MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
ABDELHAMID IBN BADIS UNIVERSITY OF MOSTAGANEM
FACULTY OF EXACT SCIENCES AND COMPUTER SCIENCE
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, ACSY TEAM-LABORATORY OF PURE AND APPLIED MATHEMATICS, FACULTY SEI-BP 227/118 MOSTAGANEM 27000, ALGERIA

UNIVERSITE
Abdelhamid Ibn Badis
MOSTAGANEM

Polycopy Presented by:

Kamel BENYETTOU

To Obtain Academic Habilitation

Descriptive Statistics and Introduction to Probability

Reminder of course and corrected exercises

The first year of Preparatory Cycle Department (CP1)

UMAB–MOSTAGANEM FSEI–MOSTAGANEM

Academic Year: 2025/2026 UMAB–MOSTAGANEM

Contents

Notations

1. N: is the total number for the complete set of all individuals, events under study or items.
2. $\bar{Z}$: denote the average (a mean) of the decision variable Z.
3. $\tilde{M}$: represent the median value for the variable being studied M.
4. F_i: represent the frequency value of the variable for the variable being studied X
5. $Var(T)$: denote the variance value of the variable T.
6. $\sigma = SD(Y)$: represents the value of the standard deviation of the variable being studied Y.
7. $CV(X)$: corresponds to the numerical value of the coefficient of variation associated with the variable under investigation. X.
8. $sk(Z)$: represents the value of the skewness coefficient of the variable being studied Z.
9. X, Y: Random variables
10. $f(y)$: Probability density function (PDF) which depend on variable y.
11. $F(y)$: The cumulative distribution function (CDF) which depend on variable y.

Chapter 1

Introduction

This work is presented as part of the requirements to obtain the University Habilitation, focusing on the subject **Descriptive Statistics and Introduction to Probability.** It is designed for first-year students in the Preparatory Cycle Department (CP1) at the National Higher School of Mathematics. This course, conducted during the first semester of the academic years 2023–2024 and 2024–2025, introduces foundational concepts in statistics and probability to build a strong mathematical base for students.

The content is structured into three chapters:

1. **Fundamentals of Statistics and Numerical Representation of Data**: This chapter lays the groundwork by introducing the basic principles of statistics and methods for organizing and representing data numerically and graphically.

2. **Two-Way Statistics**: Here, the focus shifts to analyzing relationships between two variables using statistical tools such as contingency tables and correlation measures.

3. **Introduction to Probability**: This chapter provides an entry point into probability theory, covering essential concepts like probability measures, events, and basic properties.

4. Midterm and Final Exam and Solutions

   The midterm exam and final exam evaluates the knowledge and skills acquired in the course. Below are sample questions along with their solutions:

To reinforce the theoretical concepts, each chapter is complemented by a series of exercises with detailed solutions, enabling students to apply and deepen their understanding. Additionally, the course includes a midterm exam and a final exam to assess students' progress and mastery of the material.

I would like to express my sincere gratitude to Dr. Kessira Abdelrahim for his valuable assistance and support in the preparation of the course material included in this course. His guidance and contributions have been greatly appreciated and have significantly enhanced the quality of the content.

Chapter 2

Fundamentals of Statistics and Numerical representation of data

2.1 Definitions

1. Statistics: statistics is a set of methods used to organize tests providing observations, to analyze these and to interpret the results. Statistical analysis is divided into two parts.

2. Descriptive statistics: aims to describe, i.e. to take up or represent the data. Typical questions on descriptive statistic:

   - Graphical representation
   - Parameters of position, dispersion, relationship.

3. Inferential statistics: all the methods used to formulate a judgment. It requires more advanced mathematical tools (probability theory). Typical questions

   - Estimation of parameters
   - Confidence interval
   - Hypothesis tests
   - Modeling (example linear regression)

2.2 Basic notions:

1. *POPULATION: The collection of objects or people studied (students, residents, cars, etc.).

2. INDIVIDUAL: element of the population studied. (a student, a resident, a car, etc.).

3. SAMPLE: part of the population studied. Number of individuals in a sample noted n is called the sample size

4. VARIABLE (CHARACTER): property common to the individuals of the population, that we want to study. A character can be:

   - qualitative: we cannot associate either a numerical value or a natural order (type of car, hair color, etc.).
   - quantitative can be:
     - Continuous: can take all the numerical values of a determined interval (size, weight, etc.).
     - Discontinuous (discrete): can only take isolated numerical values (number of rooms in the house, number of damaged fruits, etc.).

5. MODALITY: one of the particular forms of a character. Eye color is a character, its modalities are: blue, green, brown, ...

6. EFFECTIVE OR ABSOLUTE FREQUENCY: (noted n_i) number of appearances of the value associated with a character in a sample.

7. RELATIVE FRENQUENCY(denoted by f_i): $f_i = \frac{n_i}{N}$

2.3 Central Tendency Statistics (Measures)

2.3.1 The Mean

Ungrouped informations (data)

$$\bar{Z} = \frac{\sum_{i=1}^{m} Z_i}{m}.$$

Grouped informations (data)

$$\bar{Z} = \frac{\sum_{j}^{m} z_j m_j}{m} = \sum_{j}^{m} z_j f_i,$$

where z_i represents the midpoint of the class interval under study. Z , m_j it is the frequency value of the required category. j, $f_j = \dfrac{m_i}{m}$ is the values of relative frequency, and $m = \sum_j m_j$.

2.3.2 Median

At most half of the data values are less than or equal to the median, while at least half are greater than or equal to the median.

Ungrouped informations (data)

Let's assume that the informations is arranged in ascending (increasing) order: $Z_{(1)} \leq \cdots \leq Z_{(m)}$.

- If m is even number:

$$\text{Median} = \tilde{Z} = \frac{Z_{(m/2)} + Z_{(m/2+1)}}{2}.$$

- If m is odd number:

$$\text{Median} = \tilde{Z} = Z_{((m+1)/2)}.$$

Grouped information's (data)

Locate the class $[a_M, b_M[$ that contains the median, then interpolate using:

$$\text{Median} = a_M + \left(\frac{\frac{N}{2} - fr^*_{eq}}{n_M} \right) (b_M - a_M),$$

where N is the total number represents the observed information (observation), n_M is the frequency of the median class, and fr^*_{eq} represents the cumulative frequency for all groups before the median class.

2.3.3 Quartiles

The first quartile Q_1 satisfies the property that at least 25% of the observations are less than or equal to Q_1, while at least 75% are greater than or equal to Q_1.

Grouped data

First quartile. Locate the class $[a_{Q_1}, b_{Q_1}[$ containing the first quartile Q_1, then interpolate using:

$$Q_1 = a_{Q_1} + \left(\frac{\frac{N}{4} - f^*}{n_{Q_1}} \right) (b_{Q_1} - a_{Q_1}),$$

where n_{Q_1} corresponds to the frequency associated with the class interval that includes the first quartile and f^* is the value of cumulative frequency of the classes preceding that class.

Third quartile. Locate the class $[a_{Q_3}, b_{Q_3}[$ containing the third quartile Q_3, then interpolate using:

$$Q_3 = a_{Q_3} + \left(\frac{\frac{3N}{4} - f^*}{n_{Q_3}} \right) (b_{Q_3} - a_{Q_3}),$$

where n_{Q_3} corresponds to the frequency associated with the class interval that includes the first quartile and f^* is the value of cumulative frequency for the classes preceding that class.

2.3.4 Mode and Modal Class

Ungrouped information's (data)

The mode is the number that corresponds to the highest frequency (or number of times).

Grouped information's (data)

- When the sizes of the categories are identical, the most common category is the one with the highest frequency (or number)..
- Conversely, the modal category is the one that appears in the graph after adjusting the heights so that the areas of the rectangular shapes are proportional to the frequencies (or numbers).

2.4 Measuring (evaluating) dispersion

2.4.1 Range

$$\text{Range} = \text{maximum} - \text{minimum}.$$

2.4.2 Variance

Ungrouped (information's) data

Main definition

$$s^2(Y) = \frac{\sum_{j=1}^{m}(y_i - \bar{Y})^2}{m-1}.$$

Computational (Practical) formula

$$s^2(Y) = \frac{\sum_{j=1}^{m} Y_j^2 - \frac{\left(\sum_{j=1}^{m} Y_i\right)^2}{m}}{m-1}.$$

Grouped (information's) data

Main definition

$$s^2(Y) = \frac{\sum_{j=1}^{m}(y_j - \bar{Y})^2 m_j}{m-1}.$$

Computational (Practical) formula

$$s^2(Y) = \frac{\sum_{j=1}^{m} y_j^2 m_j - \frac{\left(\sum_{j=1}^{m} y_j m_j\right)^2}{m}}{m-1}.$$

2.4.3 Standard deviation

$$s = \sqrt{s^2}.$$

2.5 Other coefficients

2.5.1 Coefficient of variation

$$CV = \frac{|s|}{|\bar{X}|} \times 100\%.$$

Homogeneity criteria (for this course):

- An **industrial process** is considered homogeneous if its coefficient of variation is less than 10%.
- Any **other phenomenon** is considered homogeneous if its coefficient of variation is less than 30%.

2.5.2 Skewness coefficient

Ungrouped (information's) data

$$S_{Kc} = \frac{m\sum_{i=1}^{m}(y_i - \bar{Y})^3}{(m-1)(m-2)s^3}.$$

Grouped (information's) data

$$S_{Kc} = \frac{m\sum_{j=1}^{m}(y_i - \bar{Y})^3 m_i}{(m-1)(m-2)s^3}.$$

Interpretation (for this course):

- If $S_K < 0$: negative skewness (left-skewed distribution),
- If $S_K > 0$: positive skewness (right-skewed distribution),
- If $|S_K| \leq 0.5$: skewness is negligible,
- If $0.5 \leq |S_K| \leq 2$: skewness is moderate,
- If $|S_K| > 2$: skewness is pronounced.

2.5.3 Kurtosis coefficient

Ungrouped (information's) data

$$Ku = \frac{m(m+1)\sum_{j=1}^{m}(Y_i - \bar{Y})^4}{(m-1)(m-2)(m-3)s^4} - \frac{3(m-1)^2}{(m-2)(m-3)}.$$

Grouped data

$$Ku = \frac{m(m+1)\sum_{i}^{m}(x_i - \bar{X})^4 m_i}{(m-1)(m-2)(m-3)s^4} - \frac{3(m-1)^2}{(m-2)(m-3)}.$$

2.6 Exercises

Exercice 2.6.1. *From these proposition (assertions), make a choice (decision) and specify which ones are false and which ones are true.*

1. *We call variable, a characteristic that we study.*
2. *The task of descriptive statistics is to collect data.*
3. *The task of descriptive statistics is to present data in the form of tables, graphs and statistical indicators.*
4. *In Statistics, variables are classified into different types.*
5. *The values of variables are also called modalities.*
6. *For a qualitative variable, each statistical individual can have only one modality.*
7. *To perform statistical treatments, a quantitative variable is sometimes transformed into a qualitative variable.*
8. *In practice, when a quantitative (discrete) variable has a large variety of distinct measurements (values), it is treated as continuous.*

Exercice 2.6.2. *Specify for each of these variables whether it is ordinal or nominal qualitative data, discrete or continuous quantitative variable.*
Area of an apartment - Number of subjects acquired in a semester - Eye color - Grades in higher education - Cholesterol level - Speed - Height - Type of car - Number of defective parts made by a machine - Number of languages spoken - Annual income - Citizenship - Distance Place of residence - Age.

Exercice 2.6.3. *In this upcoming subject areas of investigation specify: the population being studied, the statistical variable, the statistical unit, as well as its nature.*

1. *Study of the life span of electric lamps of a brand ALPHA.*

2. *Analysis of the absence ratio of the staff members, in days, for a factory.*

3. *Distribution of students according to the mention obtained at the Baccalaureate.*

4. *The number of collisions which involved two cars on a set of 100 intersections of roads picked at arbitrarily in a city.*

Exercice 2.6.4. *The distribution of 1600 students in a high school according to the foreign language studied is given in the following table :*

Languages	*German*	*English*	*Spanish*	*Italian*	*Various*
Size	*351*	*934*	*205*	*69*	*41*

1. *Give the population and the character studied by specifying its nature.*

2. *Represent graphically the distribution of the 1600 students according to the language studied by the appropriate diagrams.*

Exercice 2.6.5. *Give an example of a real situation which*
(i) the mean represents an adequate measure (value) of central tendency.
(ii) the mean cannot be an adequate measure (value) of central tendency however the medians is a good indicator of central tendency.
2. The rate or level of inflation in fire successive year in a country was $5\%, 8\%, 12\%, 25\%$ *and* 34%. *What was the average rate of inflation per year?*

Exercice 2.6.6. *The temperature (in* °C *) recorded each day in a given city was measured during the entire month of January and the following results were obtained:*
$12,5-14,2-15,6-14,5-13,3-11,9-10,7-10,4-9,9-9,4-10,8-14,9-15,6-15,8-14,3$ $-13,6-12,9-13,6-16,2-16,8-17,6-17,0-15,3-14,2-12,6-11,8-11,4-10,2-10,0-$ *9,1-8,8.*

1. *Specify the population in this case and also provide the character studied by indicating its nature.*

2. *Calculate the range of this series. Give the statistical table of this population (distribution).*

3. *Update the statistical table by including the frequencies, the percentages and the growing and declining (increasing and decreasing) cumulative values.*

4. *Draw a histogram for this distribution as well as the curves of the increasing and decreasing cumulative numbers.*

5. *Determine the modal class or classes, the mean and the median.*

Exercice 2.6.7. *For all the soccer matches counting for the English Cup, we will study the number of goals scored by the winning team including the draws.*

No. of goals scored	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*Total*
No. of matches	*80*	*145*	*124*	*75*	*40*	*10*	*6*	*480*

1. *What is the character studied and specify its nature.*

2. *Complete the statistical table with percentages.*

3. *Describe graphically the more than Ogive Graph and the less than Ogive Graph and then locate the position of median.*

4. *Calculate the median value and mean.*

Exercice 2.6.8. *A household survey of summer vacation budgets yielded the following results:*

Budget X(10 ^{2} DA)	*Cumulative frequency*
$[800; 1000[$	*0.08*
$[1000; 1400[$	*0.18*
$[1400; 1600[$	*0.34*
$[1600; \beta[$	*0.64*
$[\beta; 2400[$	*0.73*
$[2400; \alpha[$	*1*

1. *Calculate the real value of α knowing that the range of the given series is equal to 3200×10^2DA.*

2. *Complete the given table with the value of frequencies.*

3. *Find the value of β in the two following cases:*
   a. *The average budget is equal to 1995×10^2DA.*
   b. *The median budget is equal to 1920×10^2DA.*

Consider throughout the following β calculated from the average (mean) budget.
Display the representation of the histogram.
5. *Determine the mode and median measures.*

Exercice 2.6.9. *The following values of frequency distribution of storm duration (in minutes) for 74 storms appeared in the article "Lightning Phenomenology in the Tampa Bay Area" (J. of Geophysical Research (1984): 789-805).*

Storm duration (min)	n_i	*Storm duration (min)*	n_i
$0 \leq x < 25$	*1*	$150 \leq x < 175$	*5*
$25 \leq x < 50$	*17*	$175 \leq x < 200$	*4*
$50 \leq x < 75$	*14*	$200 \leq x < 225$	*3*
$75 \leq x < 100$	*11*	$225 \leq x < 250$	*2*
$100 \leq x < 125$	*8*	$250 \leq x < 275$	*0*
$125 \leq x < 150$	*8*	$275 \leq x < 300$	*1*

1. *Sketch a histogram of the data*

2. *As a measure (value) of central tendency for the previous distribution, can you locate the interval in the given table which contains the value of median?*

3. *From the table and the provided information's, can you calculate the mean?*

4. *Calculate the three quartiles and the variance of this data .*

5. *Suppose that the storm duration is tripled and n_i for each class is doubled (it mean that we ha have* 148 *storms), what would be the average and the absolute deviation ? (without calculus)*

Exercice 2.6.10. *Given the following grades on a test:*

93	68	86	100	75	89	62	98	81	95
52	100	71	96	79	92	86	88	93	86

- *Calculate the percentage of evaluations lie between one and two standard deviation from the average (mean)?*

- *Are these calculations and results adequate and consistent with Chebyshev's Theorem ?*

Exercice 2.6.11. *Random samples, every of length* n = 10, *were obtained of the dimensions in centimeters for three different types of business fish, using the below results:*

Sample 1	*108*	*100*	*99*	*125*	*87*	*105*	*107*	*105*	*119*	*118*
Sample 2	*133*	*140*	*152*	*142*	*137*	*145*	*160*	*138*	*139*	*138*
Sample 3	*82*	*60*	*83*	*82*	*82*	*74*	*79*	*82*	*80*	*80*

1. *Create the boxplot to each of the samples. Are that their patterns of distribution are symmetric. Calculate the pearson's coefficient of skewness to validate and confirm your answer.*

2. *Which type of fish has the smallest variance in length between them?*

2.7 Solution Of Exercices

Solution 2.7.1.

1. *True - A variable is a characteristic or attribute that we study in statistics.*

2. *False - The task of descriptive statistics is not to collect data; it is to organize, summarize, and present data that has already been collected.*

3. *True - Descriptive statistics involve presenting data in the form of tables, graphs, and statistical indicators to provide a clear understanding of the data's characteristics.*

4. *True - In statistics, variables are classified into different types, such as qualitative (categorical) and quantitative (numerical) variables.*

5. *True - The values of variables can also be called modalities, especially in the context of qualitative variables.*

6. *True - For a qualitative variable (categorical), each statistical individual can have only one modality (category).*

7. *True - In some cases, a quantitative variable can be transformed into a qualitative variable for ease of analysis or interpretation. This process is called discretization.*

8. *True - In practice, when a discrete quantitative variable has a large number of distinct values, it is often treated as continuous for simplicity and ease of analysis. This is known as approximating a continuous distribution.*

Solution 2.7.2.

1. *Let's categorize each of the variables as either ordinal or nominal qualitative data, and as either discrete or continuous quantitative variables:*

2. *Area of an apartment: Continuous quantitative variable. It can take on any real value within a range.*

3. *Number of subjects acquired in a semester: Discrete quantitative variable. You can count the number of subjects, and it takes on whole-number values.*

4. *Eye color: Nominal qualitative data. Eye color is a categorical variable with no inherent order.*

5. *Grades in higher education: Ordinal qualitative data. Grades have a natural order (e.g., $A > B > C$), but the intervals between grades are not necessarily equal.*

6. *Cholesterol level: Continuous quantitative variable. Cholesterol levels can be measured with a high degree of precision and can take on a wide range of real values.*

7. *Speed: Continuous quantitative variable. Speed can be measured as a real number within a range.*

8. *Height: Continuous quantitative variable. Height can take on any real value within a range.*

9. *Type of car: Nominal qualitative data. Car types are categorical and do not have a natural order.*

10. *Number of defective parts made by a machine: Discrete quantitative variable. You can count the number of defective parts, and it takes on whole-number values.*

11. *Number of languages spoken: Discrete quantitative variable. You can count the number of languages spoken, and it takes on whole-number values.*

12. *Annual income: Continuous quantitative variable. Income can take on a wide range of real values and is typically measured in dollars or other currency units.*

13. *Citizenship: Nominal qualitative data. Citizenship is a categorical variable with no inherent order.*

14. *Distance Place of residence: Continuous quantitative variable. Distance can be measured as a real number within a range.*

15. *Age: Discrete quantitative variable. Age is often measured in whole years, so it's considered discrete.*

Solution 2.7.3. ***1.*** *Study of the life span of electric lamps of a brand ALPHA:*

- *Population: All electric lamps of brand ALPHA produced or in use.*
- *Statistical Unit: Each individual electric lamp of brand ALPHA.*

- *Statistical Variable: Life span (or lifespan) of electric lamps.*
- *Nature of Variable: Continuous quantitative variable (measured in hours, days, etc.).*

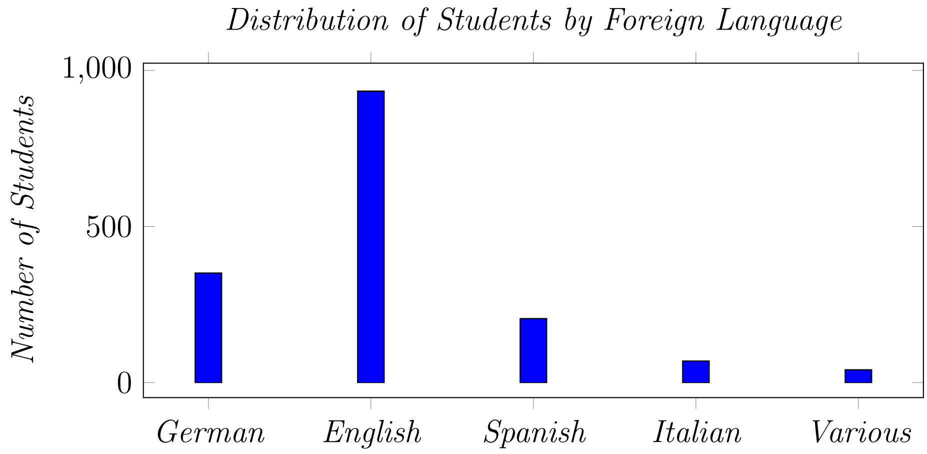
2. *Study of the absenteeism of the workers, in days, of a factory:*
   - *Population: All workers in the factory.*
   - *Statistical Unit: Each individual worker.*
   - *Statistical Variable: Absenteeism (number of days absent) of each worker.*
   - *Nature of Variable: Discrete quantitative variable (whole number of days).*
3. *Distribution of students according to the mention obtained at the Baccalaureate:*
   - *Population: All students who took the Baccalaureate exam.*
   - *Statistical Unit: Each individual student.*
   - *Statistical Variable: Mention obtained at the Baccalaureate (e.g., "Pass," "Distinction," "High Distinction").*
   - *Nature of Variable: Nominal qualitative data (categorical).*
4. *Number of accidents involving two cars on a set of 100 road intersections chosen at random in a city:*
   - *Population: All road intersections in the city.*
   - *Statistical Unit: Each individual road intersection.*
   - *Statistical Variable: Number of accidents involving two cars at each road intersection.*
   - *Nature of Variable: Discrete quantitative variable (whole number of accidents).*

Solution 2.7.4. *Let's address each part of the exercise:*

Population and Character Studied:

Population: The population in this case is the 1600 students in the high school. Character Studied: The character studied is the choice of foreign language studied by each student. Nature of Character: The character, in this case, is a categorical variable representing the different languages, so it's nominal qualitative data.

The bar chart showing the distribution of students by foreign language.



Here's how you can calculate the angles:

Calculate the Total: Add up the values of the categories to find the total. In your case, the total number of students is 1600.

