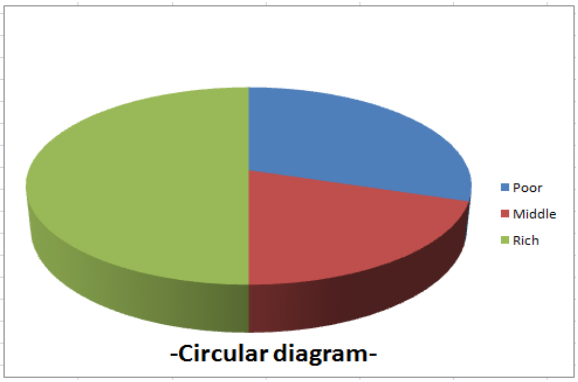

Table 13 *Figure* 6

4.4 Positional Characteristics

Quartiles, deciles, and percentiles are characteristics that correspond to the same kind of concern as the median. They are values of the variable that correspond to cumulative frequencies:

$\frac{n}{4}, \frac{2n}{4}, \frac{3n}{4}$ for quartiles; the 2^{rd} quartile is the median.

$\frac{n}{10}, \frac{2n}{10}, ..., \frac{9n}{10}$ for deciles; the 5^{rd} decile is the median.

$\frac{n}{100}, \frac{2n}{100}, ..., \frac{99n}{100}$ for percentiles; the 50^{rd} percentile is the median

They are called position characteristics since they allow the placement of variable values.

Example 4.15 Case of a continuous variable:

Consider the following statistical series (Values in grams):

Values in g	[700,750[	[750,800[	[800,840[	[840,880[	[880,900[	>900
ni	10	23	4	15	32	16
Ni ↑	10	33	37	52	84	100

Table 14

750	10
Q_1	25
800	33

Hence,

$$\frac{Q_1 - 750}{800 - 750} = \frac{25 - 10}{33 - 10}$$

Which gives $Q_1 = 782,61g$.

Similarly, we can find the 3^{rd} quartile:

$$\frac{Q_3 - 880}{920 - 880} = \frac{75 - 52}{84 - 52}$$

which gives $Q_3 = 908,75$, namely $909g$.

We would calculate the deciles in the same way. Let's provide the value for $D_1 = 750g$ and $D_9 = 935g$.

Example 4.16 Case of a discrete variable:

Consider the following statistical series:

We recorded the height in centimeters of basketball players on a team, 302, 187, 185, 206, 180, 188, 198, 195, 200, 195, 218, 210.

Let's arrange the series in ascending order.

xi	180	185	187	188	195	198	200	203	206	210	218
ni	1	1	1	1	2	1	1	1	1	1	1
Ni ↑	1	2	3	4	6	7	8	9	10	11	12

Table 15

The order of the value of Q_1 is: $\frac{1}{4} \times 12 = 3$, so $Q_1 = 187$.
The order of the value of Q_3 is: $\frac{3}{4} \times 12 = 9$, so $Q_1 = 203$.

4.5 Measures of Dispersion

The extent: The extent, denoted as E, represents the difference between the extreme values of the distribution:

$$E = Max\left(x_i\right) - Min\left(x_i\right). \tag{4.18}$$

Variance: Variance is the most commonly used measure of dispersion, denoted as: $var\left(X\right)$

$$\begin{aligned} var\left(X\right) &= \frac{\sum_{i=1}^{k} n_i \left(x_i - \bar{X}\right)^2}{n} \\ &= \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - \bar{X}^2 = \sum_{i=1}^{k} f_i x_i^2 - \bar{X}^2 \end{aligned} \tag{4.19}$$

Standard deviation: The standard deviation, σ_X, is the square root of the Variance:

$$\sigma_X = \sqrt{var\left(X\right)}. \tag{4.20}$$

The coefficient of variation: The coefficient of variation (CV) is given by:

$$CV = \left(\frac{\sigma}{\bar{X}}\right) \times 100. \tag{4.21}$$

- If $CV < 0.25 \rightarrow$ concentration.
- If $CV > 0.25 \rightarrow$ dispersion.

4.6 Shape Characteristics

There are two shape measures that characterize the shape of curves representing distributions:

- Asymmetry coefficient
- Kurtosis coefficient

4.6.1 Asymmetry coefficient

A coefficient is used to measure the asymmetry of a distribution

* If the coefficient is zero ($S_k = 0$), then it represents a perfectly symmetrical distribution. And mean = median = mode.

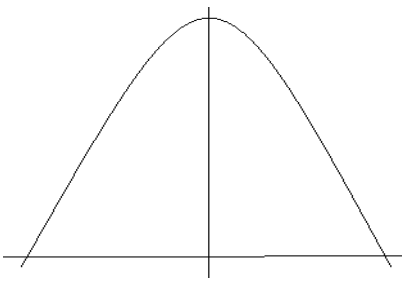

Figure 7

- Values are equally distributed on either side of the central value.

** If the coefficient is less than 0, then the distribution is on the lower side (spread to the left), then, mean < median < mode.

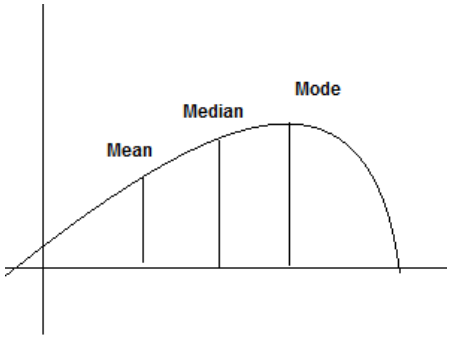

Figure 8

- The highest frequency value is located to the right of the mean.