Calculate the Angle for Each Category: To calculate the angle for a specific category, divide the value of that category by the total and multiply it by 360 degrees (a full circle). The formula is:

$$Angle\ for\ a\ Category = \frac{Value\ of\ category}{Total} \times 360^0 \tag{2.1}$$

- *German:*
  - *Angle for German:*

$$Angle\ for\ German = \left(\frac{351}{1600}\right) \times 360°$$

$$Angle\ for\ German \approx 78.975°$$

- *English:*
  - *Angle for the English language:*

$$Angle\ for\ English = \left(\frac{934}{1600}\right) \times 360°$$

$$Angle\ for\ English \approx 210.15°$$

- *Spanish:*

- *Angle for the Spanish language:*

$$\text{Angle for Spanish} = \left(\frac{205}{1600}\right) \times 360^\circ$$

$$\text{Angle for Spanish} \approx 46.12^\circ$$

- *Italian:*
  - *Angle for the Italian language:*

$$\text{Angle for Italian} = \left(\frac{69}{1600}\right) \times 360^\circ$$

$$\text{Angle for Italian} \approx 15.52^\circ$$

- *Various:*
  - *Angle for Various:*

$$\text{Angle for Various} = \left(\frac{41}{1600}\right) \times 360^\circ$$

$$\text{Angle for Various} \approx 9.22^\circ$$

A pie chart showing the distribution of students by foreign language.

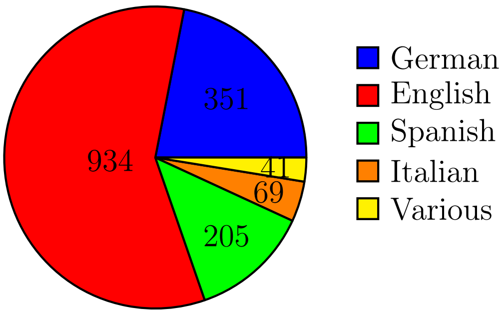


Figure 2.1: Distribution of Students by Foreign Language

Solution 2.7.5.

1. Situation for Mean and Median:

(i) *The mean is an appropriate measure of central tendency when dealing with data that has a relatively symmetric or normal distribution. For example "Exam Scores in a Classroom".*

Consider a classroom where students have taken a math exam. Each student's score on the exam represents a quantitative variable, and you want to understand the average performance of the class.

***Mean (Average):** - The mean (average) score is calculated by adding up all the scores and dividing by the number of students. It is a suitable measure of central tendency when there are no extreme outliers or skewed data.*

$$Mean = \frac{86 + 90 + 88 + 87 + 91}{5} = 88.4$$

- The mean score is 88.4, which accurately represents the central tendency of the class's performance on the math exam. It provides a good sense of how well the class performed on average.

(ii) The mean may not be appropriate when dealing with data that has outliers or is skewed, as it can be sensitive to extreme values. In such cases, the median can be a better measure of central tendency. For example:

Let's consider a situation where a group of kids is attending a summer camp, and their ages are predominantly 10 and 12 years old. However, one child joined the camp whose age is 25 years old.

In this scenario:

- *Most of the kids have ages of either 10 or 12 years old, which is typical for the group.*
- *There is one outlier, the child with an age of 25 years, who is significantly older than the rest of the group.*

Now, let's discuss the choice of central tendency measures:

1. Mean (Average)

If we calculate the mean age for this group of kids, we add up all the ages and divide by the number of kids. However, the mean can be heavily influenced by outliers. In this case, the presence of the 25-year-old child significantly skews the mean.

The mean age would be calculated as:

$$Mean = \frac{10 + 12 + 25}{3} = 15.67 \ years$$

The mean age of 15.67 years doesn't accurately represent the typical age of the kids in the camp because it's influenced by the outlier.

2. Median

The median age is the middle value when the ages are arranged in ascending order. When we arrange the ages in ascending order: 10, 12, 25. The median age is 12 years old. The median provides a more appropriate measure of central tendency because it's not affected by outliers. In this case, the median age of 12 years accurately represents the typical age of the kids in the camp, which is either 10 or 12 years old.

Solution 2.7.6.

1/the population, the character studied and its nature.
Population: the daily temperatures for that month in the specified city.
Character studied: The temperature (in °C *).*
Nature: Quantitative and continuous.
2 / the range of this series
Range $(R) = 17.6 - 8.8 = 8.8$°C
# of classes $(k) = \sqrt{n} = \sqrt{31} = 5.6 \approx 6$
Class size $(w) = R/k = 1.46 \approx 2$

Class	Frequency	Relative frequency %	Inc. Cumulative	Dec. Cumulative
$[8, 10[$	4	12.90%	4	31
$[10, 12[$	8	25.80%	12	27
$[12, 14[$	6	19.35%	18	19
$[14, 16[$	9	29.03%	27	13
$[16, 18[$	4	12.90%	31	4

Modal Classes: [14, 16[
Mean: $\frac{\sum Ci*ni}{\sum ni} = 13.06$
Median $= L + \frac{n/2 - CFb}{fm} * w$ *where*

- *L is the lower boundary of the median class.*
- *N is the total number of data points.*
- *CF is the cumulative frequency of the class before the median class.*
- *f is the frequency of the median class.*
- ***w*** *is the width of the class (bin width).*

Since $n = 31 \rightarrow$ *median class is [12, 14[*

By using the definition

Median $= 12 + \frac{\frac{31}{2} - 12}{6} * 2 = 13.17$

For the Histogram of this data,

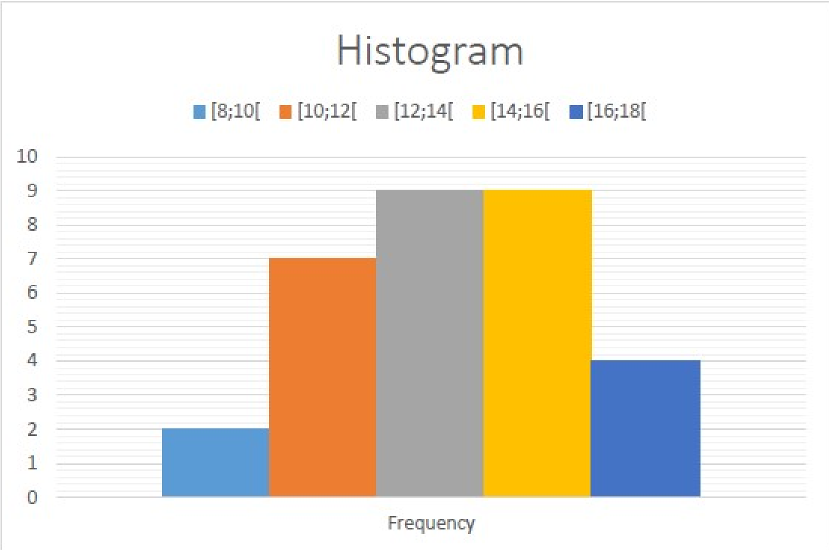


Figure 2.2

The Increasing and decreasing cumulative frequencies

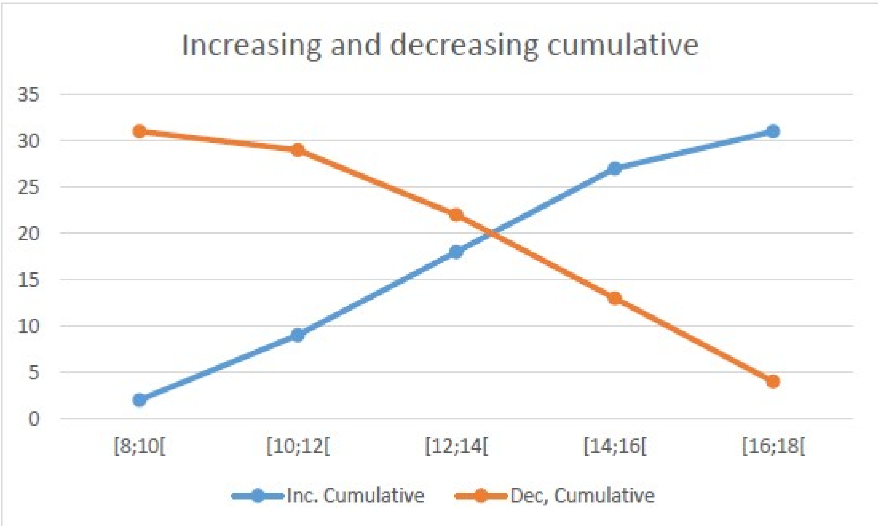


Figure 2.3

Solution 2.7.7. *1. Step 1: Character Studied and Its Nature*

The character studied is the "number of goals scored by the winning team in soccer *matches." Its nature is discrete, as it represents specific integer values (0, 1, 2, 3, 4, 5, or 6) and can be counted.*

Step 2: Complete the Statistical Table with Percentages

To complete the table with percentages, you can calculate the percentage of matches for each number of goals scored by dividing the number of matches for each category by the total number of matches (480) and then multiplying by 100.

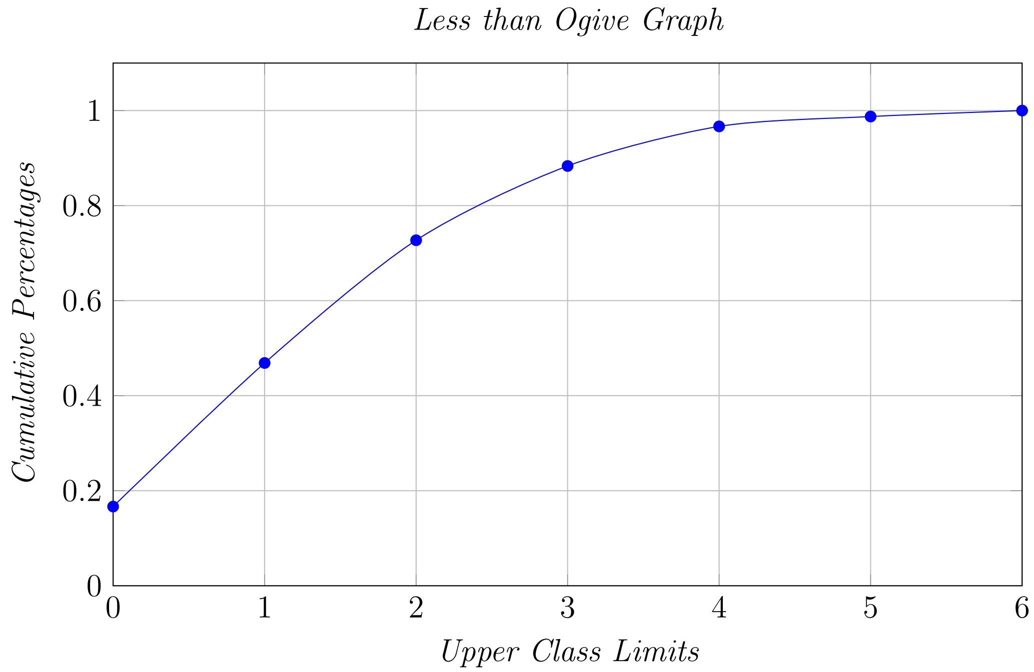
No. of goals scored	*No. of matches*	*Percentage*	*Inc. Cumulative*
0	*80*	$\frac{80}{480} \times 100\% = 16.67\%$	*0.1667*
1	*145*	$\frac{145}{480} \times 100\% = 30.21\%$	*0.4688*
2	*124*	$\frac{124}{480} \times 100\% = 25.83\%$	*0.7271*
3	*75*	$\frac{75}{480} \times 100\% = 15.63\%$	*0.8834*
4	*40*	$\frac{40}{480} \times 100\% = 8.33\%$	*0.9667*
5	*10*	$\frac{10}{480} \times 100\% = 2.08\%$	*0.9875*
6	*6*	$\frac{6}{480}1 \times 100\% = 1.25\%$	*1.0000*
Total	*480*	*100.00%*	

and for the secode case,

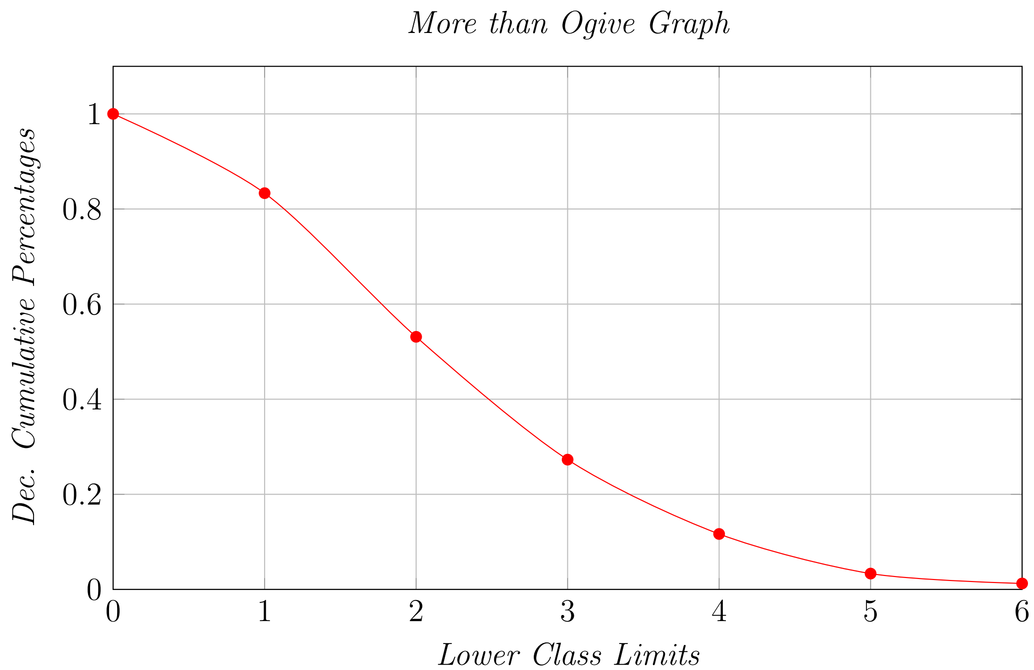
No. of goals scored	*No. of matches*	*Percentage*	*Dec. Cumulative*
0	*80*	$\frac{80}{480} \times 100\% = 16.67\%$	*1*
1	*145*	$\frac{145}{480} \times 100\% = 30.21\%$	*0.8333*
2	*124*	$\frac{124}{480} \times 100\% = 25.83\%$	*0.5312*
3	*75*	$\frac{75}{480} \times 100\% = 15.63\%$	*0.2729*
4	*40*	$\frac{40}{480} \times 100\% = 8.33\%$	*0.1166*
5	*10*	$\frac{10}{480} \times 100\% = 2.08\%$	*0.0333*
6	*6*	$\frac{6}{480}1 \times 100\% = 1.25\%$	*0.0125*
Total	*480*	*100.00%*	

Step 3: Represent Graphically the Less than Ogive and More than Ogive Graphs

Less than Ogive Graph:



More than Ogive Graph:



Locate the Median: The median is approximately 2 goals scored, where the two curves intersect.

Step 4: Calculate the Mean and Median

$$Mean = \bar{X} = \frac{(0 \times 80) + (145 \times 1) + (124 \times 2) + (75 \times 3) + (40 \times 4) + (5 \times 10) + (6 \times 6)}{480} = 1.8 \tag{2.2}$$

$$Median \approx 2 \tag{2.3}$$

Solution 2.7.8.

1. *Part 1: Determine the value of α knowing that the range of the series is equal to* $3200 \times 10^2\text{DA}$*:*

   The range of the series is given as $3200 \times 10^2\text{DA}$. *The range is the difference between the maximum and minimum values of the data. In this case, the minimum value is* $800 \times 10^2\text{DA}$ *(from the first interval), and we need to find the maximum value, which is* α.

   So, we can set up the equation for the range as follows:

   Range = Maximum - Minimum $3200 \times 10^2\text{DA} = \alpha - (800 \times 10^2\text{DA})$

   Now, solve for α*:*

   $\alpha = 3200 \times 10^2\text{DA} + (800 \times 10^2\text{DA})$ $\alpha = 4000 \times 10^2\text{DA}$

   Therefore, $\alpha = 4000 \times 10^2\text{DA}$.

2. *cumulative frequencies.*

   Let's calculate the frequencies for each interval:

   For the first interval $[800; 1000[$*: Frequency = Cumulative Frequency of* $[800; 1000[$ *- Cumulative Frequency of the previous interval Frequency* $= 0.08 - 0$ *(since there is no previous interval) Frequency* $= 0.08$

   For the second interval $[1000; 1400]$*: Frequency = Cumulative Frequency of* $1000; 1400$ *- Cumulative Frequency of the previous interval Frequency* $= 0.18 - 0.08$ *(the previous cumulative frequency) Frequency* $= 0.10$

   For the third interval $[1400; 1600]$*: Frequency = Cumulative Frequency of* $1400; 1600$ *- Cumulative Frequency of the previous interval Frequency* $= 0.34 - 0.18$ *(the previous cumulative frequency) Frequency* $= 0.16$

   For the fourth interval $[1600; \beta$*: Frequency = Cumulative Frequency of* $[1600; \beta$ *- Cumulative Frequency of the previous interval Frequency* $= 0.64 - 0.34$ *(the previous cumulative* *frequency) Frequency* $= 0.30$

For the fifth interval $[\beta; 2400]$*: Frequency* $=$ *Cumulative Frequency of* $[\beta; 2400]$ *- Cumulative Frequency of the previous interval Frequency* $= 0.73 - 0.64$ *(the previous cumulative frequency) Frequency* $= 0.09$

For the last interval $[2400; \alpha$*: Frequency* $=$ *Cumulative Frequency of* $[2400; \alpha$ *- Cumulative Frequency of the previous interval Frequency* $= 1 - 0.73$ *(the previous cumulative frequency) Frequency* $= 0.27$

Budget $X(10^2\text{DA})$	Cumulative Frequency	Frequency
$[800; 1000[$	0.08	0.08
$1000; 1400$	0.18	0.10
$1400; 1600$	0.34	0.16
$[1600; \beta]$	0.64	0.30
$\overline{[\beta; 2400]}$	0.73	0.09
$[2400; \alpha]$	1	0.27

3a) *Determine the value of* β *in the case of the average budget is equal to* 1920×10^2 *DA:*

$$\begin{aligned}
&\bar{X} = \frac{\sum x_i f_i}{f_i} \\
&1995 = \frac{0.08 * 900 + 0.1 * 1200 + 0.16 * 1500 + 0.3 * (\frac{\beta+1600}{2}) + 0.09 * (\frac{\beta+2400}{2}) + 0.27 * 3200}{0.08 + 0.10 + 0.16 + 0.30 + 0.30 + 0.09 + 0.27} \\
&1995 = 72 + 120 + 0.15(\beta + 1600) + 0.045(\beta 2400) + 864 \\
&1995 = 432 + 0.15\beta + 240 + 0.045\beta + 108 + 864 \\
&1995 = 1644 + 0.195\beta \\
&\beta = \frac{1995 - 1644}{0.195} = \frac{351}{0.195} \\
&\beta = 1800
\end{aligned} \tag{2.4}$$

3b) *The median budget is equal to* 1920×10^2*:(First we must calculate the value of* w*)*

$$
\begin{aligned}
\widetilde{X} =& L + \left(\frac{n/2 - Cf_b}{f_m}\right) w \\
1920 =& 1600 + \left(\frac{1/2 - 0.34}{0.3}\right) w \\
w =& (1920 - 1600) \times (\frac{0.3}{0.5 - 0.34}) \\
w =& 600
\end{aligned}
\tag{2.5}
$$

Where,

(a) L *=Lower class boundary of the median class*

(b) Cf_b *=Cumulative frequency before the median class*

(c) f_m*=Frequency of the median class*

(d) w*=Class size or width*

Then, we compute the value of β*,*

$$w = \beta - 1600 \tag{2.6}$$

$$\beta = w + 1600 = 600 + 1600 = 2200 \tag{2.7}$$

4) Draw the histogram:

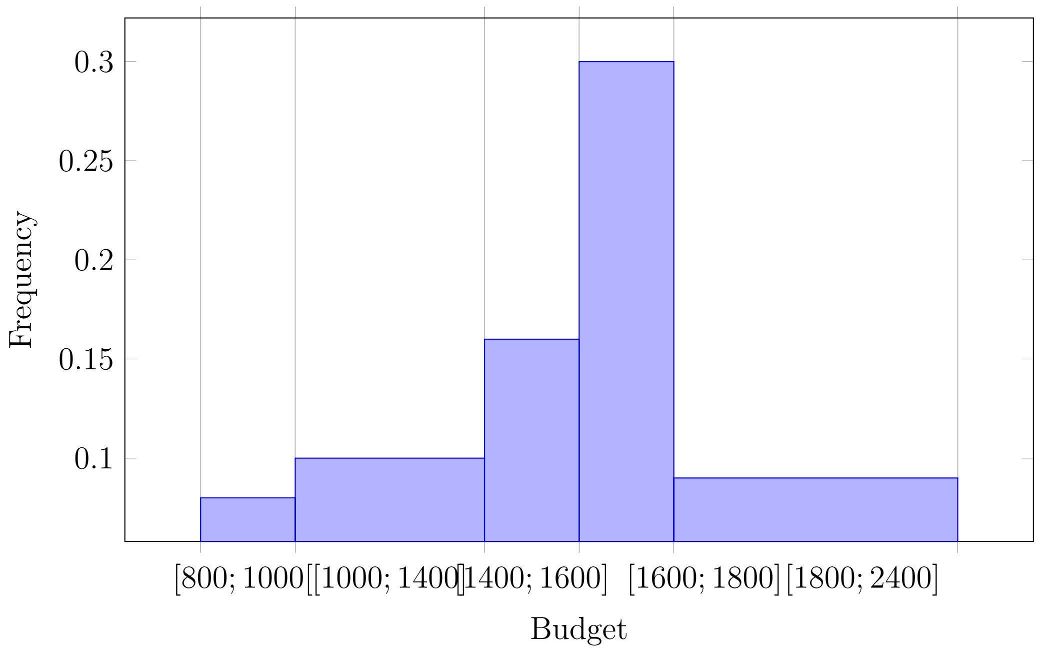


Figure 2.4: Histogram of Budget

5 *Calculate the median and the mode.*

Firstly, the median

$$\begin{aligned} Med &= L + \left(\frac{n/2 - Cf_b}{f_m}\right) w \\ &= 1600 + \left(\frac{0.5 - 0.34}{0.3}\right)(1800 - 1600) \\ &\approx 1706.66 \end{aligned} \tag{2.8}$$

Secondly, The mode

$$\begin{aligned} Mode &= L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) w \\ &= 1600 + \left(\frac{0.5 - 0.34}{0.3}\right)(1800 - 1600) \\ &\approx 1706.66 \end{aligned} \tag{2.9}$$

L_{mo} = *lower class boundary of the modal*

Δ_1 = *Difference between the frequency of the modal class and the class before it*

Δ_2 = *Difference between the frequency of the modal class and the class after it*

w = *Class size*

Solution 2.7.9. *a. Histogram*

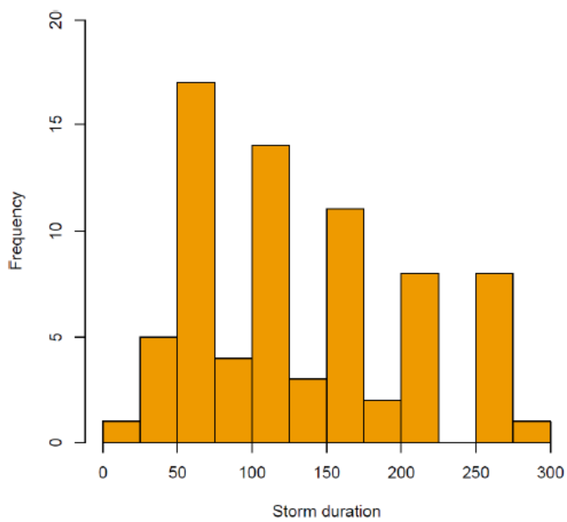


Figure 2.5

b. There are 74 observations (storm durations in minutes) and the median Me will be the average of the middle two observations in the ordered array. Thus, having 74 observations, in order to locate the median we need to find the interval that contains the 37-th and the 38-th ordered observations. From the frequency distribution, note that: - the interval $[0, 25)$ *contains one observation (the smallest); - the interval* $[0, 50)$ *contains* $1 + 17 = 18$ *observations; - the interval* $[0, 75)$ *contains* $1 + 17 + 14 = 32$ *observations; - the interval* $[0, 100)$ *contains* $1 + 17 + 14 + 11 = 44$ *observations.*

This implies that the 33 -rd, 34 -th, $\ldots, 44$*-th ordered observations will fall in the interval* $[75, 100)$. *So Me is contained in the interval* $[75, 100)$.