*** If the coefficient is greater than 0, then the distribution is on the upper side (spread to the right), then mean > median > mode.

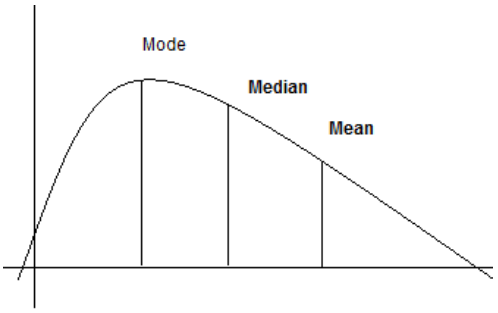

Figure 9

- The highest frequency value is located to the left of the mean.

Calculation of the coefficient: ♦ *Peasonıs skewness(asymitrical) coefficient*

It is denoted by S_P

$$S_P = \frac{3\left(\overline{X} - Med\right)}{\sigma_X} \text{ or } S_P = \frac{\overline{X} - Mod}{\sigma_X} \tag{4.22}$$

- If $S_P < 0$ left-asymmetrical (left-skewed) distribution.
- If $S_P = 0$ symmetrical distribution.
- If $S_P > 0$ right-asymmetrical distribution.

♦ *Yuleıs skewness coefficient*

It is denoted by S_Y

$$S_Y = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \tag{4.23}$$

- If $-1 \leq S_Y < 0$ left-asymmetrical (left-skewed) distribution.
- If $S_Y = 0$ symmetrical distribution.
- If $0 < S_Y \leq 1$ right-asymmetrical distribution.

♦ *Fisher skewness coefficient*
It is denoted by S_F

$$S_F = \frac{M_3}{\sigma_X^3} \tag{4.24}$$

where $M_3 = \frac{1}{n}\sum n_i \left(x_i - \overline{X}\right)^3$ is the third-order central moment.

- If $S_F < 0$ left-asymmetrical (left-skewed) distribution.
- If $S_F = 0$ symmetrical distribution.
- If $S_F > 0$ right-asymmetrical distribution.

4.6.2 Kurtosis Coefficient

To measure the kurtosis of the curve, we use

The Peason coefficient P_P: The Peason coefficient P_P is based on the fourth-order central moment

$$P_P = \frac{M_4}{\sigma_X^4} \tag{4.25}$$

- $P_P > 3$ leptokurtic or hypernormal curve.
- $P_P = 0$ normal curve.
- $P_P < 3$ platykurtic or hypornormal curve.

The Fisher coefficient P_F: The Fisher coefficient P_F is given by

$$P_F = \frac{M_4}{\sigma_X^4} - 3 \tag{4.26}$$

- $P_F > 0$ leptokurtic or hypernormal curve.
- $P_F = 0$ normal curve.
- $P_F < 0$ platykurtic or hypornormal curve.

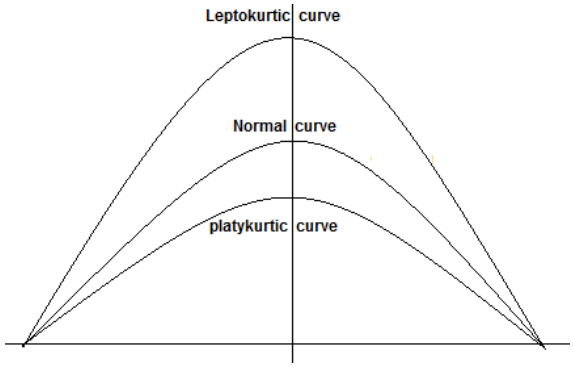

Figure 10

4.7 Statistical Tables

Data from a study can be represented in a table. However, it is possible to create a clearer table by **grouping the data into classes**. One selects classes that are not too numerous but sufficient to avoid any loss of information. It is important that these classes cover all the results and have no overlap; the difference between the two extremes is called **the class width**.

The number of classes can be determined using one of the following two formulas:

i) The Sturge Rule:

$$\text{number of classes} = 1 + (3.3 \log n) \tag{4.27}$$

ii) The Yule Rule:

$$\text{number of classes} = 5.2\sqrt[4]{n} \tag{4.28}$$

With n being the sample size

The class width is then given by:

$$\frac{\text{maximum value-minimum value}}{\text{number of classes}}. \tag{4.29}$$

Remark 4.1 In the case of a continuous statistical variable, we replace the x_i with the centers c_i of the i^{th} classes.

CHAPTER 5

Statistical Series with Two Variables

Definition 5.1 A statistical series with two numerical variables, X and Y, is a function that associates the value taken by both characteristics with each individual.

$$\begin{array}{rcl} (X,\ Y) & : & \Omega \to \mathbb{R} \\ \omega & \longmapsto & (X(\omega),\ Y(\omega)), \end{array} \tag{5.1}$$

with $\text{card}(X(\Omega)) \geq 2$ and $\text{card}(Y(\Omega)) \geq 2$.

5.1 Data Tables (Contingency Table) and Scatterplots

The main problem encountered in bivariate statistical series is primarily the question of whether or not there is a relationship between each of the variables.

Problem Statement For the sake of clarity, this course is developed based on the following example:

Example 5.1 The following table shows the evolution of the number of members in a rugby club from 2001 to 2006.

Scatterplot

The first step is to create a graph that represents the two statistical series above.

Definition 5.2 Let X and Y be two numerical statistical variables observed on n individuals. In an orthogonal coordinate system $\left(o;\ \vec{i}\,;\ \vec{j}\right)$, the set of n points with coordinates $(x_i,\ y_i)$ forms the scatterplot associated with this statistical series.

Definition 5.3 The joint distribution of the frequencies of X and Y will be called the set of information $(x_i,\ y_j,\ n_{ij})$ for $i = 1,\ ...,\ k$ and $j = 1,\ ...,\ \ell$.

5.2 Marginal Distributions, Conditional Distributions and Covariance

5.2.1 Marginal Distributions

We add the row and column totals to the contingency table.

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_ℓ	totaux
x_1	n_{11}		n_{1j}		$n_{1\ell}$	$n_{1\bullet}$
$\cdots$					$\cdots$	$\cdots$
x_i	n_{i1}		n_{ij}		$n_{i\ell}$	$n_{i\bullet}$
$\cdots$					$\cdots$	$\cdots$
x_k	n_{k1}		n_{kj}		$n_{k\ell}$	$n_{k\bullet}$
totaux	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet \ell}$	$N = n_{\bullet\bullet}$

Table 21

- In the right margin (row totals): the distribution of X: for each index i, the frequency $n_{i\bullet}$ is the total number of observations for the modality x_i of X, regardless of the category of Y. That is to say

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij} = \text{ total of row } i. \tag{5.2}$$

Definition 5.4 The k pairs $(x_i,\ n_{i\bullet})$ define the marginal distribution of the variable X.

- In the bottom margin (column totals): the distribution of Y: for each index j, the frequency $n_{\bullet j}$ is the total number of observations for the modality y_j of Y, regardless of the category of X. That is to say

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} = \text{ total of column } j. \tag{5.3}$$

Definition 5.5 The ℓ pairs $(y_j,\ n_{\bullet j})$ define the marginal distribution of the variable Y.

Remark 5.1

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = N. \tag{5.4}$$

Example 5.2 Consider the following table of a group of 100 people categorized by age group (X) and gender (Y).

X\Y	Men	Women	ni.
[0,18[	20	40	/60/
[18.50[	10	30	/40/
n.j	/30/	/70/	100

Table 22

Example 5.3 Let's add a row and a column to the table, and fill them with $n_{i.}$ and $n_{.j}$. These added row and column are the marginal distributions of the contingency table. Thus, the column $n_{i.}$ represents the marginal distribution of x, which means the possible values of x regardless of y. Similarly, the row $n_{.j}$ represents the marginal distribution of y, which means the values of y regardless of x.

5.2.2 Marginal Frequencies

The definitions of marginal frequencies are as follows:

Marginal frequencies of X:

$$f_{i.} = \frac{n_{i.}}{n_{..}} \qquad i = 1, ..., p \tag{5.5}$$

Marginal frequencies of Y::

$$f_{.j} = \frac{n_{.j}}{n_{..}} \qquad j = 1, ..., q \tag{5.6}$$

Example 5.4

X\Y	Men	Women	fi.
[0,18[	20	40	0.6
[18.50[	10	30	4
f.j	0.3	0.7	1

Table 23

5.2.3 Marginal Means and Marginal Variances

Marginal Means

The marginal means of X and Y are given by the following formulas:

$$\overline{\overline{X}} = \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} x_i \tag{5.7}$$

$$\overline{\overline{Y}} = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} y_j \tag{5.8}$$

Example 5.5 Consider the following contingency table

X\Y	4	9	ni.
5	5	4	9
5	2	3	5
n.j	7	7	14

Table 24

$$\begin{aligned} \overline{\overline{X}} &= \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} x_i = \frac{1}{14} (9 \times 3 + 5 \times 5) = \frac{26}{7} \\ \overline{\overline{Y}} &= \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} y_j = \frac{1}{14} (7 \times 5 + 7 \times 4) = \frac{9}{2} \end{aligned}$$

Marginal Variances

The marginal variances of X and Y are given by the following formulas:

$$\sigma_X^2 = \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} \left(x_i - \overline{\overline{X}} \right)^2 = \frac{1}{n_{..}} \sum_{i=1}^{p} n_{i.} x_i^2 - \overline{\overline{X}}^2 \tag{5.9}$$

$$\sigma_Y^2 = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} \left(y_j - \overline{\overline{Y}} \right)^2 = \frac{1}{n_{..}} \sum_{j=1}^{q} n_{.j} y_j^2 - \overline{\overline{Y}}^2 \tag{5.10}$$

5.2.4 Conditional Distributions

Definition 5.6 The distribution of observations according to the modalities of variable Y, given that variable X takes the modality x_i, is called the conditional distribution of Y for $X = x_i$.

- In row i of the contingency table, we read the distribution of variable Y given that $X = x_i$, denoted as $Y|_{X=x_i}$.

  Definition 5.7 The distribution of observations according to the modalities of variable X, given that variable Y takes the modality y_j, is called the conditional distribution of X for $Y = y_j$.

- In column j of the contingency table, we read the distribution of variable X given that $Y = y_j$, denoted as $X|_{Y=y_j}$.