How to compute Me within the focussed class? Given that the median is the 50^{th} *percentile (the cumulated frequency of our interest is 0.5) and assuming the homogeneity of storm duration distribution within classes, then*

$$Me = x_{low} + w \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}$$

where x_{low} *is the lower bound of the i-th class, which contains* $Me; F_{i-1}$ *is relative cumulative frequency of class* $(i-1)$*; and w is the width of the i-th class. Therefore,*

$$\begin{aligned} Me &= 75 + 25 \frac{0.5 - 32/74}{44/74 - 32/74} \\ &= 75 + 25 \frac{0.5 - 0.4324}{0.594 - 0.4324} \\ &= 85.49 \end{aligned}$$

c. Then, the mean is approximated by

$$M = \frac{1}{n} \sum \bar{x}_i \cdot n_i = \frac{1}{74} \cdot 7450 = 100,68$$

d. The quartiles :

Storm Duration (min)	n_i	*CF*	m_i	$m_i - \bar{X}$	$(m_i - \bar{X})^2$	$n_i(m_i - \bar{X})^2$
$0 \leq x < 25$	*1*	*1*	*12.5*	*-88.18*	*7785.71*	*7785.71*
$25 \leq x < 50$	*17*	*18*	*37.5*	*-63.18*	*3988.21*	*67898.69*
$50 \leq x < 75$	*14*	*32*	*62.5*	*-38.18*	*1459.77*	*20437.89*
$75 \leq x < 100$	*11*	*43*	*87.5*	*-13.18*	*173.65*	*1900.13*
$100 \leq x < 125$	*8*	*51*	*112.5*	*11.82*	*139.65*	*1117.22*
$125 \leq x < 150$	*8*	*59*	*137.5*	*36.82*	*1356.61*	*10852.85*
$150 \leq x < 175$	*5*	*64*	*162.5*	*61.82*	*3816.17*	*19080.85*
$175 \leq x < 200$	*4*	*68*	*187.5*	*86.82*	*7534.99*	*30139.96*
$200 \leq x < 225$	*3*	*71*	*212.5*	*111.82*	*12518.85*	*37556.55*
$225 \leq x < 250$	*2*	*73*	*237.5*	*136.82*	*18723.72*	*37447.45*
$250 \leq x < 275$	*0*	*73*	*262.5*	*161.82*	*26190.31*	*0.00*
$275 \leq x < 300$	*1*	*74*	*287.5*	*186.82*	*34996.95*	*34996.95*
Total					*99126.03*	*293829.35*

The quartile For a grouped data

$$Q_i = L_{qi} + \left(\frac{\frac{i \times n}{4} - Cf_{bi}}{f_{qi}} \right) \times w, \quad i = 1, 2, 3$$

Where $i = 1, 2, 3$ *and refers to the quartile number,* L_{qi} *: Lower class boundary of the class counting the quartile* Cf_{bi} *: Cumulative frequency before the* Q_i *class* f_{qi} *: The frequency of the* Q_i *class* w *: Class size of the* Q_i *class.*

1. Q_1 *: Since we want to calculate Q1 we have that* $i = 1$*.*

   Now, since $i = 1$*, we have that* $(iN/4) = (1 * 74/4) = 18.5$*.*

   The value of cumulative frequency just greater than 18.5 *is* 32*. Therefore,* Q_1 *lies in the class interval* $50 - 75$*.*

   Here, $L_{qi} = 50, f_{qi} = 14, w = 25, Cf_{bi} = 18$*. We substitute all these values in the formula.*

$$Q_1 = 50 + \frac{1 * 74/4 - 18}{14} * 25 = 50.8929 \tag{2.10}$$

2. Q_2 *: Since we want to calculate Q1 we have that* $i = 2$*.*

   Now, since $i = 2$*, we have that* $(iN/4) = (2 * 74/4) = 37$*.*

   The value of cumulative frequency just greater than 37 *is* 43*. Therefore,* Q_2 *lies in the class interval* $75 - 100$*.*

   Here, $L_{qi} = 75, f_{qi} = 11, w = 25, Cf_{bi} = 32$*. We substitute all these values in the formula.*

$$Q_2 = 75 + \frac{2 * 74/4 - 32}{11} * 50 = 86.3636 \tag{2.11}$$

3. Q_3 : Since we want to calculate Q1 we have that $i = 3$.

Now, since $i = 1$, we have that $(iN/4) = (3 * 74/4) = 55.5$.

The value of cumulative frequency just greater than 55.5 is 59. Therefore, Q_3 lies in the class interval $125 - 150$.

Here, $L_{qi} = 125, f_{qi} = 8, w = 25, Cf_{bi} = 51$. We substitute all these values in the formula.

$$Q_3 = 125 + \frac{3 * 74/4 - 51}{8} * 25 = 139.0625 \tag{2.12}$$

d.1. Variance:

$$s^2 = \frac{1}{\left(\sum_{i=1}^{n} n_i\right) - 1} \sum_{i=1}^{n} n_i \left(m_i - \bar{X}\right)^2 \tag{2.13}$$

$$s^2 = \frac{293829.35}{74 - 1} \approx 4025.1 \tag{2.14}$$

e. The new average: $m_i^{'} = 3 * m_i, n_i^{'} = 2 * n_i$

$$\bar{X}' = \frac{\sum_{i=1}^{n} n_i^{'} * m_i^{'}}{\left(\sum_{i=1}^{n} n_i^{'}\right)} = \frac{\sum_{i=1}^{n} 2 * n_i * 3 * m_i}{\left(\sum_{i=1}^{n} 2 * n_i\right)} = 3 * \bar{X} \tag{2.15}$$

$$AD^{'} = \frac{1}{\left(\sum_{i=1}^{n} n_i^{'}\right)} \sum_{i=1}^{n} n_i^{'} \mid m_i^{'} - \bar{X}^{'} \mid = \frac{1}{\left(\sum_{i=1}^{n} 2 * n_i\right)} \sum_{i=1}^{n} 2 * n_i \mid 3 * m_i - 3 * \bar{X} \mid = 3 * AD \tag{2.16}$$

Solution 2.7.10. *Given the following grades on a test:*

$$86, 92, 100, 93, 89, 95, 79, 98, 68, 62, 71, 75, 88, 86, 93, 81, 100, 86, 96, 52$$

1. To find out what percentage of scores lie within one standard deviation from the mean and within two standard deviations, you can use Chebyshev's Theorem.

$$I = [\bar{X} - k\sigma \quad \bar{X} + k\sigma]$$

Chebyshev's Theorem states that for any set of data (regardless of its shape), at least $1 - \frac{1}{k^2}$ *of the data falls within* k *standard deviations of the mean.*

2. One standard deviation: - $k = 1$ *- At least* $1 - \frac{1}{1^2} = 0\%$ *of the data falls within one standard deviation from the mean.*

3. Two standard deviations: - $k = 2$ *- At least* $1 - \frac{1}{2^2} = 75\%$ *of the data falls within two standard deviations from the mean.*

In this case :

$$\bar{X} = 84.5$$

$$\sigma^2 = 164.375$$

$$\sigma = 12.83$$

$$I = [58.82110.84]$$

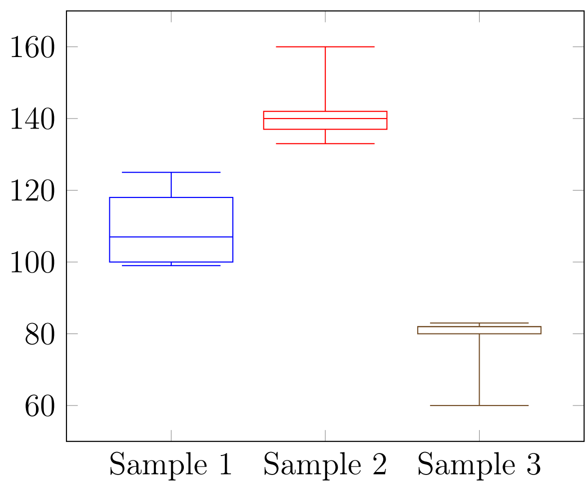
The percentage of the values that lies in $I = [58.82110.84]$ *is*

$$P = \frac{19}{20} * 100 = 95\%$$

So, according to Chebyshev's Theorem, at least 75*pourcent of the scores lie within two standard deviations from the mean. This result is consistent with the theorem.*

Solution 2.7.11. ***Exercise 3:***

The global box-plot:



Sample 1 *Arranging Observations in the ascending order, We get :*

$$87, 99, 100, 105, 105, 107, 108, 118, 119, 125$$

Here, $n = 10$

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ *value of the observation* $= \left(\frac{11}{4}\right)^{th}$ *value of the observation* $= (2.75)^{th}$ *value of the observation* $= 2^{nd}$ *observation* $+0.75\left[3^{rd} - 2^{nd}\right] = 99 + 0.75[100 - 99] = 99 + 0.75(1) = 99 + 0.75 = 99.75$

$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th}$ *value of the observation*

$= \left(\frac{2 \cdot 11}{4}\right)^{th}$ *value of the observation*

$= (5.5)^{th}$ *value of the observation*

$= 5^{th}$ *observation* $+0.5\left[6^{th} - 5^{th}\right]$

$= 105 + 0.5[107 - 105]$

$= 105 + 0.5(2)$

$= 105 + 1$

$= 106$

$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th}$ *value of the observation*

$= \left(\frac{3 \cdot 11}{4}\right)^{th}$ *value of the observation*

$= (8.25)^{th}$ *value of the observation*

$= 8^{th}$ *observation* $+0.25\left[9^{th} - 8^{th}\right]$

$= 118 + 0.25[119 - 118]$

$= 118 + 0.25(1)$

$= 118 + 0.25$

$= 118.25$

Sample 2

Arranging Observations in the ascending order, We get:

$$133, 137, 138, 138, 139, 140, 142, 145, 152, 160$$

Here, $n = 10$ $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ *value of the observation* $= \left(\frac{11}{4}\right)^{th}$ *value of the observation* $= (2.75)^{th}$ *value of the observation* $= 2^{nd}$ *observation* $+0.75\left[3^{rd} - 2^{nd}\right] = 137 + 0.75[138 - 137] = 137 + 0.75(1) = 137 + 0.75 = 137.75$

$$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th} \ \ value\ of\ the\ observation$$
$$= \left(\frac{2 \cdot 11}{4}\right)^{th} \ \ value\ of\ the\ observation$$
$$= (5.5)^{th} \ \ value\ of\ the\ observation$$
$$= 5^{th} \ \ observation \ + 0.5\left[6^{th} \ - 5^{th}\ \right]$$
$$= 139 + 0.5[140 - 139]$$
$$= 139 + 0.5(1)$$
$$= 139 + 0.5$$
$$= 139.5$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} \ \ value\ of\ the\ observation$$
$$= \left(\frac{3 \cdot 11}{4}\right)^{th} \ \ value\ of\ the\ observation$$
$$= (8.25)^{th} \ value\ of\ the\ observation$$
$$= 8^{th} \ observation \ + 0.25\left[9^{th} - 8^{th}\right]$$
$$= 145 + 0.25[152 - 145]$$
$$= 145 + 0.25(7)$$
$$= 145 + 1.75$$
$$= 146.75$$

Sample 3

Arranging Observations in the ascending order, We get:

$$60, 74, 79, 80, 80, 82, 82, 82, 82, 83$$

Here, $n = 10$

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \ \textit{value of the observation}$$

$$= \left(\frac{11}{4}\right)^{th} \ \textit{value of the observation}$$

$$= (2.75)^{th} \ \textit{value of the observation}$$

$$= 2^{nd} \ \ \textit{observation} \ + 0.75 \left[3^{rd} \ - 2^{nd}\ \right]$$

$$= 74 + 0.75[79 - 74]$$

$$= 74 + 0.75(5)$$

$$= 74 + 3.75$$

$$= 77.75$$

$$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th} \ \textit{value of the observation}$$

$$= \left(\frac{2 \cdot 11}{4}\right)^{th} \ \textit{value of the observation}$$

$$= (5.5)^{th} \ \textit{value of the observation}$$

$$= 5^{th} \ \ \textit{observation} \ + 0.5 \left[6^{th} \ - 5^{th}\ \right]$$

$$= 80 + 0.5[82 - 80]$$

$$= 80 + 0.5(2)$$

$$= 80 + 1$$

$$= 81$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} \ \textit{value of the observation}$$

$$= \left(\frac{3 \cdot 11}{4}\right)^{th} \ \textit{value of the observation}$$

$$= (8.25)^{th} \ \textit{value of the observation}$$

$$= 8^{th} \ \ \textit{observation} \ + 0.25 \left[9^{th} \ - 8^{th}\ \right]$$

$$= 82 + 0.25[82 - 82]$$

$$= 82 + 0.25(0)$$

$$= 82 + 0$$

$$= 82$$

1. To determine if the distributions are symmetric, we can use Pearson's coefficient of skewness. Here are the results for the three samples:

- *Sample 1:* $Mean = 107.3$*,* $Median = 106$*,* $SD = 10.5361$ *Coefficient of Skewness ≈ 0.3702 (approximately symmetric)*
- *Sample 2:* $Mean = 142.4$*,* $Median = 139.5$*,* $SD = 7.6315$ *Coefficient of Skewness ≈ 1.14 (moderate positive skewness.)*
- *Sample 3:* $Mean = 78.4$*,* $Median = 81$*,* $SD = 6.6061$ *Coefficient of Skewness ≈ -1.1807 (not symmetric)*

Based on the coefficients of skewness, Samples to be approximately symmetric, as their coefficient are close to 0. However, Sample 3 are left-skewed because its coefficient is significantly negative.

2. To determine which kind of fish has the lowest variation in length, we can compare the coefficient of variation of the three samples:

- *Sample 1: Standard Deviation* (σ_1) *≈ 11.10 ,* $\bar{X} = 107.3$, $CV = \frac{sd}{\bar{X}} = 0.1035 = 10.35\%$
- *Sample 2: Standard Deviation* (σ_1) *≈ 8.044322,* $\bar{X} = 142.4$, $CV = \frac{sd}{\bar{X}} = 0.056491 = 5.6491\%$
- *Sample 3: Standard Deviation* (σ_1) *≈ 6.963396,* $\bar{X} = 78.4$, $CV = \frac{sd}{\bar{X}} = 0.088819 = 8.8819\%$

Sample 2 has the lowest coefficient of variation. Therefore, the second kind of fish has the lowest variation in length.

Chapter 3

Two ways statistics

Two-dimensional descriptive statistics allow for the simultaneous study of two statistical variables on the same sample. To highlight certain relationships, it is necessary to use specific graphs (scatter plots, bar charts parallel to the two variables, etc.).

3.1 Concepts and definitions

In the case of two statistical variables, we consider a statistical table called a contingency table. The x_i variables are arranged in rows and the y_j variables in columns. k_p is the dimension of this table, and n_{ij} is the number of joint values of the values x_i and y_j. The quantities n_{ij} are called the joint frequencies.

Contingency Table

$X\backslash Y$	y_1	$\cdots$	y_j	$\cdots$	Total
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	$n_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	$n_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	n

The quantities $n_{i\cdot}$ $(i = 1, 2, \ldots, k)$ and $n_{\cdot j}$ $(j = 1, 2, \ldots, p)$ are called **marginal frequencies**. They are defined by

$$n_{i\cdot} = \sum_{j=1}^{p} n_{ij} \quad \text{and} \quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}.$$

They satisfy

$$\sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{p} n_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = n.$$

The marginal relative frequencies and the joint relative frequency are defined by

$$f_{i\cdot} = \frac{n_{i\cdot}}{n}, \qquad f_{\cdot j} = \frac{n_{\cdot j}}{n}, \qquad f_{ij} = \frac{n_{ij}}{n}.$$

Conditional Series with Respect to X

Definition 3.1.1. *The variable X conditioned on $Y = y_j$ is denoted by*

$$X/Y = y_j \quad \textit{or simply} \quad X/y_j,$$

and is also called the ***conditional series of*** X ***given*** $Y = y_j$.

In this case, we define the conditional frequency, the conditional mean, and the conditional variance.

Definition 3.1.2. *The* ***conditional frequency*** $f_{i/j}$ *(frequency of x_i knowing y_j), for $i = 1, \dots, k$, is defined by*

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}} = \frac{n_{ij}}{n_{\cdot j}}.$$

The **conditional mean** $\overline{X}_{/y_j}$ is the mean of the values of X under the condition $Y = y_j$. It is given by

$$\overline{X}_{/y_j} = \sum_{i=1}^{k} f_{i/j} x_i = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} n_{ij} x_i.$$

The **conditional variance** is given by

$$s^2_{X/y_j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} n_{ij}(x_i - \overline{X}_{/y_j})^2 = \sum_{i=1}^{k} f_{i/j}(x_i - \overline{X}_{/y_j})^2.$$

Covariance

Definition 3.1.3. *The covariance of two discrete statistical variables X and Y is the quantity denoted by* $\mathrm{Cov}(X, Y)$ *or* s_{XY}, *defined by*

$$s_{XY} = \frac{1}{n} \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})(y_j - \overline{Y}).$$

Proposition 3.1.1. *The covariance can also be expressed by the following formula:*

$$s_{XY} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i y_j - \overline{X}\,\overline{Y}.$$

Proof. We know that

$$\sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} = n, \qquad \sum_{j=1}^{p} n_{ij} = n_{i\cdot}, \qquad \sum_{i=1}^{k} n_{ij} = n_{\cdot j}.$$

Starting from the definition of covariance, we have

$$\begin{aligned} s_{XY} &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} (x_i - \overline{X})(y_j - \overline{Y}) \\ &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i y_j - \frac{\overline{Y}}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i - \frac{\overline{X}}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} y_j + \frac{\overline{X}\,\overline{Y}}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}. \end{aligned}$$

Using the relations

$$\sum_{j=1}^{p} n_{ij} = n_{i\cdot}, \qquad \sum_{i=1}^{k} n_{ij} = n_{\cdot j},$$

we obtain

$$s_{XY} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i y_j - \frac{\overline{Y}}{n} \sum_{i=1}^{k} n_{i\cdot} x_i - \frac{\overline{X}}{n} \sum_{j=1}^{p} n_{\cdot j} y_j + \overline{X}\,\overline{Y}.$$

Since

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} n_{i\cdot} x_i, \qquad \overline{Y} = \frac{1}{n} \sum_{j=1}^{p} n_{\cdot j} y_j,$$

we finally get

$$s_{XY} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} x_i y_j - \overline{X}\,\overline{Y}.$$

□

Linear Correlation Coefficient

Definition 3.1.4. *The* ***correlation coefficient*** *of two statistical variables X and Y is the quantity denoted by r_{XY}, defined by*

$$r_{XY} = \frac{s_{XY}}{s_X s_Y},$$

where s_{XY} is the covariance between X and Y, and s_X, s_Y are the standard deviations of X and Y respectively.

Proposition 3.1.2. *For any quantitative statistical variables X and Y, the following properties hold:*

1. $r_{XY} = r_{YX}$,
2. $r_{XY} = \mathrm{Cov}\left(\frac{X-\overline{X}}{s_X}, \frac{Y-\overline{Y}}{s_Y}\right)$,
3. $-1 \leq r_{XY} \leq 1$.

3.2 Bivariate distribution exercises

Exercice 3.2.1. *the following table shows the different values obtained for two quantitative variables X and Y in a sample of size 8*

X	1.3	1.5	2.5	2.5	2.7	3	4	5
Y	2.3	3.5	3.5	4.5	4.7	3	2	1

1. *Draw the scatter plot.*
2. *Calculate the covariance and the linear correlation coefficient.*
3. *Can we conclude in this case that there is a linear link between X and Y ? What does the scatter plot suggest instead?*

Exercice 3.2.2. *A survey is carried out with a clientele (533 individuals) to study their willingness to pay (X in DA) a telephone plan. This clientele was segmented according to a criterion Y which takes 3 modalities A, B and C. After consultation, we obtain the following contingency table*

$X\backslash Y$	A	B	C	Total
]0-5]	38	11	0	49
]5-10]	55	63	0	118
]10-15]	53	76	0	129
]15-20]	32	62	1	95
]20-25]	6	24	8	38
125-30]	5	4	40	49
]30-35]	2	1	37	40
]35-40]	1	0	12	13
]40-45]	0	0	2	2
Total	192	241	100	533

We give the following partial results

$$\sum_i n_{iA}x_{iA} = 2105, \sum_i n_{iA}x_{iA}^2 = 31750$$

$$\sum_i n_{iB}x_{iB} = 3217.5, \sum_i n_{iB}x_{iB}^2 = 50706.25$$

$$\sum_i n_{iC}x_{iC} = 3035, \sum_i n_{iC}x_{iC}^2 = 94175$$

1. *Determine the conditional means of X . Deduce the marginal mean of X .*
2. *Determine the marginal and conditional variances.*
3. *Compute the correlation between the three modalities of Y two by two. Do you think that this sigmentation makes sense?*

Exercice 3.2.3. *The following table contains the information of age, monthly income, health condition and frequency of exercise last month of 10 respondents.*

Age (years)	*Monthly Income (DA)*	*Health Condition*	*Frequency of exercise per month*
62	*68080*	*Very good*	*22*
36	*48000*	*Very good*	*22*
55	*84100*	*Good*	*15*
55	*77000*	*Fair*	*16*
27	*48510*	*Excellent*	*30*
37	*38400*	*Poor*	*5*
49	*68560*	*Poor*	*10*
49	*56380*	*Poor*	*10*
26	*22970*	*Good*	*22*
38	*44330*	*Fair*	*13*

1. *Calculate the correlation coefficient between the age and the monthly income. Hence, interpret the value obtained.*

2. *Draw a scatter diagram to represent the relationship between the age and the monthly income and show the point* $(\bar{X}, \bar{Y})$*. Interpret your result.*

3. *Draw the cross tabulation then calculate the Spearman's rank correlation coefficient between health condition and frequency of exercise per month. Hence, comment on the result obtained*

Exercice 3.2.4. *Let* X *and* Y *be two quantitative statistical variables. Then*

$$r_{XY} = \frac{n\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_iy_j - \left(\sum_{i=1}^{k} n_{i.}x_i\right)\left(\sum_{j=1}^{p} n_{.j}y_j\right)}{\sqrt{\left[n\sum_{i=1}^{k} n_{i.}x_i^2 - \left(\sum_{i=1}^{k} n_{i.}x_i\right)^2\right]\left[n\sum_{j=1}^{p} n_{.j}y_j^2 - \left(\sum_{j=1}^{p} n_{.j}y_j\right)^2\right]}} \tag{3.1}$$

3.3 Solution of exercises

Solution 3.3.1. 1. *Draw the scatter plot:*

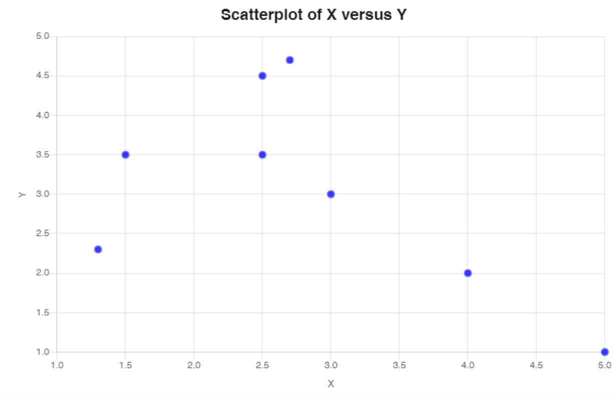


Figure 3.1

2. *Calculate the covariance and the linear correlation coefficient.*

x_i	$x_i - \mu_x$	y_i	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
1.3	*-1.5125*	*2.3*	-0.7625	*1.15328*
1.5	*-1.3125*	*3.5*	*0.4375*	*-0.5742 .*
2.5	-0.3125	*3.5*	*0.4375*	*-0.1367:*
2.5	-0.3125	*4.5*	*1.4375*	*-0.4492:*
2.7	*-0.1125*	*4.7*	*1.6375*	*-0.1842 .*
3	*0.1875*	*3*	*-0.0625*	*-0.0117 .*
4	*1.1875*	*2*	*-1.0625*	*-1.2617:*
5	*2.1875*	*1*	*-2.0625*	*-4.5117:*

$$\sigma_{XY} = -\frac{5.97625}{8}$$
$$\sigma_{XY} = -0.747$$

The linear correlation coefficient.