Example 5.6 We keep the previous example

X\Y	4	9	ni.
5	/5/	4	9
5	/2/	3	5
n.j	7	7	14

Table 25

The boxed column represents the distribution of X when $Y = 4$.

X\Y	4	9	ni.
5	/5/	/4/	9
5	2	3	5
n.j	7	7	14

Table 26

The boxed row represents the distribution of Y when $X = 3$.

5.2.5 Conditional Means and Conditional Variances

Conditional Means

For each conditional distribution, we can calculate a mean. For example, in the previous table, we have two conditional means for X

$$\overline{X}_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} n_{ij} x_i \qquad 1 \leq j \leq p \tag{5.11}$$

$\overline{X}_1$ to denote the conditional mean of X when $Y = 4$.

$\overline{X}_2$ to denote the conditional mean of X when $Y = 9$.

In the same way

$$\overline{Y}_i = \frac{1}{n_{i\cdot}} \sum_{j=1}^{q} n_{ij} y_j \qquad 1 \leq i \leq q \tag{5.12}$$

$\overline{Y}_1$ to denote the conditional mean of Y when $X = 3$.

$\overline{Y}_2$ to denote the conditional mean of Y when $X = 5$.

Example 5.7 The previous example

$$\overline{X}_1 = \frac{1}{7} (5 \times 3 + 2 \times 5) \approx 3.57$$

and

$$\overline{Y}_2 = \frac{1}{5} (2 \times 4 + 3 \times 9) = 7$$

Conditional Variances

For each conditional distribution, one can calculate a variance. For example, in the previous table, we have two conditional variances of X.

$$Var\left(X_j\right) = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} n_{ij} \left(x_i - \overline{X}_j\right)^2 = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} n_{ij} x_i^2 - \overline{X}_j^2 \tag{5.13}$$

$Var\left(X_1\right)$ to refer to the conditional variance of X when $Y = 4$.

$Var\left(X_2\right)$ to refer to the conditional variance of X when $Y = 9$.

Similarly,

$$Var\left(Y_i\right) = \frac{1}{n_{i.}}\sum_{j=1}^{q} n_{ij}\left(y_j - \overline{Y}_i\right)^2 = \frac{1}{n_{i.}}\sum_{j=1}^{q} n_{ij}y_j^2 - \overline{Y}_i^2 \tag{5.14}$$

$Var\left(Y_1\right)$ to refer to the conditional variance of Y when $X = 3$.

$Var\left(Y_2\right)$ to refer to the conditional variance of Y when $X = 5$.

Example 5.8 Let's calculate the conditional variances for the data in the previous table.

5.2.6 Covariance

All statistical individuals would therefore have the same value on this variable, a value that would also happen to be the mean. Let's imagine that this variable starts to vary a little. Consequently, we will find some values that are a little higher than the mean, and others that are a little lower than the mean. If the variable varies a lot, we will find more and more individuals producing a value very different from the mean. Based on this, it seems quite natural to use the deviation from the mean to quantify the degree of variation exhibited by a given individual.

Let's suppose

$$x_i,\ \overline{x},\ y_i,\ \overline{y}$$

respectively the mean value observed on variable X, the mean value observed on variable Y, the measurement of variable X obtained for the i^{th} individual, and finally the measurement of variable Y obtained for the i^{th} individual.

So, the measure of variation on variable X for individual i can be directly given by the value

$$(x_i - \overline{x})$$

Similarly, the measure of variation on variable Y for individual i can be given by the value

$$(y_i - \overline{y})$$

Note that both of these values are signed, in the sense that a deviation in the direction where the value is greater than the mean will have a positive sign, whereas a variation where the measurement is smaller than the mean will result in a negative deviation.

Definition 5.8 The covariance of the double statistical series of variables x and y is defined as the real number.

$$Cov\,(x,\ y) = \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})\,. \tag{5.15}$$

For calculations, you can also use:

$$Cov\,(x,\ y) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{xy}. \tag{5.16}$$

Remark 5.2

$$Cov\,(x,\ x) = \sigma_{x^2} = V\,(x) = [\sigma\,(x)]^2\,. \tag{5.17}$$

5.3 Linear correlation coefficient

The linear correlation coefficient is a number that helps determine the strength of a linear relationship between two interval- or ratio-scale quantitative variables.

Definition 5.9 The linear correlation coefficient of a statistical series of variables x and y is the number r defined by:

$$r = \frac{\sigma_{xy}}{\sigma\,(x) \times \sigma\,(y)}. \tag{5.18}$$

Remark 5.3

- The correlation coefficient r is a dimensionless value that always falls between -1 and $+1$.
- A positive linear correlation coefficient indicates a positive linear relationship, whereas if r is negative, the linear relationship between the two variables is negative.
- The closer the value of r is to -1 or $+1$, the stronger the linear relationship between the two variables.

5.4 Regression Line and Mayer's Line

5.4.1 Regression Line

A regression line is the line that best fits a scatterplot showing a linear correlation. The regression line is used for making predictions. We talk about a linear correlation when the points in a scatterplot tend to align. The stronger the tendency, the stronger the linear correlation.