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{13}$$
$$-1 \leq \rho(X,Y) \leq 1$$

$(x_i-\mu_X)^2$	$(y_i-\mu_Y)$	$(y_i-\mu_Y)^2$	$(x_i-\mu_X)(y_i-\mu_Y)$
2.28765625	*-0.7625*	*0.58140625*	*1.15328125*
1.72265625	*0.4375*	*0.19140625*	*-0.57421875*
0.09765625	*0.4375*	*0.19140625*	*-0.13671875*
0.09765625	*1.4375*	*2.06640625*	*-0.44921875*
0.01265625	*1.6375*	*2.68140625*	*-0.18421875*
0.03515625	*-0.0625*	*0.00390625*	*-0.01171875*
1.41015625	*-1.0625*	*1.12890625*	*-1.26171875*
4.78515625	*-2.0625*	*4.25390625*	*-4.51171875*
$\sum(x_i-\mu_X)^2$ $= 11.09875$	$\sum(y_i-\mu_Y)$ =*-5.97625*	$\sum(y_i-\mu_Y)^2$ =*11.094*	$\sum(x_i-\mu_X)(y_i-\mu_Y)$ $= 10.44875$

$$
\begin{aligned}
\sigma_X &= \sqrt{\frac{10.44875}{8}} \\
&= \sqrt{1.3061} \\
\sigma_X &= 1.1428 \\
\sigma_Y &= \sqrt{\frac{11.09875}{8}} \\
&= \sqrt{1.3873} \\
\sigma_Y &= 1.1779 \\
\rho_{XY} &= \frac{1}{8} \times \frac{-5.97625}{1.1428 \times 1.1779} \\
&= \frac{1}{8} \times \frac{-5.97625}{1.3461} \\
&= \frac{-5.97625}{10.7688} \\
\rho_{XY} &= -0.555
\end{aligned}
$$

3. Since $\mathrm{Cov}(X;Y) < 0$*; then we can say that the relationship between the two variables is negative. In this case, where* $\rho = -0.5555$*, it suggests a moderate negative linear correlation. This means that there is a tendency for one variable to decrease as the other increases, but the relationship is not perfect*

Solution 3.3.2.

(a) *We Compute the conditional means (averages) of the statistical variable X*

$$E(X \mid Y = A) \approx \frac{\sum_i n_{iA} x_{iA}}{n_A} = \frac{2105}{192} \approx 10.9635$$
$$E(X \mid Y = B) \approx \frac{\sum_i n_{iB} x_{iB}}{n_B} = \frac{3217.5}{241} \approx 13.3506$$
$$E(X \mid Y = C) \approx \frac{\sum_i n_{iC} x_{iC}}{n_C} = \frac{3035}{100} \approx 30.35$$

Using the previous calculus, we can easily deduce the marginal average (mean) of X

$$E(X) = \overline{X} = \sum_y P(Y = y) \cdot E(X \mid Y = y)$$

$$E(X) = \frac{192}{533} \cdot 10.9635 + \frac{241}{533} \cdot 13.3506 + \frac{100}{533} \cdot 30.35 = 15.6801$$

(b) *Determine the conditional and marginal variances:*

i. *The Conditional Variances:*

$$\operatorname{Var}(X \mid Y = A) \approx \frac{\sum_i n_{iA} x_{iA}^2}{n_A} - [E(X \mid Y = A)]^2 \approx \frac{31750}{192} - (10.9635)^2 \approx 45.1662$$
$$\operatorname{Var}(X \mid Y = B) \approx \frac{\sum_i n_{iB} x_{iB}^2}{n_B} - [E(X \mid Y = B)]^2 \approx \frac{50706.25}{241} - (13.3506)^2 \approx 32.1608$$
$$\operatorname{Var}(X \mid Y = C) \approx \frac{\sum_i n_{iC} x_{iC}^2}{n_C} - [E(X \mid Y = C)]^2 \approx \frac{94175}{100} - (30.35)^2 \approx 20.6275$$

ii. *The Marginal Variance:*

Step 1: Let us calculate $E\left(X^2\right)$ *:*

$$\begin{aligned} E\left(X^2\right) = \overline{X^2} = & \frac{192}{533} \cdot [\operatorname{Var}(X \mid Y = A) + [E(X \mid Y = A)]^2] \\ & + \frac{241}{533} \cdot \left[\operatorname{Var}(X \mid Y = B) + [E(X \mid Y = B)]^2\right] \\ & + \frac{100}{533} \cdot [\operatorname{Var}(X \mid Y = C) + [E(X \mid Y = C)]^2] \qquad (14) \\ E\left(X^2\right) = \overline{X^2} = & = \frac{192}{533} \cdot \left[45.1662 + (10.9635)^2\right] \\ & + \frac{241}{533} \cdot \left[32.1608 + (13.3506)^2\right] \\ & + \frac{100}{533} \cdot \left[20.6275 + (30.35)^2\right] \approx 331.907 \end{aligned}$$

Step 2: We then calculate $\operatorname{Var}(X)$ *:*

$$\text{Var}_{marg}(X) = E\left(X^2\right) - [E(X)]^2 = \overline{X^2} - \overline{X}^2$$

$$\text{Var}_{marg}(X) = 331.907 - (15.6801)^2 = 86.0415$$

(c) *We calculate the correlation Between Modalities of Y Two by Two:(the case between A and B)*

We use the following formula

$$\mathrm{r} = \frac{\sum\left(\mathrm{x_i} - \overline{\mathrm{X}}\right)\left(\mathrm{y_i} - \overline{\mathrm{Y}}\right)}{\sqrt{\sum\left(\mathrm{x_i} - \overline{\mathrm{X}}\right)^2 \sum\left(\mathrm{y_i} - \overline{\mathrm{Y}}\right)^2}}$$

Now we have to find mean of both data sets:
Mean for Data set X:
$\bar{X} = \frac{192}{9}$
$\overline{\mathrm{X}} = 21.333$
Mean for Data set Y: $\overline{\mathrm{Y}} = \frac{241}{9}$
$\overline{\mathrm{Y}} = 26.778$

$\mathbf{x_i} - \overline{\mathbf{X}}$	$\left(\mathbf{x_i} - \overline{\mathbf{X}}\right)^2$	$\mathbf{y_i} - \overline{\mathbf{Y}}$	$\left(\mathbf{y_i} - \overline{\mathbf{Y}}\right)^2$	$\left(\mathbf{x_i} - \overline{\mathbf{X}}\right)\left(\mathbf{y_i} - \overline{\mathbf{Y}}\right)$
16.667	*277.789*	*-15.778*	*248.945*	*-262.971926*
33.667	*1133.467*	*36.222*	*1312.033*	*1219.486074*

and then we construct the following table,

$\mathbf{x_i} - \overline{\mathbf{X}}$	$\left(\mathbf{x_i} - \overline{\mathbf{X}}\right)^2$	$\mathbf{y_i} - \overline{\mathbf{Y}}$	$\left(\mathbf{y_i} - \overline{\mathbf{Y}}\right)^2$	$\left(\mathbf{x_i} - \overline{\mathbf{X}}\right)\left(\mathbf{y_i} - \overline{\mathbf{Y}}\right)$
16.667	*277.789*	*-15.778*	*248.945*	*-262.971926*
33.667	*1133.467*	*36.222*	*1312.033*	*1219.486074*
31.667	*1002.799*	*49.222*	*2422.805*	*1558.713074*
10.667	*113.785*	*35.222*	*1240.589*	*375.713074*
-15.333	*235.101*	*-2.778*	*7.717*	*42.595074*
-16.333	*266.767*	*-22.778*	*518.837*	*372.033074*
-19.333	*373.765*	*-25.778*	*664.505*	*498.366074*
-20.333	*413.431*	*-26.778*	*717.061*	*544.477074*
-21.333	*455.097*	*-26.778*	*717.061*	*571.255074*

$\sum\left(\mathrm{x_i} - \overline{\mathrm{X}}\right)^2 = 4272.000001, \quad \sum\left(\mathrm{y_i} - \overline{\mathrm{Y}}\right)^2 = 7849.555556,$
$\sum\left(\mathrm{x_i} - \overline{\mathrm{X}}\right)\left(\mathrm{y_i} - \overline{\mathrm{Y}}\right) = 4919.666666$

The general formula of the correlation coefficient:

$$\mathrm{r} = \frac{\sum\left(\mathrm{x_i}-\overline{\mathrm{X}}\right)\left(\mathrm{y_i}-\overline{\mathrm{Y}}\right)}{\sqrt{\sum\left(\mathrm{x_i}-\overline{\mathrm{X}}\right)^2\sum\left(\mathrm{y_i}-\overline{\mathrm{Y}}\right)^2}}$$

Now putting values in the above equation:

$$r = \frac{4919.666666}{\sqrt{(4272.000001)(7849.555556)}}$$

$$r = \frac{4919.666666}{5790.7945}$$

$$r = \frac{4919.666666}{5790.7945}$$

$$r = 0.8496$$

The value of r is 0.8496 which give a very Strong and Positive Correlation.

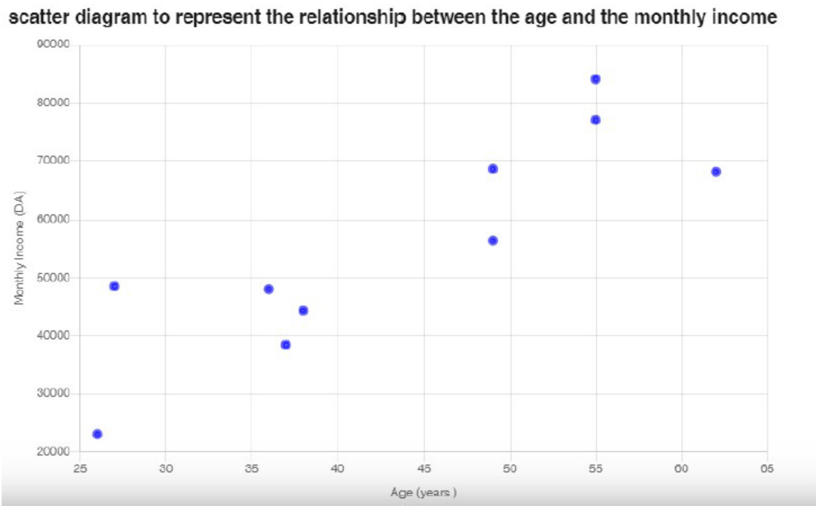
Solution 3.3.3.

1. *Calculate the correlation coefficient between the age and the monthly income:*

$$\bar{X} = 43.4, \bar{Y} = 55633 \tag{15}$$

$$r = \frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\left(\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2\right)}}$$
$$r = \frac{1799128}{\sqrt{(1374.4\times 3190117410)}} = \mathbf{0.8592}$$

2. *Draw a scatter diagram to represent the relationship between the age and the monthly income*



3. Draw the cross tabulation then calculate the Spearman's rank correlation coefficient between health condition and frequency of exercise per month.

Step 1: Cross-tabulation

Health Condition	R_1	*Frequency*	R_2	$d_i = R_2 - R_1$	d_i^2
Very good	*2.5*	*22*	*3*	*0.5*	*0.25*
Very good	*2.5*	*22*	*3*	*0.5*	*0.25*
Good	*4.5*	*15*	*6*	*1.5*	*2.25*
Fair	*6.5*	*16*	*5*	*-1.5*	*2.25*
Excellent	*1*	*30*	*1*	*0*	*0*
Poor	*9*	*5*	*10*	*1*	*1*
Poor	*9*	*10*	*8.5*	*-0.5*	*0.25*
Poor	*9*	*10*	*8.5*	*-0.5*	*0.25*
Good	*4.5*	*22*	*3*	*-1.5*	*2.25*
Fair	*6.5*	*13*	*7*	*0.5*	*0.25*
					$\sum d_i^2 = 9$

Hence the formula can is written as follows:

$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{m_1 \times \left(m_1^2 - 1\right)}{12} + \frac{m_2 \times \left(m_2^2 - 1\right)}{12} + \dots\right]}{N\left(N^2 - 1\right)}$$

In the first rank R_1, we have the following repetitions: 2.5 two times, 4.5 two times, 6.5 two times and 9 three times.

In the second rank R_2, we have the following repetitions: 3 three times, 8.5 two times. So,

$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{2\times\left(2^2-1\right)}{12} + \frac{2\times\left(2^2-1\right)}{12} + \frac{2\times\left(2^2-1\right)}{12} + \frac{2\times\left(2^2-1\right)}{12} + \frac{3\times\left(3^2-1\right)}{12} + \frac{3\times\left(3^2-1\right)}{12}\right]}{N\left(N^2-1\right)}$$

Then,

$$r_s = 1 - \frac{6[9+2+4]}{10\times 99}$$
$$r_s = 0.9106$$

Solution 3.3.4.

$$\begin{aligned}
r_{XY} = \frac{s_{XY}}{s_x s_y} &= \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_i y_j - \overline{XY}}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i^2 - \bar{X}^2\right)\left(\frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j^2 - \bar{Y}^2\right)}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_i y_j - \left(\frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i\right)\left(\frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j\right)}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{k} n_{i.}x_i\right)^2\right]\left[\frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j^2 - \left(\frac{1}{n}\sum_{j=1}^{p} n_{.j}y_j\right)^2\right]}} \\
&= \frac{\frac{1}{n^2}\left[n\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_i y_j - \left(\sum_{i=1}^{k} n_{i.}x_i\right)\left(\sum_{j=1}^{p} n_{.j}y_j\right)\right]}{\left[\sqrt{n\sum_{i=1}^{k} n_{i.}x_i^2 - \left(\sum_{i=1}^{k} n_{i.}x_i\right)^2}\right]\left[n\sum_{j=1}^{p} n_{.j}y_j^2 - \left(\sum_{j=1}^{p} n_{.j}y_j\right)^2\right]} \qquad (3.2) \\
&= \frac{\sqrt{n\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}x_i y_j - \left(\sum_{i=1}^{k} n_{i.}x_i\right)\left(\sum_{j=1}^{p} n_{.j}y_j\right)]}}{\sqrt{\left[n\sum_{i=1}^{k} n_{i.}x_i^2 - \left(\sum_{i=1}^{k} n_{i.}x_i\right)^2\right]\left[n\sum_{j=1}^{p} n_{.j}y_j^2 - \left(\sum_{j=1}^{p} n_{.j}y_j\right)^2\right]}}
\end{aligned}$$

3.4 Regression study for statistical variables

Exercice 3.4.1. *We carried out the adjustment of a points of a couple of statistical variables (X, Y). The equations obtained are the following:*
For the regression line $D_{Y/X} : y = x + 30$,
For the regression line $D_{X/Y} : x = \frac{1}{4}y + 60$.

1. *Deduce the value of the linear correlation coefficient.*
2. *Calculate the arithmetic means of the variables X and Y.*
3. *Calculate the covariance between X and Y, given that the variance of Y is equal to 40 .*

Exercice 3.4.2. *Let P be the percentage of executives of a company whose net monthly salary is greater than the value R (in $10^4 DA$)*

R	8	10	15	20	35	50
P	80.3	55.8	28.1	15.2	5.1	2.5

1. *Plot the scatter plot. What kind of adjustment can we suggest?*
2. *Let $x_i = \log(R_i)$ and $y_i = \log(P_i)$. Draw the scatter plot (x_i, y_i).*
3. *Calculate the linear correlation coefficient $\rho(X, Y)$. Comment.*
4. *Give the equation of the regression line $y = ax+b$ and the equation of the curve $P = f(R)$.*
5. *Calculate P if $R = 77060$.*

Exercice 3.4.3. *A sample of the evolution of electricity consumption in mega Wh between 1975 and 2003 is given by*

X (Year)	1975	1979	1983	1987	1991	1995	1999	2003
Y (consumption)	60	82	112	146	194	246	330	414

1. *Represent the scatter plot corresponding to these observations.*
2. *We set $u_i = \frac{x_i - 1989}{9.798}$ and $v_i = \log(y_i)$. Calculate the $(u_i, v_i)'s$, represent the results using a table then in a scatter plot of the corresponding points. In view of the result obtained, what do you think is the purpose of the previous transformations?*
3. *We denote by (U, V) the pair whose observations are the $(u_i, v_i)'s$. Calculate $\bar{U}, \bar{V}, \sigma_U^2, \sigma_V^2$, $\operatorname{cov}(U, V)$ and $\rho(U, V)$.*
4. *Write the linear regression equation of V on U then measure its goodness of fit.*
5. *Deduce a relation between Y and X.*
6. *Assuming the evolution of electricity consumption followed during the years after the same rhythm, give an estimate of the consumption in the year 2007.*

3.5 Solution of exercises

Solution 3.5.1. *1.*

2. *Deduce the value of the linear correlation coefficient.:*

$$r = \sqrt{slope\ of\ D_{Y/X} \times slope\ of\ D_{X/Y}}$$

$$r = \sqrt{1 \times 0.25} = 0.5$$

So, the linear correlation coefficient is 0.5.

3. *Calculate the arithmetic means of the variables X and Y:*

Solving the systems of two equations:

$$\begin{cases} y = x + 30 \\ x = \frac{1}{4}y + 60 \end{cases} \tag{3.3}$$

we find the solution :$\bar{X} = 90$ and $\bar{Y} = 120$

4. *We know that,*

$$a' = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} \tag{3.4}$$

Form the given $a^{'} = \frac{1}{4}$ in the regression $D_{X/Y}$ and $Var(Y) = 40$ we can observe that

$$0.25 = \frac{\mathrm{Cov}(X,Y)}{40} \tag{3.5}$$

$\Rightarrow$

$$Cov(X,Y) = 10 \tag{3.6}$$

Solution 3.5.2.

1. *Calculate the means $\bar{X}$ and $\bar{Y}$ and the standard deviations σ_X and σ_Y of the variables X and Y as well as their correlation coefficient:*

   (a) *Calculate the means $\bar{X}$:*

$$\bar{X} = \frac{\sum_{i=1}^{10} x_i}{10} = 75$$

$$\bar{Y} = \frac{\sum_{i=1}^{10} y_i}{10} = 62.8$$

$\mathbf{x_i - \overline{X}}$	$\left(x_i - \overline{\mathbf{X}}\right)^2$	$y_i - \overline{\mathbf{Y}}$	$\left(y_i - \overline{\mathbf{Y}}\right)^2$	$\left(x_i - \overline{X}\right)\left(y_i - \overline{\mathbf{Y}}\right)$
-45	2025	-51.8	2683.24	2331
-35	1225	-47.8	2284.84	1673
-25	625	-37.8	1428.84	945
-15	225	-27.8	772.84	417
-5	25	-15.8	249.64	79
5	25	-0.8	0.64	-4
15	225	15.2	231.04	228
25	625	32.2	1036.84	805
35	1225	52.2	2724.84	1827
45	2025	82.2	6756.84	3699

(3.7)

$$\sum \left(x_i - \overline{X}\right)^2 = 8250, \quad \sum \left(y_i - \overline{Y}\right)^2 = 18169.6$$
$$\sum \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = 12000$$
$$SD_X = \sqrt{\tfrac{8250}{10}} = 28.72 \quad SD_Y = \sqrt{\tfrac{18169.6}{10}} = 42.62$$

Let use the the general expression to calculate r the correlation coefficient

$$r = \frac{\sum \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum \left(x_i - \overline{X}\right)^2 \sum \left(y_i - \overline{Y}\right)^2}}$$

Now let us substitute the values into the previous equation

$$r = \tfrac{12000}{\sqrt{8250 \times 18169.6}}$$
$$r = \tfrac{12000}{12243.3329}$$
$$r = 0.9801$$

The value of r is 0.9801 .

Analyze: Very Strong and Positive Correlation.

2. *The scatter plot and regression:*

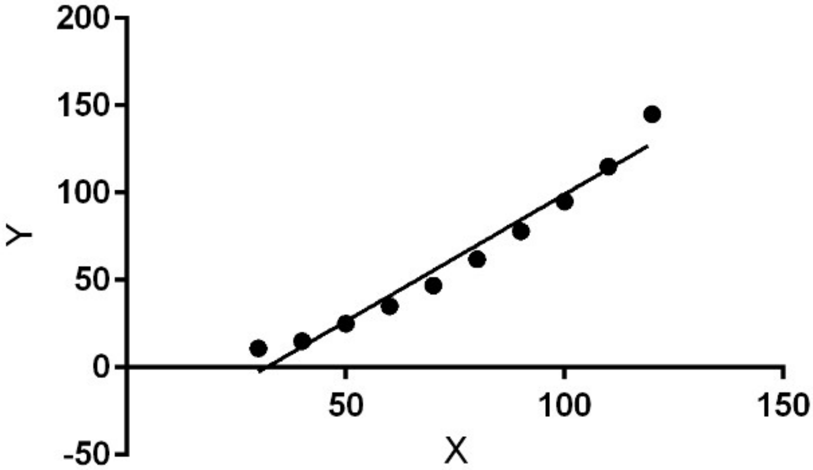


Figure 3.2

The regression line is defined by the equation: $D_{Y/X} y = ax + b$, *where*

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = 1.45 \tag{3.8}$$

and

$$b = \bar{Y} - a\bar{X} = -45.95 \tag{3.9}$$

So

$$D_{Y/X} : y = 1.45x - 45.96 \tag{3.10}$$

3.

$$\rho(X,Y) = r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = 0.9801.$$

Calculate the residual variance of Y *and the variance explained by the line* $D_{Y/X}$

$$V_R(Y) = \mathrm{Var}(Y)\left(1 - \rho(X,Y)^2\right) = 1816.96(1 - 0.9801) = 1816.9401$$
$$V_E(Y) = \mathrm{Var}(Y)\rho(X,Y)^2 = 1816.96 * 0.9801^2 = 1745.36$$

then we calculate the coefficient of determination

$$R^2 = \frac{V_E(Y)}{\mathrm{Var}(Y)} = \frac{1745.36}{1816.96} = 0.9605$$

The regression lines explain 96.05% *of the variance (of* X *and* Y *) which gives that the* *goodness of fit is very good.*

4. *We use this equation to estimate the braking distance if the truck is traveling at 130 km/h:*

$$y = 1.415 * 130 - 45.96 = 142.5m$$

Solution 3.5.3. 1. *Plot the given data in a scatter plot. The x-axis represents the years (X), and the y-axis represents electricity consumption (Y).*

X	1975	1979	1983	1987	1991	1995	1999	2003
Y	60	82	112	146	194	246	330	414

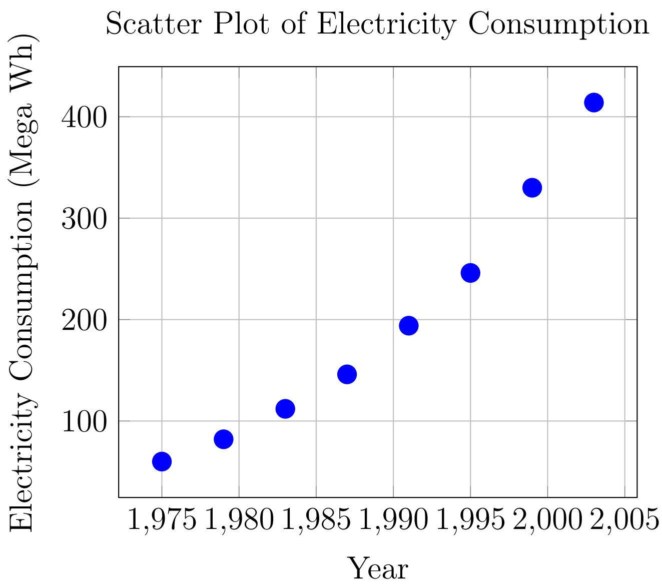


Figure 3.3: Scatter Plot of Electricity Consumption

2. *Define* $u_i = \frac{z_i - 1989}{9.798}$ *and* $v_i = \log(y_i)$*. Calculate the transformed values* (u_i, v_i)*.*

X	1975	1979	1983	1987	1991	1995	1999	2003
Y	60	82	112	146	194	246	330	414
u	-1.42	-1.02	-0.61	-0.20	0.20	0.61	1.02	1.42
v	4.09	4.40	4.71	4.98	5.26	5.50	5.79	6.02

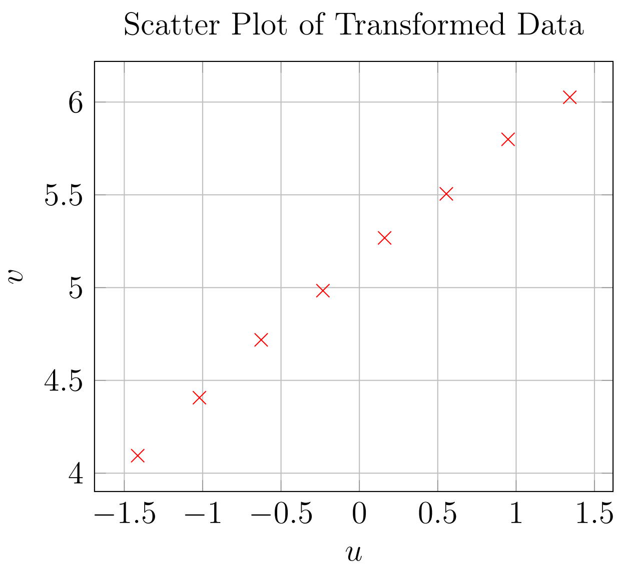


Figure 3.4: Scatter Plot of Transformed Data

The purpose of the transformations seems to be linearizing the relationship between the variables.

3. *Calculate the sample mean ($\bar{U}$, $\bar{V}$), sample variances (σ_U^2, σ_V^2), covariance ($cov(U,V)$), and the correlation coefficient ($\rho(U,V)$).*

$$\bar{U} = 0, \quad \bar{V} = 5.0938$$

$$\sigma_U^2 = s^2 = \frac{\sum_{i=1}^{n} \left(u_i - \bar{U}\right)^2}{n} = \frac{6.9378}{8} = 0.8672$$

$$\sigma_V^2 = s^2 = \frac{\sum_{i=1}^{n} \left(v_i - \bar{V}\right)^2}{n} = \frac{3.1843}{8} = 0.3980$$

$$cov(U,V) = 0.58$$

$$\rho(U,V) = 0.9992$$

4. *Write the linear regression equation for V on U ($V = aU + b$).*

$$V = 0.6769U + 5.094 \tag{3.11}$$

The coefficient of determination (R^2) is a measure of the goodness of fit.

First wr calculate,

$$V_E(V) = Var(V)\rho(U,V)^2 = 0.3980 * 0.9992^2 = 0.3973 \tag{3.12}$$

$$R^2 = \frac{V_E(V)}{\mathrm{Var}(V)} = \frac{0.3973}{0.3980} = 0.9984$$

the goodness of fit is very good.

5. *Relation between V and U:*

   The relation between U and V is perfect.

6. *Assuming the same rhythm of consumption, estimate the consumption in the year 2007 by plugging $X = 2007$ into the regression equation:*

$$Y_{2007} \approx e^{5.09+0.67\left(\frac{2007-1989}{9.798}\right)\approx 556.0529}$$

Exercice 3.5.1. ***Proposed exercise***

The braking distance of a truck driving on a dry road with the respect to its speed is given by the following table

Speed (km/h)	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*
Distance (m)	*11*	*15*	*25*	*35*	*47*	*62*	*78*	*95*	*115*	*145*

1. *Calculate the means $\bar{X}$ and $\bar{Y}$ and the standard deviations σ_X and σ_Y of the variables X and Y as well as their correlation coefficient.*

2. *Draw the scatter pllot. What kind of connection is there between X and Y ? Justify.*

3. *Determine the regression line of Y on X. Plot this line on the same graph. Show the center of gravity $(\bar{X}, \bar{Y})$.*

4. *Does the regression line fit the data closely? Measure it.*

5. *Use this equation to estimate the braking distance if the truck is traveling at 130 km/h.*

We give $\sum_{i=1}^{10} x_i = 750, \sum_{i=1}^{10} y_i = 628, \sum_{i=1}^{10} x_i y_i = 59100, \sum_{i=1}^{10} x_i^2 = 64500, \sum_{i=1}^{10} y_i^2 = 57608$.

Chapter 4

Introduction to probability

Probabilities are used to model a random experiment, that is, a phenomenon whose outcome cannot be predicted with certainty and for which the final result is considered to be governed by chance.

Examples:

- the unborn child will be a girl.
- the football team ESM will defeat MCA in their next match.
- the die will show an even number.

We study a **random experiment**. The **sample space**, also called the **set of possible outcomes**, describes the set of all possible results of the experiment. Each of these results is called an **elementary event** (or elementary outcome).

The sample space is usually denoted by Ω, and an elementary outcome by ω.

An **event** is a subset of Ω, or equivalently a union of elementary events. An event is said to **occur** if one of the elementary events that compose it occurs.

Events are sets and are usually represented by **capital letters**.

- We toss a coin three times. The **elementary events** describe the outcome of this experiment as precisely as possible. Thus, an elementary event ω is an ordered triple (r_1, r_2, r_3) giving the results of the three tosses (in order).

  The event B: *"we obtain heads on the second toss"* is

  $$B = \{(t, h, t), (t, h, h), (h, h, t), (h, h, h)\}.$$

  The event B occurs if one of the elementary events listed above occurs.

Sometimes it is not necessary to keep all these details. One may instead choose ω to represent the number of heads obtained. Then,

$$\Omega = \{0, 1, 2, 3\}.$$

This model is much simpler, but it does not allow us to describe events such as B.

Notation	**Set-theoretic vocabulary**	**Probabilistic vocabulary**
Ω	universal set	certain event
$\varnothing$	empty set	impossible event
ω	element of Ω	elementary event
A	subset of Ω	event
$\omega \in A$	ω belongs to A	A occurs
$A \subset B$	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 or $\overline{A}$	complement of A	complementary event of A
$A \cap B = \varnothing$	A and B are disjoint	A and B are incompatible

A probability is a function on $\mathcal{P}(\Omega)$, the power set of Ω, such that:

1. $0 \leq P(B) \leq 1$, for all events $A \subset \Omega$
2. $P(A) = \sum_{\omega \in B} P(\omega)$, for all events B
3. $P(\theta) = \sum_{\omega \in \theta} P(\theta) = 1$

Proposition 4.0.1. *Let B et A be two events.*

1. *If B et A are mutually exclusive, $P(B \cup A) = P(B) + P(A)$.*
2. $P(B^c) = 1 - P(B)$.
3. $P(\emptyset) = 0$.
4. $P(B \cup A) = P(B) + P(A) - P(b \cap A)$.

Definition 4.0.1. *Let Ω be a random experiment and Ω the associated space of possibilities. A probability on Ω is a function, defined on the set of events, that satisfies:*
- *Axiom 1 : $0 \leq P(B) \leq 1$, for any event B*
- *Axiom 2: For any sequence of pairwise incompatible events $(B_i)_{i \in \mathbb{N}}$),*

$$P\left(\bigcup_{i\in\mathbb{N}} B_i\right) = \sum_{i\in\mathbb{N}} P\left(B_i\right)$$

- *Axiom 3:* $P(\Omega) = 1$

The probability of an event A is easily calculated:

$$P(B) = \sum_{\omega\in B} P(\theta) = \frac{\text{card}(B)}{\text{card}(\Omega)}$$

Definition 4.0.2. *For a given two events A and B, with $P(A) > 0$, the probability of B conditional on A, is denoted $P(B \mid A)$ and defined by:*

$$P(A \mid B) = \frac{P(B\cap A)}{P(B)}$$

We can also write $P(B\cap A) = P(A \mid B)P(B)$.
Use 1: When $P(B)$ and $P(B\cap A)$ are easy to calculate, we can deduce $P(A \mid B)$.
Use 2: When $P(A \mid B)$ and $P(B)$ are easy to find, we can obtain $P(B\cap A)$.

Example 4.0.1. *An urn contains r red balls and v green balls. Two balls are drawn one after the other (without replacement). What is the probability of obtaining two red balls?*

Let us choose a sample space Ω that describes the outcomes of the experiment precisely:

$$\Omega = \{\textit{red, green}\} \times \{\textit{red, green}\}.$$

An elementary event is an ordered pair (x, y), where x is the color of the first ball drawn and y is the color of the second one.
Let A be the event "the first ball is red" and B the event "the second ball is red".
Then,

$$P(A\cap B) = P(B \mid A)\, P(A) = \frac{r-1}{r+v-1}\cdot\frac{r}{r+v}.$$

What is the probability that the second ball drawn is red? We keep the same formalism.

$$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$$

$$= \frac{r-1}{r+v-1}\cdot\frac{r}{r+v} + \frac{r}{r+v-1}\cdot\frac{v}{r+v}$$

$$= \frac{r}{r+v}.$$

Proposition 4.0.2. *(The Law of Total Probability) Let B be an event such that $0 < P(B) < 1$. For any event A, we have:*

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$$

Proposition 4.0.3. *(Generalized Law of Total Probability) Let $(A_i)_{i \in I}$ be a partition of Ω, such that $P(A_i) > 0$, for all $i \in I$. Then, for any event B,*

$$P(B) = \sum_{i \in I} P(B \mid A_i) P(A_i)$$

The law of total probability allows us to follow the steps of a random experiment in chronological order. We will now look at a formula for going back in time...

Theorem 4.0.1. *(Bayes' Theorem) Let A and B be two events such that $0 < P(A) <$1 and $P(B) > 0$. Therefore,*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}$$

Proof:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

and we conclude by substituting $P(B)$ with its expression given by the law of total probability.

Theorem 4.0.2. *(Generalized Bayes formula) Let $(A_i)_{i \in I}$ be a partition of Ω, such that $P(A_i) > 0$, for all $i \in I$. Consider an event B, such that $P(B) > 0$. Then, for all $i \in I$,*

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{\sum_{j \in I} P(B \mid A_j) P(A_j)}$$

Definition 4.0.3. *Two events A and B are said to be independent if:*

$$P(B \cap A) = P(B)P(A)$$

If they have non-zero probability, then:

$$P(A \mid B)=P(A) \Longleftrightarrow P(B \mid A)=P(B) \Longleftrightarrow P(B \cap A)=P(B) P(A)$$

4.1 Exercices

Exercice 4.1.1. *1. An urn contains 12 balls numbered from 1 to 12. We randomly draw one, and consider the events*

$$\begin{aligned} A= & \text{ "drawing an even number"}, \\ B= & \text{ "drawing a multiple of 3"}. \end{aligned}$$

Are events A and B independent?

2. Repeat the question with an urn containing 13 balls.

Exercice 4.1.2. *Your neighbor has two children whose sex you do not know. Consider the following three events - $A=$ "the two children are of different sexes" - $B=$ "the eldest is a girl" - $C=$ "the youngest is a boy".*

Show that A, B and C are two by two independent, but are not mutually independent. Assume that the probability at birth of having a girl (respectively a boy) is equal to $1/2$.

Exercice 4.1.3. *Let $B_1, \ldots, B_n$ be n events in a probability space (Ω, P). Assume that they are mutually independent and have respective probabilities $p_i=P\left(B_i\right)$. Give a simple expression for $P\left(B_1 \cup \cdots \cup B_n\right)$ as a function of $p_1, \ldots, p_n$.*

Application: Assume that a person is subjected to n experiences that are independent of each other and that for each experience, they have a probability $\boldsymbol{p}$ of having an accident. What is the probability that they have at least one accident?

Exercice 4.1.4. *1. Let Q, S, and D be three events. Show that:*

$$P(Q \cup S \cup D)=P(Q)+P(S)+P(D)-P(Q \cap S)-P(Q \cap D)-P(S \cap D)+P(Q \cap S \cap D)$$

2. We have three electrical components C_1, C_2, and C_3 with probabilities of functioning p_i, which are completely independent of one another. Determine the probability of the circuit functioning:

1. If the components are arranged in series.

2. *If the components are arranged in parallel.*

3. *If the circuit is mixed: C_1 is arranged in series with the sub-circuit consisting of C_2 and C_3 in parallel.*

Exercice 4.1.5. *Let $(Y_i)_{i\in\mathbb{N}}$ be a sequence of independent random variables, all following the same distribution. We set, for $n \geq 1, Z_n = Y_1 + \cdots + Y_n$ and we consider $a \in \mathbb{R}$. Prove the following equivalence:*

$$P(Y_1 \geq a) > 0 \Longleftrightarrow \forall n \geq 1, P(Z_n \geq na) > 0$$

Exercice 4.1.6. *An urn contains 5 identical red balls and 3 identical black balls. We draw two balls without replacement.*

1. *What is the probability that the first is red and the second is black?*

2. *What is the probability that at least one of the two balls is red?*

3. *Knowing that at least one of the two balls is red, what is the probability that the other is black?*

Exercice 4.1.7. *In a laboratory, the following observations were made:*

** if a mouse carries antibody T, then 2 times out of 5 it also carries antibody R;*

** if a mouse does not carry antibody T, then 4 times out of 5 it does not carry antibody R.*

Half of the population carries antibody A.

1. *Calculate the probability that, if a mouse carries antibody R, then it also carries antibody T.*

2. *Calculate the probability that, if a mouse does not carry antibody R, then it does not carry antibody T.*

Exercice 4.1.8. *Three brands of coffee, X, Y, and Z, are to be ranked according to taste by a judge. Define the following events:*

- *A: Brand X is preferred to Y .*

- *B : Brand X is ranked best.*

- *C: Brand X is ranked second best.*

- *D: Brand X is ranked third best.*

If the judge actually has no taste preference and randomly assigns ranks to the brands, is event A independent of events B, C*, and D ?*

Exercice 4.1.9. *If A and B are independent events, show that A and $\bar{B}$ are also independent. Are $\bar{A}$ and $\bar{B}$ independent?*

Exercice 4.1.10. *It is known that a patient with a disease will respond to treatment with probability equal to 0.9. If three patients with the disease are treated and respond independently, find the probability that at least one will respond.*

Exercice 4.1.11. *A local fraternity is conducting a raffle where 50 tickets are to be sold one per customer. There are three prizes to be awarded. If the four organizers of the raffle each buy one ticket, what is the probability that the four organizers win*

1. *all of the prizes?*
2. *exactly two of the prizes?*
3. *exactly one of the prizes?*
4. *none of the prizes?*

4.2 Solution

Solution 4.2.1.

1. *We have:*

$$A = \{2, 4, 6, 8, 10, 12\}$$
$$B = \{3, 6, 9, 12\}$$
$$A \cap B = \{6, 12\}.$$

So we have $P(A) = 1/2, P(B) = 1/3$ and $P(A \cap B) = 1/6 = P(A)P(B)$. The events A and B are independent.

2. *The events A, B and $A \cap B$ are again described in exactly the same way. However, this time we have: $P(A) = 6/13, P(B) = 4/13$ and $P(A \cap B) = 2/13 \neq 24/169$.*

The events A and B are not independent. This is consistent with intuition. There is no longer the same distribution of even balls and odd balls, and in the multiples of 3 between 1 and 13, the distribution of even and odd numbers has remained unchanged.

Solution 4.2.2. *Clearly, we have* $P(B) = P(C) = \frac{1}{2}$*. The four possibilities for the two children, assumed to be equiprobable, are* $(\mathrm{F}, \mathrm{G}), (\mathrm{F}, \mathrm{F}), (\mathrm{G}, \mathrm{G}), (\mathrm{G}, \mathrm{F})$*. We deduce that* $P(A) = \frac{1}{2}$*. Then* $A \cap B$ *corresponds to* (F, G) *and therefore* $P(A \cap B) = \frac{1}{4} = P(A)P(B)$*. We deduce that* A *and* B *are independent. We prove in the same way that* A *and* C *are independent, and that* B *and* C *are independent. Finally,* $A \cap B \cap C = B \cap C$ *and therefore* $P(A \cap B \cap C) = \frac{1}{4} \neq P(A)P(B)P(C)$*. The three events are not mutually independent.*

Solution 4.2.3. *If the events* $B_1, \ldots, B_n$ *are mutually independent, the complementary events* $\overline{B_1}, \ldots, \overline{b_n}$ *are also independent. We deduce that*

$$\begin{aligned} P\left(B_1 \cup \cdots \cup B_n\right) &= 1 - P\left(\overline{B_1} \cap \cdots \cap \overline{B_n}\right) \\ &= 1 - \prod_{i=1}^{n} P\left(\overline{B_i}\right) \\ &= 1 - \prod_{i=1}^{n} (1 - p_i) \end{aligned}$$

In this particular case, the probability of having at least one accident is given by $1 - (1-p)^n$

Solution 4.2.4.

1. We proceed the solution in two steps. Firstly:

$$P((Q \cup S) \cup D) = P(Q \cup S) + P(D) - P((Q \cup S) \cap D)$$

We know that,

$$P((Q \cup S) \cap D) = P((Q \cap D) \cup (S \cap D)) = P(Q \cap D) + P(S \cap D) - P(Q \cap S \cap D)$$

and

$$P(Q \cup S) = P(Q) + P(S) - P(Q \cap S)$$

This is also called the Poincaré sieve formula, it generalizes to several events by recurrence.

2. We denote by K_i *the event: 'component* C_i *works'. By hypothesis, events* F_i *are independent of each other. We must calculate for the first case* $P\left(K_1 \cap K_2 \cap K_3\right)$ *(the circuit formed by the three components arranged in series works if and only if the three components work), for the second* $P\left(K_1 \cup K_2 \cup K_3\right)$ *(the circuit is formed by the three components arranged in parallel* *works if and only if one of the three components works), and for the third* $P\left(K_1 \cap \left(K_2 \cup K_3\right)\right)$*. We then have:*

2.1. By independence of the events:

$$P\left(K_1 \cap K_2 \cap K_3\right) = p_1 p_2 p_3$$

2.2. According to the previous formula, and independently of the events:

$$P\left(K_1 \cup K_2 \cup K_3\right) = p_1 + p_2 + p_3 - p_1 p_2 - p_1 p_3 - p_2 p_3 + p_1 p_2 p_3$$

2.3. The event $K_2 \cup K_3$ *is independent of* K_1*. We therefore have:*

$$P\left(K_1 \cap \left(K_2 \cup K_3\right)\right) = P\left(K_1\right) P\left(K_2 \cup K_3\right) = P\left(K_1\right)\left(P\left(K_2\right) + P\left(K_3\right) - P\left(K_2 \cap K_3\right)\right)$$

and

$$P\left(F_1 \cap \left(F_2 \cup F_3\right)\right) = p_1\left(p_2 + p_3 - p_2 p_3\right)$$

Solution 4.2.5. *The converse is trivial since it corresponds only to the case* $\boldsymbol{n} = 1$*. For the direct sense, we notice that*

$$\left(S_n \geq na\right) \supset \left(Y_1 \geq a\right) \cap \left(Y_2 \geq a\right) \cap \cdots \cap \left(Y_n \geq a\right)$$

which implies that

$$P\left(S_n \geq na\right) \geq P\left(\left(Y_1 \geq a\right) \cap \left(Y_2 \geq a\right) \cap \cdots \cap \left(Y_n \geq a\right)\right)$$

Since the random variables are independent, it yields

$$P\left(S_n \geq na\right) \geq \prod_{j=1}^{n} P\left(y_j \geq a\right)$$

Finally, since the Y_i *all have the same law as* X_1*, we then obtain*

$$P\left(S_n \geq na\right) \geq \left(P\left(X_1 \geq a\right)\right)^n$$

Thus, if $P(Y_1 \geq a) > 0$, we have $P(S_n \geq na) > 0$ for any integer $n \geq 1$, which proves the direct meaning.

Solution 4.2.6.