Definition 5.10 In a plane equipped with an orthonormal coordinate system, consider a set of n points with coordinates $(x_i; y_i)$. The line D with the equation $y = ax + b$ is called the regression line of y on x for the statistical series if the following quantity is minimized:

$$S = \sum_{i=1}^{n} (M_i Q_i)^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 . \tag{5.19}$$

Remark 5.4 To define the coefficients a and b, we expand S and consider it successively as a trinomial in b and then, with b determined, as a trinomial in a. We obtain:

$$b = \overline{y} - a\overline{x} \tag{5.20}$$

and

$$a = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x}^2)} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}. \tag{5.21}$$

The form of the coefficient b allows us to observe that the fitting line passes through the point $(\overline{x},\ \overline{y})$ satisfying $\overline{y} = a\overline{x} + b$.

Remark 5.5 The regression line D for y on x has the equation $y = ax + b$ where:

$$\begin{cases} a = \frac{\sigma_{xy}}{[\sigma_x]^2} \\ b \text{ satisfying } \overline{y} = a\overline{x} + b. \end{cases} \tag{5.22}$$

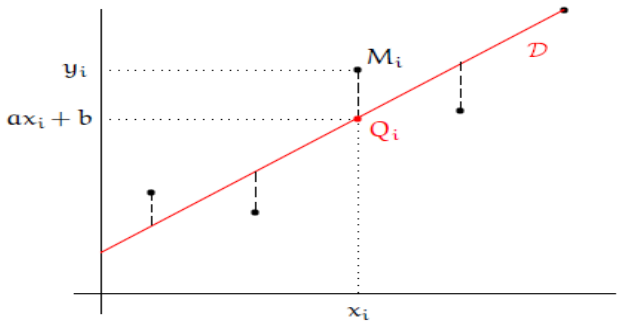

Figure 11

Remark 5.6 It would be just as wise to consider the line D' that minimizes the quantity $\sum_{i=1}^{n} [x_i - (ay_i + b)]^2$. This line is called **the regression line of** x **on** y.

Example 5.9 Here is the number of kilometers covered by cyclists as a function of the number of hours spent on the track during an endurance race.

The first column represents the number of hours, and the second column represents the distance covered in kilometers.

Hours	Distance
1	9
2	20
2	31
3	36
4	56
5	57
5	71
6	45
7	60
7	71
8	75
9	98
10	97
10	104

Table 27

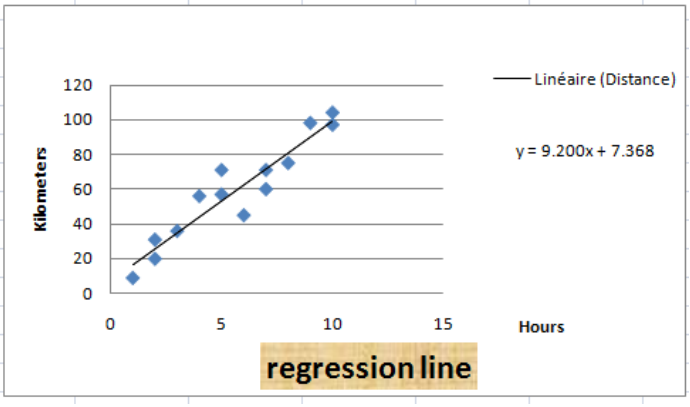

Figure 12

The equation of the regression line is: $y = ax + b$

$$a = \frac{\sigma_{xy}}{[\sigma_x]^2},\ b = \overline{y} - a\overline{x}. \tag{5.23}$$

We have$\overline{x} = 5.64$, $\overline{y} = 59.28$, then, $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy} = 77.03$ and $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = 8.37$, we obtain, $a = 9.20$.

On the other hand, $b = 59.28 - 9.20 * 5.64 = 7.39$

Finally, we have the regression line: $y = 9.20x + 7.39$.

Definition 5.11 Consider a statistical series with two variables, X and Y, whose values are pairs $(x_i; y_i)$. We call the series' mean point the point G with coordinates:

$$\begin{cases} x_G = \frac{x_1+x_2+\ldots+x_n}{n}. \\ y_G = \frac{y_1+y_2+\ldots+y_n}{n}. \end{cases} \tag{5.24}$$

5.4.2 Mayer's Line (or Method of Mean Points)

This adjustment involves determining the line passing through two mean points of the scatterplot.

Example 5.10 Find the regression line representing the results of a group of students in Mathematics and Probability using the Mayer's line method.

Students	Math(%)	Proba(%)
1	82	60
2	65	80
3	91	85
4	75	62
5	58	75
6	42	63
7	71	60
8	64	50
9	83	87
10	60	82
11	86	71
12	95	87
13	45	40
14	55	50
15	12	90

Table 28

Math(%)	12	42	45	60	55	58	64	65
Proba(%)	90	63	40	82	50	75	50	80
Math(%)	71	75	82	83	86	91	95	
Proba(%)	60	62	70	87	71	85	87	

Table 29

Therefore,

$$\begin{aligned} x_{G_1} &= \frac{12+42+45+60+55+58+64+65}{8} = \frac{391}{8} \\ y_{G_1} &= \frac{90+63+40+82+50+75+50+80}{8} = \frac{530}{8} \end{aligned}$$

So, the coordinates of the point $G_1\left(\frac{391}{8},\ \frac{265}{4}\right)$.

$$\begin{aligned} x_{G_1} &= \frac{65+71+75+82+83+86+91+95}{8} = 81 \\ y_{G_1} &= \frac{80+60+62+70+87+71+85+87}{8} = \frac{301}{4} \end{aligned}$$

So, the coordinates of the point $G_2\left(81,\ \frac{301}{4}\right)$.
The regression slope is $a = \frac{\Delta y}{\Delta x} = \frac{72}{875}$ with $b = \frac{54029}{1028}$
In the end, we have:

$$y = \frac{72}{875}x + \frac{54029}{1028}.$$

Question: What would be the probability grade of a student who scored 75% in mathematics?