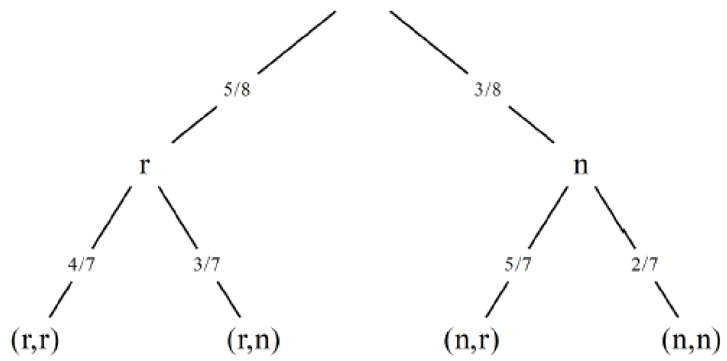


Figure 4.1

1. *$P(r,n) = \frac{5}{8} \cdot \frac{3}{7} = \frac{15}{56}$ with $A =$ "the first ball is red"; $B =$ "the second ball is black"; $A \cap B =$ "the first is red and the second is black"; $B \mid A =$ "the second is black knowing that the first is red". We will remember that, in a tree, the branches carry conditional probabilities. The probability of a path is equal to the product of the probabilities carried by the branches.*

2.

$$\begin{aligned}
&P\left(l' \text{ ' at least one of the two balls is red }\right)\\
&= P(\text{ the first is red and the second black })\\
&\quad + P \text{ (the first is black and the second red)}\\
&\quad + P \text{ (the first is red and the second red)}\\
&= P \text{ (the first is red)} \cdot P \text{ (the second is black } \mid \text{ the first is red)}\\
&\quad + P \text{ (the first is black)} \cdot P \text{ (the second is red } \mid \text{ the first is black)}\\
&\quad + P \text{ (the first is red)} \cdot P \text{ (the second is red } \mid \text{ the first is red)}\\
&= \frac{5}{8} \times \frac{3}{7} + \frac{3}{8} \times \frac{5}{7} + \frac{5}{8} \times \frac{4}{7} = \frac{25}{28}
\end{aligned}$$

3. *In order to simplify our results and expression we denote*

$$P(n,r) = P_{n,r}\ ,\ P(r,n) = P_{r,n} \text{ and } P(r,r) = P_{r,r}$$

$$
\begin{aligned}
&P(\textit{ the other ball is black / at least one of the two is red})\\
&= \frac{P(\textit{ one ball is red and the other is black })}{P(\textit{ at least one of the two balls is red })}\\
&= \frac{P_{r,n} + P_{n,r}}{P_{r,n} + P_{n,r} + P_{r,r}} = \frac{\frac{5}{8}\times\frac{3}{7}+\frac{3}{8}\times\frac{5}{7}}{\frac{5}{8}\times\frac{3}{7}+\frac{3}{8}\times\frac{5}{7}+\frac{5}{8}\times\frac{4}{7}} = \frac{3}{5} = 0.6
\end{aligned}
$$

Solution 4.2.7.

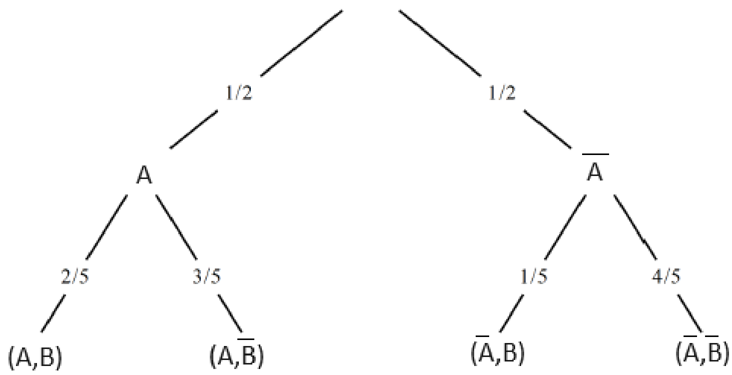


Figure 4.2

1. *For the partition* $(T, \bar{T})$*, the Bayes formula gives*

$$
\begin{aligned}
P(R) &= P(T)\cdot P(R \mid T) + P(\bar{T})\cdot P(R \mid \bar{T}) = \frac{1}{2}\cdot\frac{2}{5}+\frac{1}{2}\cdot\frac{1}{5} = \frac{3}{10}\\
P(T \mid R) &= \frac{P(T)\cdot P(R \mid T)}{P(R)} = \frac{\frac{1}{2}\cdot\frac{2}{5}}{\frac{3}{10}} = \frac{2}{3}\\
P(\bar{T} \mid R) &= \frac{P(\bar{T})\cdot P(B \mid \bar{T})}{P(R)} = \frac{\frac{1}{2}\cdot\frac{1}{5}}{\frac{3}{10}} = \frac{1}{3}
\end{aligned}
$$

Interpretation:

- *prior probabilities:* $P(T) = \frac{1}{2}, P(\bar{T}) = \frac{1}{2}$*;*
- *information contributions:* $P(R \mid T) = \frac{2}{5}, P(R \mid \bar{T}) = \frac{1}{5}$*;*
- *posterior probabilities for carriers of antibody* R *:*

$$
P(T \mid R) = \frac{2}{3}, P(\bar{T} \mid R) = \frac{1}{3}
$$

Probability of causes :

- *The* $\frac{2}{3}$ *of mice carrying antibody* R *also carry antibody* T*, indicating a clear incidence of* T *on* R*.*

2. *For the partition* $(T, \bar{T})$*, the Bayes formula gives*

$$
\begin{aligned}
P(\bar{R}) &= P(T) \cdot P(\bar{R} \mid T) + P(\bar{T}) \cdot P(\bar{R} \mid \bar{T}) = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{4}{5} = \frac{7}{10} \\
P(T \mid \bar{R}) &= \frac{P(T) \cdot P(\bar{R} \mid T)}{P(\bar{R})} = \frac{\frac{1}{2} \cdot \frac{3}{5}}{\frac{7}{10}} = \frac{3}{7} \\
P(\bar{T} \mid \bar{R}) &= \frac{P(\bar{T}) \cdot P(\bar{R} \mid \bar{T})}{P(\bar{R})} = \frac{\frac{1}{2} \cdot \frac{4}{5}}{\frac{7}{10}} = \frac{4}{7}
\end{aligned}
$$

Interpretation:

- *a priori probabilities:* $P(T) = \frac{1}{2}, P(\bar{T}) = \frac{1}{2}$*;*
- *information contributions:* $P(\bar{R} \mid T) = \frac{3}{5}, P(\bar{R} \mid \bar{T}) = \frac{4}{5}$*;*
- *posterior probabilities for non-carriers of antibody* R*:*

$$P(T \mid \bar{R}) = \frac{3}{7}, P(\bar{T} \mid \bar{R}) = \frac{4}{7}$$

Probability of causes:

- *The* $\frac{4}{7}$ *of mice that do not carry antibody* R *also do not carry antibodies* T*, which perhaps indicates a slight incidence of* $\bar{T}$ *on* $\bar{R}$*.*

Solution 4.2.8. *The six equally likely sample points for this experiment are given by*

$$
\begin{gathered}
E1 : XYZ, E3 : YXZ, E5 : ZXY \\
E2 : XZY, E4 : YZX, E6 : ZYX,
\end{gathered}
$$

where $XY\,Z$ *denotes that* X *is ranked best,* Y *is second best, and* Z *is last. Then* $A = \{E1, E2, E5\}, B = \{E1, E2\}, C = \{E3, E5\}, D = \{E4, E6\}$*, and it follows that*

$$
\begin{gathered}
P(A) = 1/2, P(A \mid B) = P(A \cap B) \\
P(B) = 1, P(A \mid C) = 1/2, P(A \mid D) = 0
\end{gathered}
$$

We then conclude that: the events A *and* C *are independent, but events* A *and* B *are dependent. The events* A *and* D *are also dependent.*

Solution 4.2.9.

$$P(A \mid \bar{B}) = P(A \cap \bar{B})/P(\bar{B}) = \frac{P(\bar{B} \mid A)P(A)}{P(\bar{B})} = \frac{[1 - P(B \mid A)]P(A)}{P(\bar{B})} = \frac{[1 - P(B)]P(A)}{P(\bar{B})} =$$

$\frac{P(\bar{B})P(A)}{P(\bar{B})} = P(A)$. *So, A and* $\bar{B}$ *are independent.*
$P(\bar{B} \mid \bar{A}) = P(\bar{B} \cap \bar{A})/P(\bar{A}) = \frac{P(\bar{A}|\bar{B})P(\bar{B})}{P(\bar{A})} = \frac{[1-P(A|\bar{B})]P(\bar{B})}{P(\bar{A})}$.

From the above, A and $\bar{B}$ *are independent. So* $P(\bar{B} \mid \bar{A}) = \frac{[1-P(A)]P(\bar{B})}{P(\bar{A})} = \frac{P(\bar{A})P(\bar{B})}{P(\bar{A})} = P(\bar{B})$. *So,* $\vec{A}$ *and* $\bar{B}$ *are independent*

Solution 4.2.10. *A: At least one of the three patients will respond. B1: The first patient will not respond. B2: The second patient will not respond. B3: The third patient will not respond. Then observe that* $\bar{A} = \mathrm{B1} \cap\ \mathrm{B2} \cap\ \mathrm{B3}$ *and* $P(A) = 1 - P(\bar{A}) = 1 - P(B1 \cap B2 \cap B3)$. *Applying the multiplicative law, we have*

$$P(B1 \cap B2 \cap B3)P(B1)P(B2 \mid B1)P(B3 \mid B1 \cap B2)$$

where, because the events are independent,

$$P(B2 \mid B1) = P(B2) = 0.1 \text{ and } P(B3 \mid B1 \cap B2) = P(B3) = 0.1$$

Substituting $P(Bi) = 0.1, i = 1, 2, 3,$ *we obtain* $P(A) = 1 - (0.1)^3 = 0.999$. *Notice that we have demonstrated the utility of complementary events. This result is important because frequently it is easier to find the probability of the complement,* P(A)*, than to find* P(A) *directly.*

Solution 4.2.11. *There are* $\binom{50}{3} = 19600$ *ways to choose the 3 winncis. Each of these is equally likely.*

1. *There are* $\binom{4}{3} = 4$ *ways for the organizers to win all of the prizes. The probability is* $\frac{4}{19600}$.
2. *There are* $\binom{4}{2}\binom{46}{1} = 276$ *ways the organizers can win two prizes and one of the other 46 people to win the third prize. So, the probability is 276/19600.*
3. $\binom{4}{1}\binom{46}{2} = 4,140$. *The probability is* $\frac{4140}{19600}$.
4. $\binom{46}{3} = 15180$. *The probability is* $\frac{15180}{19600}$.

Chapter 5

Midterm Exam and Final Exam

5.1 Midterm Exam

Midterm Exam

2022/2023

2023/2024

2022-2023

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
National Higher School of Mathematics

NATIONAL HIGHER SCHOOL OF MATHEMATICS
المدرسة الوطنية
العليا في الرياضيات

Midterm examt N°1

Name :	ID :	Groupe:

Exercice 1. Among these assertions, specify which ones are true and which ones are false.

	True	False
1- The sample mean is always higher than the population mean.		√
2- The median is always higher than the mean.		√
3- The mean of (1, 2, 8, 9) is 10.		√
4- The STD of a sample is always less than 1.		√
5- The STD of a sample can be -1.		√
6- When median > mean then the distribution is symmetrical.		√
7- Increasing all values of a sample by 5, will increase the mean by 5.	√	
8- Increasing all values of a sample by 5, will increase STD by 5.		√
9- Multipying all values by 2, will multiply the mean by 2.	√	
10- Multiplying all values by 2, will multiply the STD by 2.	√	
11- The mean is at the middle of the histogram.		√
12- There is 50% chance that one random sample will be in the IQR. (IQR=$Q_3 - Q_1$)	√	
13- We can learn the IQR by looking at the boxplot.	√	
14- If a value is not in the IQR, then it is an outlier.		√
15- If a value is higher than the mean then it is an outlier.		√
16- Removing an outlier that is positive, will increase the mean.		√
17- Removing an outlier that is negative, will most likely decrease the STD.	√	
18- A histogram of any distribution is bell-shaped.		√
19- On the x-axis, a histogram has always positive values.		√
20- The mean is equal to 0 for a symmetric histogram.		√

Exercice 2. We took note of the performance bonuses (in DA) of the employees of a company. The results are given in the following statistical table :

Bonuses	frequency	Rel. Frequency	Cum. Frequency	Cum. Rel. Frequency
$[750, 1250[$	26	**0.130**	**26**	**0.130**
$[1250, 1750[$	**71**	**0.355**	97	**0.485**
$[1750, 2250[$	**43**	**0.215**	**140**	0.700
$[2250, 2750[$	**26**	**0.130**	166	**0.830**
$[2750, 3250[$	**18**	**0.09**	**184**	**0.920**
$[3250, 3750[$	**10**	0.05	**194**	**0.970**
$[3750, 4250[$	6	0.03	**200**	**1.00**

1. Complete the statistical table above.
2. What is the bonus value received by 50% of the employees.

Bonus value received by 50% of the employees ($n = 200$) represents the median $\widetilde{X} \in [1750, 2250[$ and given by

$$\widetilde{X} = L_1 + \left(\frac{\frac{n}{2} - Cf_b}{f_m}\right) \times w = 1750 + \left(\frac{\frac{200}{2} - 97}{43}\right) \times (2250 - 1750) = \mathbf{1784.9}$$

3. Give Q_1, Q_3 and inerquartile range

The first quartile Q_1 is located in the class $[1250, 1750[$ and the third quartle Q_3 is in $[2250, 2750[$ and both are given by

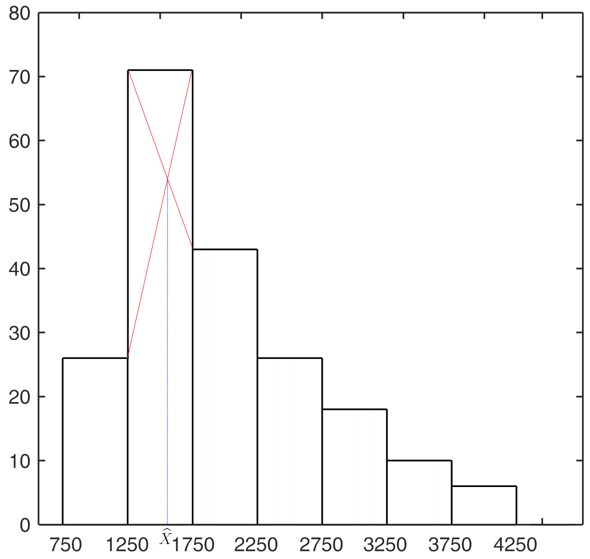
$$Q_1 = L_1 + \left(\frac{\frac{n}{4} - Cf_{q_1}}{f_{q_1}}\right) \times w = 1250 + \left(\frac{\frac{200}{4} - 26}{71}\right) \times (1750 - 1250) = \mathbf{1419}$$

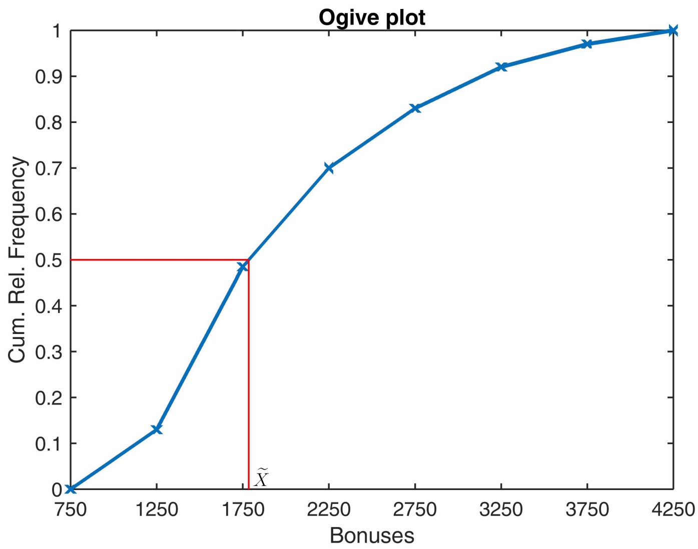
and

$$Q_3 = L_1 + \left(\frac{\frac{3n}{4} - Cf_{q_3}}{f_{q_3}}\right) \times w = 2250 + \left(\frac{\frac{600}{4} - 140}{26}\right) \times (2750 - 2250) = \mathbf{2442.3}.$$

The inerquartile range is $IQR = Q_3 - Q_1 = 2442.3 - 1419 = \mathbf{1023.3}$.

4. Plot the histogram and the Ogive plot of the above distribution then determine the mode and median of the series graphically





5. Give the modale class then calculate the mode value.

Clearly, The highest frequency is for the second class $[1250, 1750[$ to which the mode $\widehat{X}$ belongs and it's given by

$$\widehat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times w = 1250 + \left(\frac{(71 - 26)}{(71 - 26) + (71 - 43)}\right) \times (1750 - 1250) = \mathbf{1558.2}$$

6. Calculate the mean, variance, standard deviation and coefficient of variation.

Since the data are grouped, we need to compute the med point of each class and the formulas to be used are :

$$\mu = \overline{X} = \frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} = \frac{396500}{200} = \mathbf{1982.5}$$

$$\sigma^2 = Var(X) = \frac{1}{\sum_{i=1}^{N} f_i} \sum_{i=1}^{N} f_i (x_i - \mu)^2 = \frac{8874643.75}{200} = \mathbf{44373.22}$$

and

$$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{N} f_i} \sum_{i=1}^{N} f_i (x_i - \mu)^2} = \sqrt{44373.21875} = \mathbf{210.65}.$$

The coefficient of variation is given by

$$CV = \frac{\sigma}{\mu} \times 100 = CV = \frac{210.65}{1982.5} \times 100 = \mathbf{10.625\%}$$

7. Do you think the distribution of the performance bonuses is symmetric? justify.

No, the distribution of the performance bonuses is not symmetric characterized by the shape of the histogram, also by computing pearson's coefficient of skewness

$$sk = \frac{3 \times (1982.5 - 1784.9)}{210.65} = \mathbf{2.8141}$$

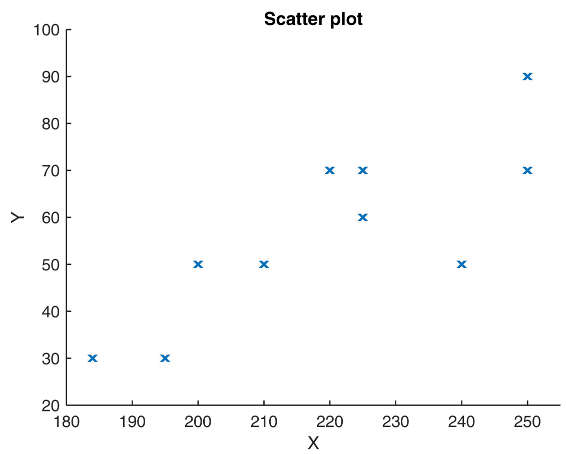
we can observe that its value is close to 3 which let us conclude that the distribution of the performance bonuses is asymmetric

Exercice 3. Wheat production (X in quintals) and the number of rainy days (Y) were recorded. The following results were obtained:

X	200	184	225	250	240	195	210	225	250	220
Y	50	30	70	90	50	30	50	60	70	70
$(x_i - \overline{X})^2$	396.01	1288.81	26.01	906.01	404.01	620.01	98.01	26.01	906.01	0.01
$(y_i - \overline{Y})^2$	49	729	169	1089	49	729	49	9	169	169
$(x_i - \overline{X})(y_i - \overline{Y})$	139.3	969.3	66.3	993.3	-140.7	672.3	69.3	15.3	391.3	1.3

$\sum x_i = 2199, \sum y_i = 570, \sum(x_i - \overline{X})^2 = 4670.9, \sum(y_i - \overline{Y})^2 = 3210, \sum(x_i - \overline{X})(y_i - \overline{Y}) = 3177.$

1. Draw the scatter plot of the given data. Describe the relationship between X and Y.



the plotted points fall in a narrow band and show a rising tendency from the lower left hand corner to the upper right hand corner, therefore, it suggests a **High Degree of Positive Linear Correlation**

2. Calculate the correlation coefficient and interpret the obtained result.

The above scatter plot suggested the presence of linear correlation, hence, we choose pearson's coefficient of correlation to measure it.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y} = \frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum(x_i - \overline{X})^2}\sqrt{\sum(y_i - \overline{Y})^2}} = \frac{3177}{\sqrt{4670.9} \times \sqrt{3210}} = \mathbf{0.820\,47}$$

which indicates the existence of a strong positive linear correlation between the two variables.

Exercice 4. Let X be a numeric and discrete statistical variable with the modalities $x_1, ..., x_n$. Denote by $f_1, ..., f_n$ the associated frequences.
Show that :

$$\sum_{i=1}^{n} f_i(x_i - \overline{X})^2 = \min_{a\in\mathbb{R}} \sum_{i=1}^{n} f_i(x_i - a)^2$$

Set

$$F(a) = \sum_{i=1}^{n} f_i(x_i - a)^2$$

clearly F is differentiable, hence, we calculate F' then we solve the equation $F'(a) = 0$ for a possible minimum(s) or maximum(s) by checking the signe of F''.

$$\begin{aligned} F'(a) = -\sum_{i=1}^{n} f_i(x_i - a) = 0 &\Longrightarrow -\sum_{i=1}^{n} f_i x_i + \sum_{i=1}^{n} f_i a = 0 \\ \Longrightarrow a\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i &\Longrightarrow a = \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} f_i x_i = \overline{X}, \end{aligned}$$

moreover, $\forall a \in \mathbb{R}, F''(a) = \sum_{i=1}^{n} f_i = 1 > 0$ which means the F has only one minimum at $a = \overline{X}$, thus

$$\sum_{i=1}^{n} f_i(x_i - \overline{X})^2 = \min_{a\in\mathbb{R}} \sum_{i=1}^{n} f_i(x_i - a)^2$$

Midterm N°2

Name :	Groupe:

Exercice 1. Among these assertions, specify which ones are true and which ones are false.

	True	False
1- If $P(A \cup B) = 0.5$ and $P(A \cap B) = 0.5$, then $P(A) = P(B)$.	$\surd$	
2- if $P(A) = 0.4, P(B) = 0.5$ and $P(A \cup B) = 0.7$ then $P(A \cap B) = 0.3$.		$\surd$
3- If $P(A\|B) = 0.5$ and $P(B) = 0.5$, then A and B are necessarily independent.		$\surd$
4- If two events are independent, then $p(A\|B) = p(B\|A)$.		$\surd$
5- If two events are independent then both $P(A\|B) = P(A)$ and $P(B\|A) = P(A)$.		$\surd$
6- If $P(A\|B) = 0.6$ and $P(B) = 0.2$ then $P(A \cap B) = 0.12$.	$\surd$	
7- If $P(A \cap B) \geq 0.10$ then $P(A) \geq 0.10$.	$\surd$	
8- If A and B are mutually exclusive events, then $P(A\|B) = 0$.	$\surd$	
9- Bayes' rule states that $p(A\|B) = p(B\|A)$.		$\surd$
10- A certainty event has a probability equal to 0.9.		$\surd$
11- Any random variable is a measurable function.	$\surd$	
12- A random variable can be qualitatif		$\surd$
13- $X \stackrel{d}{=} Y$ if $\forall x \in \mathbb{R}, F_X(x) = F_Y(x)$.	$\surd$	
14- $x_1 \leq x_2 \Longrightarrow F(x_1) \leq F(x_2)$.	$\surd$	
15- B and $A \cap B^c$ are disjoint.	$\surd$	
16- $P(A) - P(B) \neq P(A \cap B^c)$.		$\surd$
17- If $A_i =]0, \frac{1}{i}[$ then $\lim_{i\to\infty} P(A_i) = 0$.	$\surd$	
18- $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cup B^c$.		$\surd$
19- $P(A_1 \cup A_2 \cup ...) \geq P(A_1) + P(A_2) + ...$		$\surd$
20- $E[XY] = E[X]E[Y]$.		$\surd$

Exercice 2. We give the following table

x_i	1	2	3	4	5	6	7	8	9	10
y_i	1504	809	540	438	375	320	290	254	231	212
$(x_i - \overline{X})$	-4.5	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5	3.5	4.5
$(y_i - \overline{Y})$	1006.7	311.7	42.7	-59.3	-122.3	-177.3	-207.3	-243.3	-266.3	-285.3
$(x_i - \overline{X})(y_i - \overline{Y})$	-4530.15	-1090.95	-106.75	88.95	61.15	-88.65	-310.95	-608.25	-932.05	-1283.85
R_{y_i}	10	9	8	7	6	5	4	3	2	1
d_i^2	81	49	25	9	1	1	9	25	49	81

1. Calculate the correlation coefficients of pearson and spearman then explain the difference between the two obtained results

$$\overline{X} = 5.5 \quad , \quad \overline{Y} = 497.3$$

$$\sum (x_i - \overline{X})^2 = 82.5 \quad , \quad \sum (y_i - \overline{Y})^2 = 1416814.1 \quad , \quad \sum (x_i - \overline{X})(y_i - \overline{Y}) = -8801.5$$

$$\sum d_i^2 = 330$$

We have

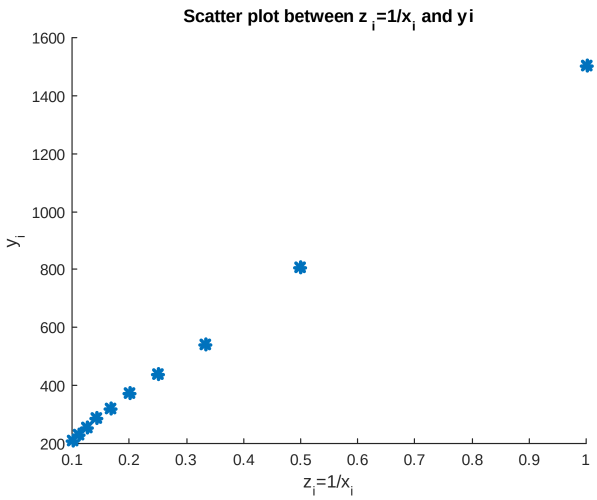
$$\rho(X,Y) = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum (x_i - \overline{X})^2 \sum (y_i - \overline{Y})^2}} = \frac{-8801.5}{\sqrt{82.5 \times 1416814.1}} = -0.81409$$

and

$$r_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)} = 1 - \frac{6 \times 330}{10 \times (10^2 - 1)} = -1.0$$

Spearman's coefficient is equal to -1, which indicates a perfect negative monotonic trend, and since pearson's coefficient is not equal to r_s, which means that the relationship is not perfect negative linear, therefore, the relationship between the number of units and the unit price is not linear.

2. Set $z_i = \frac{1}{x_i}$, calculate z_i then draw the scatter plot between z_i and the unit price y_i. Comment.



Clearly, the relationship beween z_i and y_i is linear

Exercice 3. Consider the data in the table below. In this table Y is the purity of oxygen produced in a chemical distillation process, and X is the percentage of hydrocarbons that are present in the main condenser of the distillation unit.

Obs. N°	X	Y	$(x_i - \overline{X})$	$(y_i - \overline{Y})$	$(x_i - \overline{X})(y_i - \overline{Y})$	$y_i - \widehat{y}_i$	$\widehat{y}_i - \overline{Y}$
1	0.99	90.01	−0.194	−2.247	0.436	0.652	−2.899
2	1.02	89.05	−0.164	−3.207	0.526	−0.756	−2.451
3	1.15	91.43	−0.034	−0.827	0.028	−0.319	−0.508
4	1.29	93.74	0.106	1.483	0.157	−0.101	1.584
5	1.46	96.73	0.276	4.473	1.235	0.349	4.125
6	1.36	94.45	0.176	2.193	0.386	−0.437	2.630
7	0.87	87.59	−0.314	−4.667	1.465	0.026	−4.692
8	1.23	91.77	0.046	−0.487	−0.022	−1.174	0.687
9	1.55	99.42	0.366	7.163	2.622	1.694	5.470
10	1.40	93.65	0.216	1.393	0.301	−1.835	3.228
11	1.19	93.54	0.006	1.283	0.008	1.194	0.090
12	1.15	92.52	−0.034	0.263	−0.009	0.771	−0.508
13	0.98	90.56	−0.204	−1.697	0.346	1.352	−3.049
14	1.01	89.54	−0.174	−2.717	0.473	−0.116	−2.600
15	1.11	89.85	−0.074	−2.407	0.178	−1.301	−1.106

1. Measure the correlation between the two variables then interpret it.

$$\overline{X} = 1.184 \quad , \quad \overline{Y} = 92.257$$

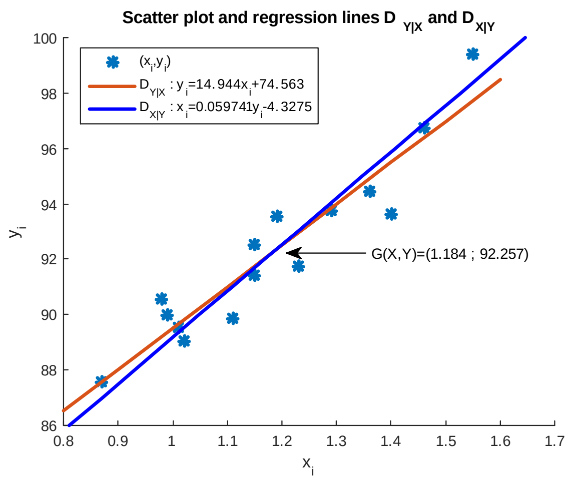
$$\sum (x_i - \overline{X})^2 = 0.54396 \quad , \quad \sum (y_i - \overline{Y})^2 = 136.0719333 \quad , \quad \sum (x_i - \overline{X})(y_i - \overline{Y}) = 8.1291$$

Pearson's coefficient of correlation is given by

$$\rho(X,Y) = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum (x_i - \overline{X})^2 \sum (y_i - \overline{Y})^2}} = \frac{8.1291}{\sqrt{0.54396 \times 136.0719333}} = 0.944\,88$$

which indicate a very strong positive linear correlation between the percentage of hydrocarbons and the purity of oxygen produced.

2. Draw the scatter plot between the two variables. Which model is the most appropriate to fit our data.



According to the scatter plot and pearson's coefficient of correlation, the suggested model will be the linear equations $Y = aX + b$ or $X = a'Y + b'$

3. We aim to explain in both sens the relationship between Y and X. Let

$$Y = aX + b + \varepsilon \qquad \text{and} \qquad X = a'Y + b' + \varepsilon'.$$

Give the values of the parameters a, a', b and b' for which $\sum \varepsilon^2$ and $\sum (\varepsilon')^2$ are minimals.

$$a = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sum (x_i - \overline{X})^2} = \frac{8.1291}{0.54396} = 14.944$$

and

$$b = \overline{Y} - a\overline{X} = 92.257 - 14.944 \times 1.184 = 74.563.$$

Thus, the regression line $D_{Y|X}$ will be $y_i = 14.944 x_i + 74.563$. Also

$$a' = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sum (y_i - \overline{Y})^2} = \frac{8.1291}{136.0719333} = 0.059741$$

and

$$b' = \overline{X} - a\overline{Y} = 1.184 - 5.974\,1 \times 10^{-2} \times 92.257 = -4.327\,5.$$

The regression line $D_{X|Y}$ will be $x_i = 0.059741 y_i - 4.327\,5$.

4. Plot in the above figure the two lines $D_{Y|X}$ and $D_{Y|X}$ represented by the two equations above. What are the coordinates of the point of intersection between the two lines and what does it represent?

the intersection point between the two lines $D_{Y|X}$ and $D_{Y|X}$ is $G(\overline{X}, \overline{Y}) = (1.184; 92.257)$ and it represents the center of the cloud of the scatter plot also known as the average/center of gravity

5. Compute the residual variance and the variance explained by the regression line $D_{Y|X}$. Do you think that the regression line fit well the data, measure it.

We call residual variance of Y or residual variance of $D_{Y/X}$, and we note $V_R(Y)$, the following expression

$$V_R(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = 0.972\,48$$

We call variance explained by the regression line $D_{Y/X}$, and we note $V_E(Y)$, the expression defined by

$$V_E(Y) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - \overline{Y})^2 = 8.099.$$

We know that $Var(Y) = V_R(Y) + V_E(Y)$ which gives that

$$V_E(Y) = \frac{Cov(X,Y)^2}{Var(X)} = Var(Y)\rho(X,Y)^2 = 9.071462 \times 0.944\,88^2 = 8.099$$

$$V_R(Y) = Var(Y)(1 - \rho(X,Y)^2) = 9.071462 \times (1 - 0.944\,88^2) = 0.972\,48.$$

Both regression lines fit well the cloud of points in the above plot, to measure the goodness of fit, we need to compute the coefficient of determination

$$R^2 = \frac{V_E(Y)}{Var(Y)} = \frac{8.099}{9.071462} = 0.893$$

which is close to 1.

6. According to you, what is the level of hydrocarbon required to achieve 100% oxygen purity and what is the least percentage of oxygen purity at a hydrocarbon level of 0.8.

To achieve 100% oxygen purity, the estimated level of hydrocarbon required is given by $D_{X|Y} : x_{100\%} = 0.059741 \times 100 - 4.327\,5 = 1.646\,6$ and at a hydrocarbon level of 0.8, the minimum percentage of oxygen purity is obtained by $D_{Y|X} : y_{0.8} = 14.944 \times 0.8 + 74.563 = 86.518\%$.

2024-2025
24-11-2024

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
National Higher School of Mathematics

NATIONAL HIGHER SCHOOL OF MATHEMATICS
المدرسة الوطنية العليا في الرياضيات

Midterm exam of Statistics & Probability (2h)

Name : Groupe:

Exercice 1. Among these assertions, specify which ones are true and which ones are false.

	True	False
1-The geometric mean is always less than or equal to the arithmetic mean.	$\surd$	
2-The range is a measure of central tendency.		$\surd$
3-The median is not affected by the presence of extreme values in a dataset.	$\surd$	
4-The STD of a sample can be -1.		$\surd$
5-The sum of deviations of data points from the mean is always zero	$\surd$	
6-The mode can be used for both numerical and categorical data	$\surd$	
7-The mean is equal to 0 for a symmetric histogram.		$\surd$
8-If a value is higher than the mean then it is an outlier.		$\surd$
9-Percentiles divide data into 100 equal parts	$\surd$	
10-The sample mean is always higher than the population mean.		$\surd$
11-$\rho(X,Y)$ is a unitless number between -1 and 1.	$\surd$	
12-The coefficient of skewness is zero for a perfectly symmetric distribution.	$\surd$	
13-Parameter is a numerical measure that describes a variable of a population.	$\surd$	
14-The harmonic mean is most appropriate for datasets with large outliers.	$\surd$	
15-Chebyshev's theorem applies to any set of data.	$\surd$	
16-Whe $\rho(X,Y)$ is close to zero, it indicates a weak linear relationship.	$\surd$	
17-A correlation of 1 or -1 means that the two variables are perfectly linearly related.	$\surd$	
18-A Box-plot can detect the outliers.	$\surd$	
19-The interquartile range is the difference between the first and second quartiles.		$\surd$
20-$\forall a; b; x_0; y_0 \in \mathbb{R} : Cov(aX + x_0; bY + y_0) \neq abCov(X;Y)$		$\surd$

Exercice 2. A teacher has collected exam scores from a class of 40 students and grouped them into classes. The data is as follows:

Class (Score)	Mid-point (x_i)	frequency (f_i)	Rel. Freq.	Cum. Freq.	Cum. Rel. Freq.
$50-59$	$\frac{50+59}{2} = 54.5$	5	$\frac{5}{40} = 0.125$	5	0.125
$60-69$	64.5	$0.2 \times 40 = 8$	0.2	13	0.325
$70-79$	74.5	12	0.3	25	0.625
$80-89$	84.5	10	0.25	35	0.875
$90-99$	94.5	5	0.125	40	1
$\sum$		**40**			

1. Complete the statistical table above.

2. Calculate the geometric, Harmonic and arithmetic mean score then compare them.

Geometric mean score :

$$GM = \sqrt[\sum_{i=1}^{5} f_i]{\prod_{i=1}^{5} x_i^{f_i}} = \sqrt[40]{54.5^5 \times 64.5^8 \times 74.5^{12} \times 84.5^{10} \times 94.5^5} = \mathbf{74.003}$$

Harmonic mean score :

$$HM = \frac{\sum_{i=1}^{5} f_i}{\sum_{i=1}^{5} \frac{f_i}{x_i}} = \frac{40}{\frac{5}{54.5} + \frac{8}{64.5} + \frac{12}{74.5} + \frac{10}{84.5} + \frac{5}{94.5}} = \mathbf{72.979}$$

Arithmetic mean score :

$$AM = \overline{X} = \frac{1}{\sum_{i=1}^{5} f_i} \sum_{i=1}^{5} f_i x_i = \frac{5 \times 54.5 + 8 \times 64.5 + 12 \times 74.5 + 10 \times 84.5 + 5 \times 94.5}{40} = \mathbf{75}$$

and we have

$$HM \leq GM \leq AM$$

3. Identify the modal class and determine the Mode

The highest frequency is for the third class $70 - 79$ to which the mode $\widehat{X}$ belongs and it's given by

$$\widehat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times w = 70 + \left(\frac{(12-8)}{(12-8)+(12-10)} \right) \times 10 = \mathbf{76.667}$$

4. Compute the variance and standard deviation of the exam scores.

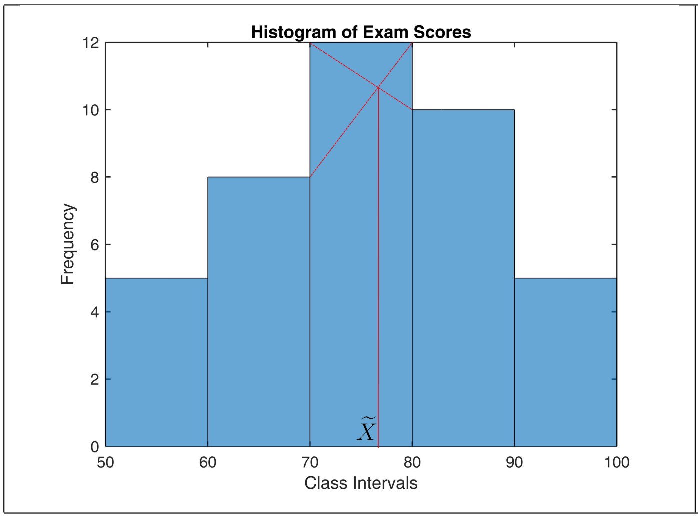
We have $\mu = \overline{X}$ and the variance is given by

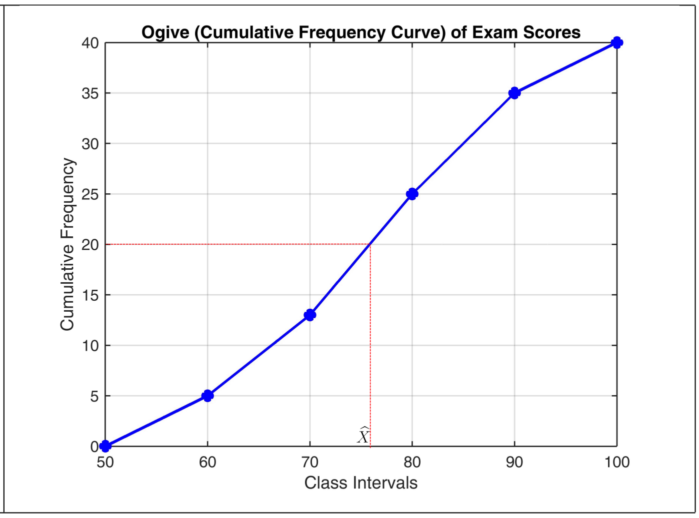
$$\sigma^2 = Var(X) = \frac{1}{\sum_{i=1}^{N} f_i} \sum_{i=1}^{N} f_i (x_i - \mu)^2 = \frac{5790}{40} = \mathbf{144.75}$$

therefore, the standard deviation of the exam scores is

$$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{N} f_i} \sum_{i=1}^{N} f_i (x_i - \mu)^2} = \sqrt{144.75} = \mathbf{12.0312.}$$

5. Construct a histogram to visualize the frequency distribution of the exam scores and the Ogive plot of the above distribution then determine the mode and median of the series graphically.





6. Determine the 10th, 25th, 50th, 75th, and 90th percentiles of the exam scores.

The 10th percentiles is also the 1st decile, it's located in the class $50-59$ and given by

$$P_{10} = D_1 = L_1 + \left(\frac{\frac{n}{10} - Cf_{d_1}}{f_{d_1}}\right) \times w = 50 + \left(\frac{\frac{40}{10} - 0}{5}\right) \times 10 = \mathbf{58}$$

and the 25th, 50th and 75th percentiles are also the 1st quartile, the median and the third quartile respectively , they are respectively located in the classes $60-69$, $70-79$ and $80-89$. They're given by

$$P_{25} = Q_1 = L_1 + \left(\frac{\frac{n}{4} - Cf_{q_1}}{f_{q_1}}\right) \times w = 60 + \left(\frac{\frac{40}{4} - 5}{8}\right) \times 10 = \mathbf{66.25}.$$

$$P_{50} = Q_2 = \widetilde{X} = L_1 + \left(\frac{\frac{n}{2} - Cf_{q2}}{f_{q2}}\right) \times w = 70 + \left(\frac{\frac{40}{2} - 13}{12}\right) \times 10 = \mathbf{75.833}.$$

$$P_{75} = Q_3 = L_1 + \left(\frac{\frac{3n}{4} - Cf_{q3}}{f_{q3}}\right) \times w = 80 + \left(\frac{\frac{3\times 40}{4} - 25}{10}\right) \times 10 = \mathbf{85}.$$

The 90th percentiles belongs to the class $90-99$ and given by

$$P_{90} = D_9 = L_1 + \left(\frac{\frac{9n}{10} - Cf_{d_9}}{f_{d_9}}\right) \times w = 90 + \left(\frac{\frac{9\times 40}{10} - 35}{5}\right) \times 10 = \mathbf{92}$$

7. Determine whether the distribution of exam scores is positively skewed, negatively skewed, or symmetrical.

By computing pearson's coefficient of skewness

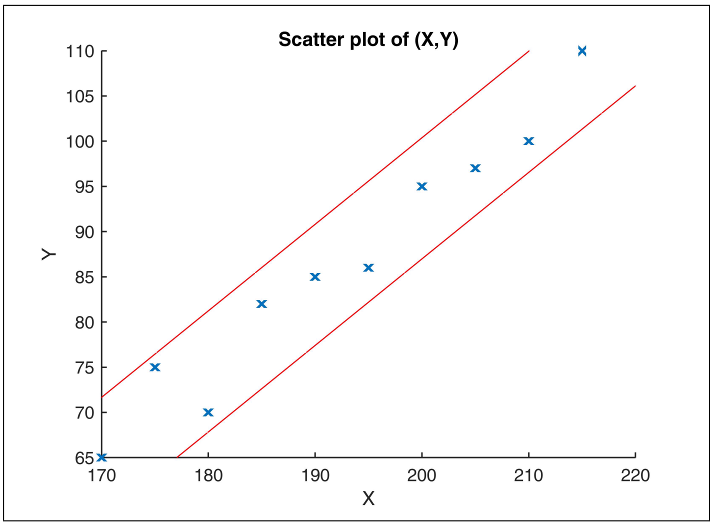
$$sk = \frac{3\left(\overline{X} - \widetilde{X}\right)}{s} = \frac{3 \times (75 - 75.833)}{12.0312} = \mathbf{-0.20771}$$

we can observe that its value is close to 0 which let us conclude that the distribution of the exam scores is almost symmetrical

8. Consider the following results:

X	170	175	180	185	190	195	200	205	210	215
Y	65	75	70	82	85	86	95	97	100	110

a. Draw the scatter plot of the given data. Describe the relationship between X and Y.



the plotted points fall in a narrow band and show a rising tendency from the lower left hand corner to the upper right hand corner, therefore, it suggests a **High Degree of Positive Linear Correlation**

b. Calculate the correlation coefficient and interpret the obtained result.

The above scatter plot suggested the presence of linear correlation, hence, we choose pearson's coefficient of correlation to measure it.
We have
$\sum x_i = 1925, \sum y_i = 865, \sum(x_i - \overline{X})^2 = 2062.5, \sum(y_i - \overline{Y})^2 = 1806.5, \sum(x_i - \overline{X})(y_i - \overline{Y}) = 1887.5.$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y} = \frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum(x_i - \overline{X})^2}\sqrt{\sum(y_i - \overline{Y})^2}} = \frac{1887.5}{\sqrt{2062.5} \times \sqrt{1806.5}} = \mathbf{0.97785}$$

which indicates the existence of a strong positive linear correlation between the two variables.

Exercice 3. 1- Set

$$Z = X - \frac{Cov(X,Y)}{\sigma_Y^2} Y$$

where X and Y are two correlated variables, σ_X^2 and σ_Y^2 are their variances.
Using Z, show that

$$|Cov(X,Y)| \leq \sigma_X \sigma_Y$$

By definition, we have

$$\begin{aligned}
0 \leq Var(Z) = Cov\,(Z,Z) &= Cov\left(X - \frac{Cov(X,Y)}{\sigma_Y^2}Y, X - \frac{Cov(X,Y)}{\sigma_Y^2}Y\right) \\
&= \underbrace{Cov\,(X,X)}_{\sigma_X^2} - 2\frac{Cov(X,Y)}{\sigma_Y^2}Cov(X,Y) + \frac{Cov(X,Y)^2}{\left(\sigma_Y^2\right)^2}\underbrace{Cov\,(Y,Y)}_{\sigma_Y^2} \\
&= \sigma_X^2 - 2\frac{Cov(X,Y)^2}{\sigma_Y^2} + \frac{Cov(X,Y)^2}{\left(\sigma_Y^2\right)^2}\sigma_Y^2 = \sigma_X^2 - \frac{Cov(X,Y)^2}{\sigma_Y^2} \geq 0 \\
&\Longrightarrow \frac{Cov(X,Y)^2}{\sigma_Y^2} \leq \sigma_X^2 \Longrightarrow Cov(X,Y)^2 \leq \sigma_X^2\sigma_Y^2 \\
&\Longrightarrow |Cov(X,Y)| \leq \sigma_X\sigma_Y. \qquad Q.E.D.
\end{aligned}$$

2- Let's consider a set of observations $x_1, x_2, ...x_n$. Define a function $S\,(m)$ that represents the sum of absolute deviations from a value m

$$S\,(m) = \sum_{i=1}^{n} |x_i - m|$$

Show that $S\,(m)$ is minimized when m is the median of the dataset.

without loss of generality, suppose that $x_1 < x_2 < ... < x_n$. If $m < x_1$, then

$$S(m) = \sum_{i=1}^{n} |x_i - m| = \sum_{i=1}^{n} (x_i - m)\,. \tag{1}$$

As m increases, each term of (1) decreases until m reaches x_1, therefore $S(x_1) < S(m)$ for all $m < x_1$.
Now suppose that $x_k \leq m \leq m + d \leq x_{k+1}$. Then

$$\begin{aligned}
S\,(m+d) &= \sum_{i=1}^{k} (m + d - x_i) + \sum_{i=k+1}^{n} (x_i - m - d) = dk + \sum_{i=1}^{k} (m - x_i) - (n-k)\,d + \sum_{i=k+1}^{n} (x_i - m) \\
&= d\,(2k - n) + \sum_{i=1}^{n} |x_i - m| = d\,(2k-n) + S\,(m)
\end{aligned}$$

so $S(m+d) - S(m) = d(2k-n)$. This is negative if $k < \frac{n}{2}$, zero if $k = \frac{n}{2}$, and positive if $k > \frac{n}{2}$. Thus, on the interval $[x_k, x_{k+1}]$

$$S(m) \text{ is } \begin{cases} \text{decreasing} & \text{if } k < \frac{n}{2} \\ \text{constant} & \text{if } k = \frac{n}{2} \\ \text{increasing} & \text{if } k > \frac{n}{2} \end{cases}$$

From here, it shouldn't be too hard to show that $S(m)$ is minimal when m is the median of the dataset $\{x_1, x_2, ...x_n\}$.