Answer: We have:

$$\begin{aligned} y &= \frac{72}{875}x + \frac{54029}{1028} \\ &= \frac{72}{875} \times 75 + \frac{54029}{1028} \\ &\simeq 73.6\% \end{aligned}$$

5.5 Regression curves, regression corridor, and correlation ratio

5.5.1 Regression curves

A regression curve allows for the analysis of the relationship between two variables (explanatory variable and response variable) and highlights the nature of this relationship without making any prior assumptions about its form. It associates the values of the first variable with the conditioned means of the second variable. Therefore, from two variables, X and Y, two regression curves can be constructed:

– The X on Y curve, denoted as $C_{X/Y}$: it relates the values of Y (y_i) to the conditional means of X: $\overline{x_i}$.

– The Y on X curve, denoted as $C_{Y/X}$: it relates the values of X (x_i) to the conditional means of Y: $\overline{y_i}$.

Aside from providing a visual representation of the relationship between two variables, regression curves have important properties:

– They best summarize the shape of the scatterplot because, on average, they are closest to the points in the scatterplot (the sum and average of the squared distances between the points in the scatterplot and the regression curves are minimal).

– They intersect at a point that represents the center of gravity of the scatterplot (i.e., a point with approximate coordinates being the marginal means of the two variables).

Definition 5.12 If, for each value of X, we plot the point corresponding to the average of the Y values at that value of X (the conditional means of Y given X), and then connect these points, we obtain a curve known as the regression curve of Y on X (see figure below). If we imagine that X perfectly predicts Y, then the average conditional value of Y at each X point should be the value predicted by the relationship. Consequently, the regression curve would actually be the curve that expresses the relationship between X and Y. In

practice, however, empirical measurements are always affected by noise, so there is always some amount of variation around the conditional mean (in other words, the conditional variance is never zero).

xi	yi
5.3	13
7.2	18
5.6	9.4
7.1	15.4
5	7.8
3.3	9.3
5.2	10.1
4.5	7.1
4	8.9
2	4.4
5.7	12.1
4.7	11.5

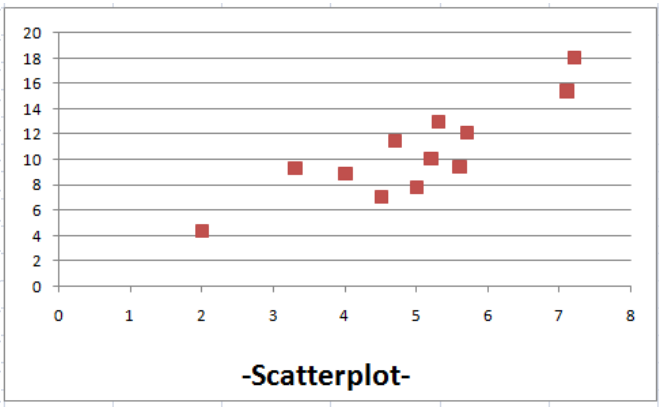

Table 30 *Figure* 13

In the example in the figure above, the regression curve does not necessarily pass through the middle of the points on a given vertical. This is because each point is weighted by the number of corresponding observations. A point present in many observations will 'pull' the regression curve more than the other points.

Conversely, one can plot the regression curve of X on Y, and this regression curve represents the best possible estimate of Y ability to predict X.

For each point in the scatterplot, its coordinate y_i (resp x_i) can be decomposed into a part corresponding to the regression curve in X (resp Y) plus a residual part. Expressed in terms of variance, this yields:

$Var(Y)$ = variance 'explained by the regression curve' in X + Residual Variance

Or alternatively:

$Var(X)$ = variance 'explained by the regression curve' in Y+ Residual Variance

5.5.2 Correlation Ratio

To study the relationship between a qualitative variable and a quantitative variable, we decompose the total variation into intergroup (or interclass) variation and intragroup (or intraclass) variation. To measure the strength of the relationship, one can calculate a parameter called the correlation ratio.

The concept of variation

The variance of a quantitative variable is, by definition, the average of the squares of the deviations from the mean. We define:

- **Total variation:** Total variation is the sum of the squares of the deviations from the mean.

$$Vartot = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{5.25}$$

-Intergroup variation: Intergroup variation is the square of the differences between the group mean and the global mean.

$$Varintr = \frac{1}{n}\sum_{k=1}^{p} n_k\,(\overline{x_k} - \overline{x})^2 \tag{5.26}$$

where $\overline{x_k}$ It designates the mean of group k, where n_k, represents the number of individuals belonging to the same group.

- Intra-groupe variation: In general, the total variation is the sum of intergroup variation and intragroup variation. The latter is the weighted sum of variances calculated within each group.

$$Varintra = \frac{1}{n}\sum_{k=1}^{p} n_k s_k^2, \tag{5.27}$$

where s_k^2 is the descriptive variance within group k. It is easily obtained by taking the difference between the total variation and the intergroup variation. We then have the following fundamental relationship in statistics:

$$Vartot = Varintr + Varintra. \tag{5.28}$$

We calculate the correlation ratio, denoted as:

$$\eta^2 = \frac{Varintr}{Vartot} = \frac{\sum_{k=1}^{p} n_k\,(\overline{x_k} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{5.29}$$

Example 5.11 We consider a population in which we observe a variable X, which represents the number of days of absence per year. This variable is observed across different professional sectors (primary sector (P), secondary sector (S), and tertiary sector (T)).

This essentially involves introducing a second variable Y, whose categories define three groups (sectors).

X	7	6	7	12	6	9	12	9	6	8	14	11	10	9
Y	P	S	S	T	P	P	S	S	T	T	T	S	P	S

Table 31

The variable X is quantitative; however, the variable Y is qualitative.

We have:

$$\overline{X} = 9, \qquad Var\,(X) = Vartot = \frac{84}{14} = 6$$

The variable Y, whose categories define 3 groups, allows us to:

-Take an intragroup approach by studying the number of days of absence by sector. This will involve means and intragroup variances.

-Take an intergroup approach by studying the variability between groups (represented by their means). This will involve intergroup variance.

For each sector, we consider the distribution of the number of days of absence, and thus, we can calculate 3 means and 3 variances, referred to as conditional means and conditional variances.

	absence	n_k	mean m_k	variance V_k	$n_k \times V_k$
P	6, 7, 9,10	4	8	10/4	10
S	6, 7, 9, 9, 11, 11	6	9	26/4	26
T	6, 8, 12, 14	4	10	40/4	40

Table 32

Therrefore,

$$Varintra = \frac{76}{14}$$

The Intra-Variance (varintra) is equal to the average of conditional variances. It measures residual variability.

	absence	n_k	mean m_k	$m_k - 9$	$n_k \times (m_k - 9)^2$
P	6, 7, 9,10	4	8	-1	4
S	6, 7, 9, 9, 11, 11	6	9	0	0
T	6, 8, 12, 14	4	10	1	4

Table 33

Then,

$$Varinter = \frac{8}{14}$$

The inter-variance (var inter) is equal to the variance of conditional means. It measures variability between groups.

Therefore,

$$\eta^2 = \frac{8}{84} = 0.095$$

Remark 5.7 - η^2 always between 0 and 1.

- $\eta^2 = 0$ if and only if varinter $= 0$, which implies that the two variables are independent.
- $\eta^2 = 1$ if and only if varintra $= 0$, indicating a perfect relationship.
- $0 < \eta^2 < 1$ represents the percentage of one variable's explanation by the other.

5.6 Functional Adjustment

When we want to model data, the next question that arises is to find the equation for the regression curves. Regression curves typically correspond to complex functions, but we can try to approximate them with simpler functions. The general principle is to start with a known functional form and seek the parameters that best fit the obtained curves to the regression curves.

For example, if we begin with the idea that the regression curve, disregarding measurement errors, would be a straight line, then it would be characterized by an equation of the type:

$$y = ax + b. \tag{5.30}$$

We would then need to identify the parameters a and b to determine the equation of the straight line.

If we imagine that the regression curve corresponds to a parabolic, the sought-after equation would be of the form:

$$y = ax^2 + bx + c. \tag{5.31}$$

And in this case, we would need to find the values of the three parameters a, b, and c.

Example 5.12

xi	yi
5.3	13
7.2	18
5.6	9.4
7.1	15.4
5	7.8
3.3	9.3
5.2	10.1
4.5	7.1
4	8.9
2	4.4
5.7	12.1
4.7	11.5

Table 34

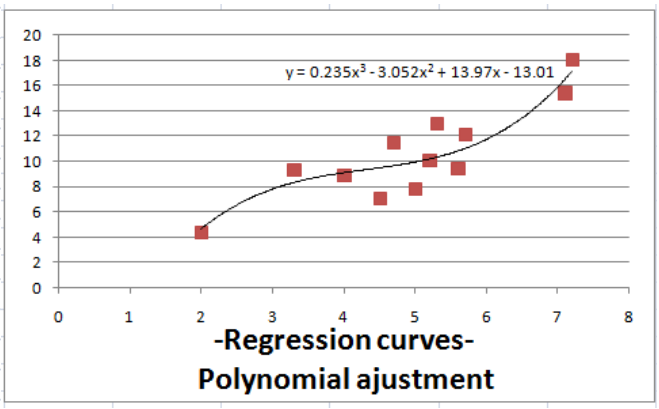

Figure 14

5.6.1 Power adjustment

Power adjustment is based on the curve represented by the equation of the type

$$y = ax^b. \tag{5.32}$$

We notice that

$$\ln y = a \ln x + \ln b,$$

we define $V = \ln y$ and $U = \ln x$, nd determine the equation of the regression line of v on u using **the Mayer method or the least squares method**. The obtained equation is of the form $v = Au + B$, from which we deduce the equation of the power function curve:

$$y = ax^b, \text{ since, } A = b \text{ and } B = \ln a.$$

Example 5.13 We use the example data 5.9 and we seek the regression curve of the following power equation:

$$y = ax^b,$$

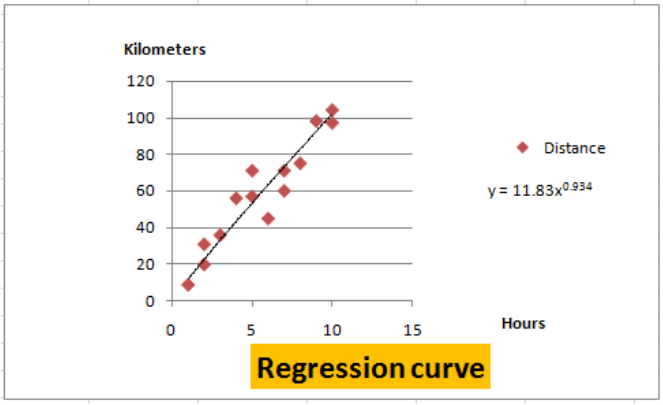

Figure 15

5.6.2 Exponential adjustment

Exponential adjustment has an exponential function curve with the equation:

$$y = ab^x. \tag{5.33}$$

We notice that

$$\ln y = x \ln b + \ln a,$$

We set $z = \ln y$,, determine the equation of the regression line of z on x using **the Mayer method or the least squares method**. The obtained equation is of the form $z = Au + B$, from which we deduce the equation of the exponential function curve.

$$z = Ax + B,$$

This leads to the equation of the exponential function curve:

$$y = ab^x, \text{ since, } A = \ln b \text{ and } B = \ln a.$$

Example 5.14 We use example 5.9, we are looking for the regression curve of the following exponential equation.

$$y = ae^{bx},$$

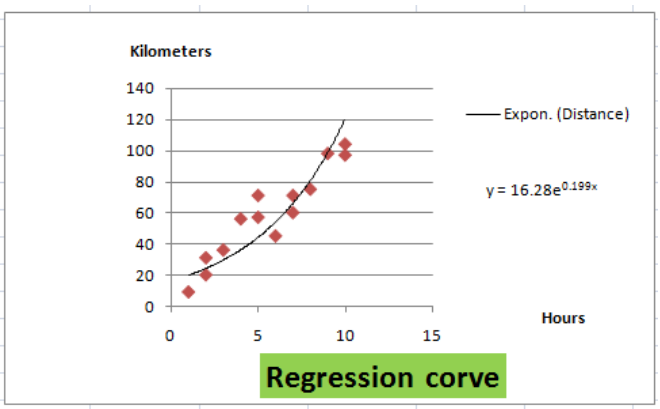
Kilometers
140
120
100
80
60
40
20
0
0
5
10
15
Expon. (Distance)
y = 16.28e^0.199x
Hours
Regression corve

Figure 16

BIBLIOGRAPHY

[1] A. KRIEF, S. LEVY. Calcul des probabilités, exercices, 1972.

[2] K. Khaled. Méthodes statistiques, Rappels de cours, exercices corrigés, Filières des sciences exactes,2 édition, 2010.

[3] F. MAZEROLLE. Statistique discriptive, séries statistiques à une et deux variables, séries chronologiques, indices, 2006.

[4] J-L ROCUET. probabilités et statistiques, édition des sciences et techniques appliquées, $4^{ème}$ édition, 1991.

[5] J. FOURASTIE, J.-F. LASLIER. Probabilités et statistique ⊙ Dunod, Paris, 1997.

[6] Y. Dodge. Premiers pas en statistique. Springer-Verlag, 2003.

[7] D. Dacunha-Castelle and M. Duflo. Probabilités et statistiques : Problèmes à temps fixe. Masson, 1982.

[8] J.-F. Delmas. Introduction au calcul des probabilités et à la statistique. Polycopié ENSTA, 2008.

[9] W. Feller. an Introduction to Probability Theory and its Applications, Volume 1. Wiley & Sons, Inc.,3rd edition, 1968.

[10] G. Grimmett, D. Stirzaker, Probability and Random Processes, Oxford University Press, 2nd edition, 1992.

[11] J. Jacod and P. Protter,Probability Essentials, Springer, 2000.

[12] A. Montfort. Cours de statistique mathématique. Economica, 1988.

[13] A. Montfort. Introduction à la statistique. Ecole Polytechnique, 1991.

[14] MURRAY R. SPIEGEL, Théorie et Applications de la Statistique, Serie Schaum, 875 exersices résolus, Paris 1991.

[15] Jérôme Depauw, Statistiques cours et exercices corrigés, Master Mathématiques, 2012.

[16] Christophe chesneau, Probabilités discrètes, recueil d'exercices corrigés, licence 1-2, 2011.

[17] P. Del Moral, B. Rémillard, S. Rubenthaler, une introduction aux probabilités, Laboratoire J. A. Dieudonné-HEC Montréal, Université de Nice-Sophia Antipolis, 2016.

[18] B. Jean-Pierre, Statistique mathématique, Applications commentées, 2010.

[19] J. Benjamin, Probabilités et statistique, 2009.

[20] D. Fredon, M. Maumy-Bertrand, F. Bertrand, Mathématiques Statistique et probabilités en 30 fichiers, 2009.

[21] F. Dominique, A. Fuchs, Calcul de probabilités, cours et exercices corrigés, deuxième édition, 1998.

[22] G. Christol, A. Decomps-Guilloux, C. Piquet, Analyse et probabilités avec rappels de cours, Capes de mathématiques. écrits 1996-1999.

[23] A. Rebbouh, Statistique descriptive calcul de probabilités et variables aléatoires avec rappels de cours et problèmes corrigés, 2009.