5.2 Final Exam

Final Exam

2022/2023

2023/2024

People's Democratic Republic of Algeria

2022-2023 Ministry of Higher Education and Scientific Research

National Higher School of Mathematics

NATIONAL HIGHER SCHOOL OF MATHEMATICS

المدرسة الوطنية العليا في الرياضيات

Final exam (2^h30)

Name :	Groupe:

Exercice 1. Among these assertions, specify which ones are true and which ones are false.

	True	False
1-The median is always higher than the mean.		$\surd$
2- If $P(A \cap B) \geq 0.10$ then $P(A) \geq 0.10$.	$\surd$	
3- If $P(A\|B) = 0.5$ and $P(B) = 0.5$, then A and B are necessarily independent.		$\surd$
4- If a value is higher than the mean then it is an outlier.		$\surd$
5- $P(A) - P(B) \neq P(A \cap B^c)$.		$\surd$
6- Multiplying all values by 2, will multiply the STD by 2.	$\surd$	
7- Removing an outlier that is negative, will most likely decrease the STD.	$\surd$	
8- If A and B are mutually exclusive events, then $P(A\|B) = P(A)$.		$\surd$
9- The mean is equal to 0 for a symmetric histogram.		$\surd$
10- $E[X + 2Y] = E[X] + 2E[Y]$.	$\surd$	
11- The probability is always non negative.	$\surd$	
12- The STD of a sample is always less than 1.		$\surd$
13- Two r.v X and Y are identically distributed if $\exists x \in \mathbb{R}, F_X(x) = F_Y(x)$.		$\surd$
14- The median is always higher than the mean.		$\surd$
15- If $A \cap B \neq \emptyset$ and the events B and C are disjoint. then $A \cap B \cap C^c = \emptyset$		$\surd$
16- If $P(A\|B) = 0.6$ and $P(B) = 0.2$ then $P(A \cap B) = 0.12$.	$\surd$	
17- $P(A_1 \cup A_2 \cup ...) \leq P(A_1) + P(A_2) + ...$	$\surd$	
18- The Coefficient of determination R^2 is a unitless number between -1 and 1.		$\surd$
19- $\forall a, b, c \in \mathbb{R}, Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X;Y)$.	$\surd$	
20- We can learn the IQR by looking at the boxplot.	$\surd$	

Exercice 2. For a sample of 15 students we observe the Time (in minutes, X) usually spent using Facebook per day, and the final grade in the Statistics exam (Y):

x_i	0	11	17	16	22	17	25	30	27	27	31	35	30	45	60
y_i	25	22	28	30	22	27	26	21	27	28	23	29	30	24	27

1. Compute mean, median and quartiles for both X and Y

$$\overline{X} = \frac{1}{15}\sum_{i=1}^{15} x_i = \frac{393}{15} = 26.2 \text{ and } \overline{Y} = \frac{1}{15}\sum_{i=1}^{15} y_i = \frac{389}{15} = 25.933.$$

There are 15 (odd) observation, the medians $\widetilde{X}$ and $\widetilde{Y}$ are the center of observation in the ordered lists. The location of the medians are the $\left(\frac{15+1}{2}\right)^{th}$ item :

$$\widetilde{X} = x_{(8)} = 27 \text{ and } \widetilde{Y} = y_{(8)} = 27.$$

Also, Q_1 and Q_3, are the $\left(\frac{15+1}{4}\right)^{th}$ and $\left(\frac{3\times(15+1)}{4}\right)^{th}$ items in the ordered lists :

$$Q_1^X = x_{(4)} = 17, Q_3^X = x_{(12)} = 31 \quad \text{and} \quad Q_1^Y = y_{(4)} = 23, Q_3^Y = y_{(12)} = 28.$$

2. Find the standard deviation and the coefficient of variation for both X and Y. Which variable is more scattered?

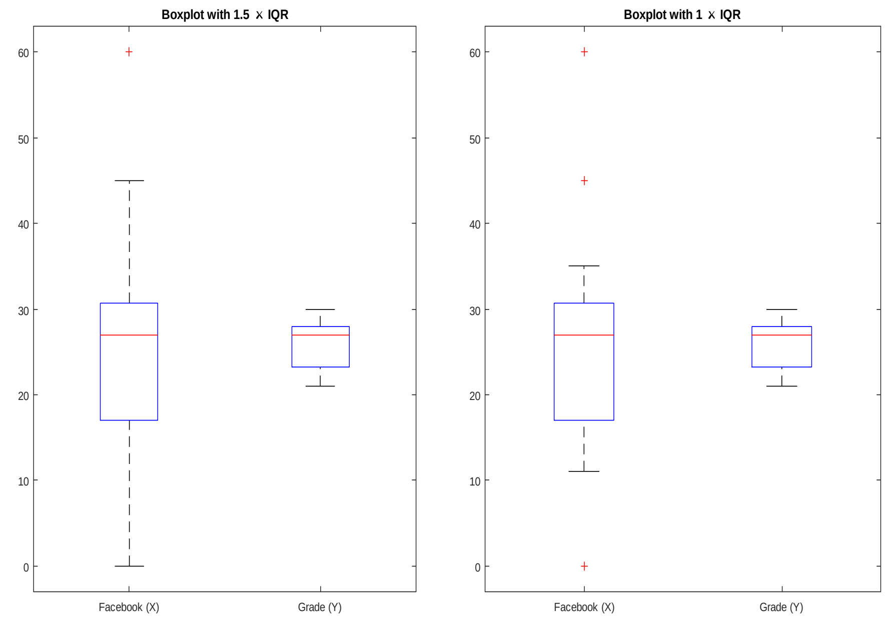
$$s_X = \sqrt{\frac{1}{15-1}\sum_{i=1}^{15}\left(x_i - \overline{X}\right)^2} = \sqrt{\frac{2836.4}{14}} = 14.234$$

and

$$s_Y = \sqrt{\frac{1}{15-1}\sum_{i=1}^{15}\left(y_i - \overline{Y}\right)^2} = \sqrt{\frac{122.93}{14}} = 2.963$$

$CV_X = \frac{s_X}{\overline{X}} \times 100\% = 54.328\%$ and $CV_Y = \frac{s_Y}{\overline{Y}} \times 100\% = 11.426\%$.
Thus the time spent on Facebook (X) is more heterogeneous than the Statistics grade (Y).

3. Plot the box-plot for both X and Y and comment them.

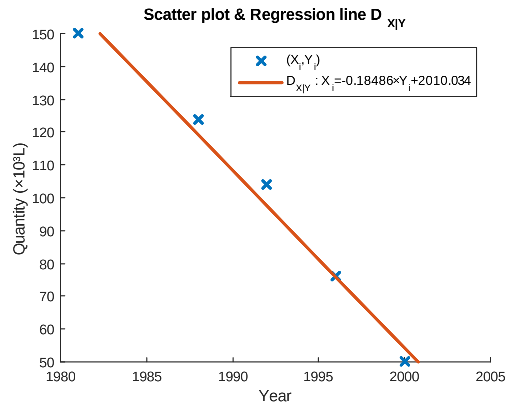


Both boxplots show an asymmetrical character with presence of outliers for the time spent on Facebook (X)

Exercice 3. Since clean-air regulations have dictated the use of unleaded gasoline, the supply of leaded gas has diminished. The table below was given

Year (X)	1981	1988	1992	1996	2000
Quantity $(\times 10^3 L)$ (Y)	150	124	104	76	50

1. Draw the scatter plot between the two variables and measure the correlation between them.



$$\overline{X} = 1991.4, \overline{Y} = 100.8,$$

$$Cov(X,Y) = \frac{1}{5}\sum\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = -227.92$$

$$Var(X) = \frac{1}{5}\sum\left(x_i - \overline{X}\right)^2 = 43.04$$

$$\sigma_X = \sqrt{43.04} = 6.56$$

$$Var(Y) = \frac{1}{5}\sum\left(y_i - \overline{Y}\right)^2 = 1232.96(\times 10^3 L)^2$$

$$\sigma_Y = \sqrt{1232.9} = 35.113(\times 10^3 L)$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-227.92}{6.56 \times 35.113} = -0.98949$$

2. Create a linear model to predict the quantity that would be available in 2005.

To predict the quantity that would be available in 2005, we need first to fit the line

$$Y = aX + b + \varepsilon.$$

therefore, $\widehat{a} = \frac{Cov(X,Y)}{Var(X)} = \frac{-227.92}{43.04} = -5.2955$ and $\widehat{b} = \overline{Y} - \widehat{a}\overline{X} = 100.8 - (-5.2955) \times 1991.4 = 10646.26$. Thus

$$Y = -5.2955X + 10646.26 + \varepsilon$$

and

$$Y_{2005} = -5.2955 \times 2005 + 10646.26 = 28.783(\times 10^3 L)$$

3. Create a model that would estimate the year when leaded gasoline will first become unavailable then draw this line on the above scatter plot.

The year when leaded gasoline will first become unavailable is the year with $Y = 0$ in the model

$$X = a'Y + b' + \varepsilon'.$$

with $\widehat{a'} = \frac{Cov(X,Y)}{Var(Y)} = \frac{-227.92}{1232.96} = -0.18486$ and $\widehat{b'} = \overline{X} - \widehat{a'}\overline{Y} = 1991.4 - (-0.18486) \times 100.8 = 2010.034$. Thus

$$X_0 = -0.18486 \times (0) + 2010.034 \simeq 2010.$$

4. Give the value of R^2 for the line of the 3^{rd} question then briefly interpret its value.

$$R^2 = \frac{V_E(X)}{Var(X)} = \frac{Var(X)\rho(X,Y)^2}{Var(X)} = (-0.98949)^2 = 0.97909$$

The coefficient of determination R^2 summarizes the part of the variance explained by the regression lines. It is a unitless number, between 0 and 1.The regression lines here explain 97.91% of the variance (of X and Y) which gives that the goodness of fit is very good.

Exercice 4. Consider a given population, a certain disease D and a certain symptom S. We know that 85% of the population members who have the disease D do have the symptom S, while the remaining 15% do not. Suppose that 95% of those who do not have the disease D do not have the symptom S, while 5% do have the symptom S (due to some other cause).
Suppose that 10% of the given population have the disease.

1. What is the probability that a random person chosen from the population has the symptom?

We have

$$P(D) = 0.1, P(D^c) = 0.9$$
$$P(S|D) = 0.85, P(S^c|D) = 0.15$$
$$P(S^c|D^c) = 0.95, P(S|D^c) = 0.05$$

and we're looking for $P(S)$ which can be obtained by the **Law of Total Probability**

$$P(S) = P(S|D)P(D) + P(S|D^c)P(D^c) = 0.85 \times 0.1 + 0.05 \times 0.9 = 0.13$$

2. Given that a random person was sampled, and he has the symptom, what is the probability that he or she has disease?

Since we already know that he has the symptom, the probability that he or she has disease is given by **Bayes' Theorem**

$$P(D|S) = \frac{P(D \cap S)}{P(S)} = \frac{P(S|D)P(D)}{P(S)} = \frac{0.85 \times 0.1}{0.13} = 0.65$$

3. What is the probability that a random person does not have the disease and does not have the symptom?

$$P(D^c \cap S^c) = P(S^c|D^c)P(D^c) = 0.95 \times 0.9 = 0.855$$

2023-2024
28-01-2024

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
National Higher School of Mathematics

NATIONAL HIGHER SCHOOL OF MATHEMATICS
المدرسة الوطنية العليا في الرياضيات

Final exam (2^h30)

Name : Groupe:

Exercice 1. A household survey of gas consumption yielded the following results

Gas consumption	Frequency (f_i)	med-class (x_i)	Cum. Freq.	Rel.Cum. Freq.
0 – 9	1	**4.5**	**1**	**0.01**
10 – 19	2	**14.5**	**3**	**0.03**
20 – 29	1	**24.5**	**4**	**0.04**
30 – 39	5	**34.5**	**9**	**0.09**
40 – 49	8	**44.5**	**17**	**0.17**
50 – 59	16	**54.5**	**33**	**0.33**
60 – 69	19	**64.5**	**52**	**0.52**
70 – 79	20	**74.5**	**72**	**0.72**
80 – 89	17	**84.5**	**89**	**0.89**
90 – 99	11	**94.5**	**100**	**1**

1. Complete the above statistical table.

2. What is the average gas consumption of these household ? Calculate the variance, standard deviation and coefficient of variation.

Since the data are grouped, we have :

$$\mu = \overline{X} = \frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} = \frac{6650}{100} = \mathbf{66.5}$$

$$\sigma^2 = Var(X) = \frac{1}{\sum_{i=1}^{N} f_i - 1} \sum_{i=1}^{N} f_i \left(x_i - \mu\right)^2 = \frac{37800}{99} = \mathbf{381.82}$$

and

$$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{N} f_i - 1} \sum_{i=1}^{N} f_i \left(x_i - \mu\right)^2} = \sqrt{381.82} = \mathbf{19.54}.$$

The coefficient of variation is given by

$$CV = \frac{\sigma}{\mu} \times 100 = CV = \frac{19.54}{66.5} \times 100 = \mathbf{29.383\%}$$

3. Give the Q_1, Q_3, D_5 and P_5.

The first quartile Q_1 is located in the class $50 - 59$ and the third quartle Q_3 is in $80 - 89$ and both are given by

$$Q_1 = L_1 + \left(\frac{\frac{n}{4} - Cf_{q_1}}{f_{q_1}}\right) \times w = 50 + \left(\frac{\frac{100}{4} - 17}{16}\right) \times 10 = \mathbf{55}$$

and

$$Q_3 = L_1 + \left(\frac{\frac{3n}{4} - Cf_{q3}}{f_{q3}}\right) \times w = 80 + \left(\frac{\frac{300}{4} - 72}{17}\right) \times 10 = \mathbf{81.765}.$$

Also, D_5 and P_5 are located in $60-69$ and $30-39$ respectively, and both are given by

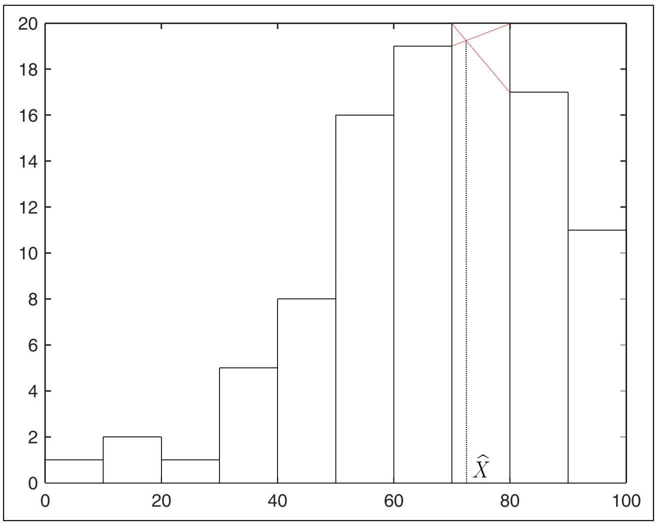
$$D_5 = L_1 + \left(\frac{\frac{5n}{10} - Cf_{d_5}}{f_{d_5}}\right) \times w = 60 + \left(\frac{\frac{500}{10} - 33}{19}\right) \times 10 = \mathbf{68.947}$$

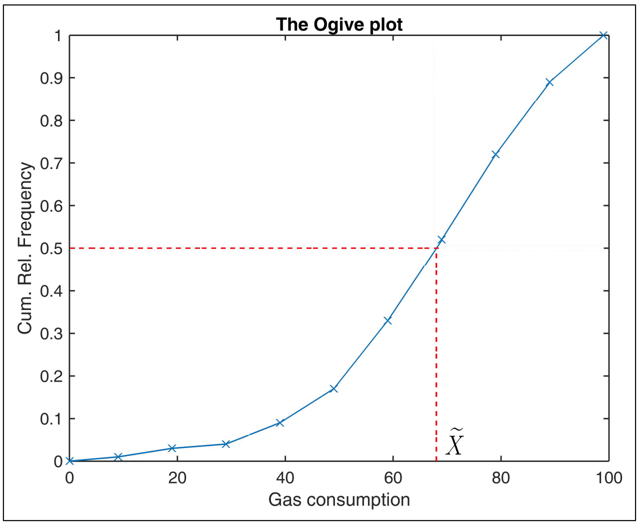
and

$$P_5 = L_1 + \left(\frac{\frac{5n}{100} - Cf_{p_5}}{f_{p_5}}\right) \times w = 30 + \left(\frac{\frac{500}{100} - 4}{5}\right) \times 10 = \mathbf{32}.$$

Notice that $D_5 = Q_2$ =the Median $\widetilde{X}$

4. Plot the histogram and the Ogive plot of the above distribution then determine the mode and median of the series graphically





5. Give the modale class then calculate the mode value.

Clearly, The highest frequency is for the class $70-79$ to which the mode $\widehat{X}$ belongs and it's given by

$$\widehat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times w = 70 + \left(\frac{(20-19)}{(20-19)+(20-17)}\right) \times 10 = \mathbf{72.5}$$

Exercice 2. A statistician carries out a study on 200 executives of a company and seeks to study the link between the age of the executives and the monthly salary (in thousands of DA) that they receive. He presents his results in a contingency table, the statistical variable X representing age and the statistical variable Y monthly salary

$X \backslash Y$	$[18,22[$	$[22,26[$	$[26,30[$	$[30,34[$	$[34,38[$	$\mathbf{n}_{i.}$
$[20,28[$	12	5	4	3	0	**24**
$[28,36[$	7	11	9	4	1	**32**
$[36,44[$	4	9	10	9	8	**40**
$[44,52[$	1	6	12	11	15	**45**
$[52,60[$	1	4	15	17	22	**59**
$\mathbf{n}_{.j}$	**25**	**35**	**50**	**44**	**46**	**200**

1. Give the percentage of executives whose salary is greater than or equal to 26,000 DA.

$$\frac{\mathbf{n}_{.3} + \mathbf{n}_{.4} + \mathbf{n}_{.5}}{N} \times 100\% = \frac{50+44+46}{200} \times 100\% = \mathbf{70\%}$$

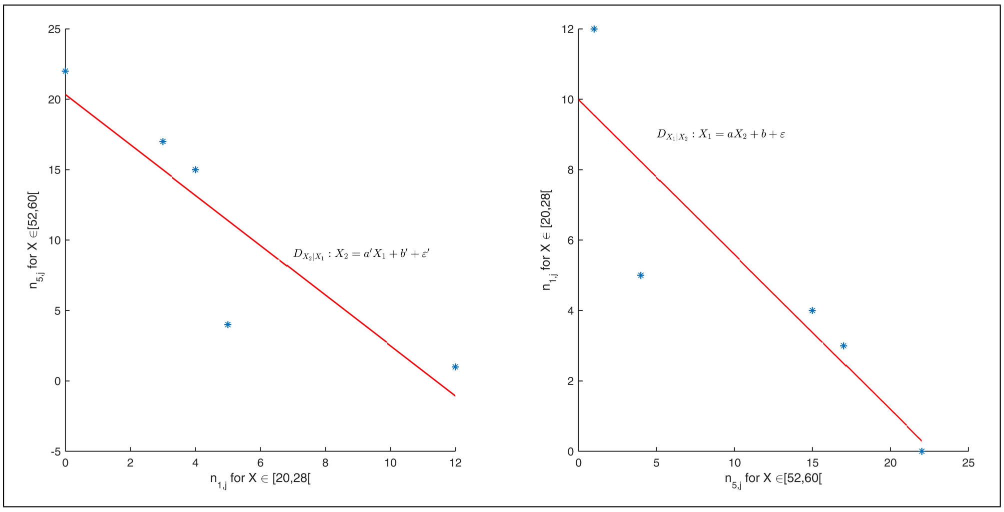
2. Among executives aged at least 36, give the percentage of those who receive less than 30,000 DA.

$$\frac{\sum_{i=3}^{5}\sum_{j=1}^{3} n_{ij}}{N} \times 100\% = \frac{4+9+10+1+6+12+1+4+15}{200} \times 100\% = \mathbf{31\%}$$

3. Give the conditional distribution of the age variable knowing that the salary received is in class [26,30[.

$X \mid Y \in [26,30[$	$[20,28[$	$[28,36[$	$[36,44[$	$[44,52[$	$[52,60[$
$Frequency$	4	9	10	12	15

4. Plot the scatter plot between age class $[20,28[$ and age class $[52,60[$. Comment.



The scatter plot shows a relatively strong negative linear relation between the two classes

5. Calculate the correlation coefficient and give the regression line between the two classes then plot it and give the value of the coefficient of determination.

$X_1 \equiv n_{ij}$ of $X \in [20,28[$	12	5	4	3	0
$X_2 \equiv n_{ij}$ of $X \in [52,60[$	1	4	15	17	22

The above scatter plot suggested the presence of linear correlation, hence, we choose pearson's coefficient of correlation to measure it.
We have
$\sum x_{1i} = 24, \sum x_{2i} = 59, \sum(x_{1i} - \overline{X_1})^2 = 78.8, \sum(x_{2i} - \overline{X_2})^2 = 318.8, \sum(x_{1i} - \overline{X_1})(x_{2i} - \overline{X_2}) = -140.2.$

$$\rho(X_1, X_2) = \rho(X_2, X_1) = \frac{Cov(X_1, X_2)}{\sigma_{X_1}.\sigma_{X_2}} = \frac{\sum(x_{1i} - \overline{X_1})(x_{2i} - \overline{X_2})}{\sqrt{\sum(x_{1i} - \overline{X_1})^2}\sqrt{\sum(x_{2i} - \overline{X_2})^2}}$$
$$= \frac{-140.2}{\sqrt{78.8} \times \sqrt{318.8}} = \mathbf{-0.88456}$$

which indicates the existence of a strong negative linear correlation between the two variables.
Set

$$X_1 = aX_2 + b + \varepsilon \qquad \text{and} \qquad X_2 = a'X_1 + b' + \varepsilon'$$

we have

$$a = \frac{Cov(X_1, X_2)}{Var(X_2)} = \frac{\sum(x_{1i} - \overline{X_1})(x_{2i} - \overline{X_2})}{\sum(x_{2i} - \overline{X_2})^2} = \frac{-140.2}{318.8} = \mathbf{-0.43977},$$
$$b = \overline{X_1} - a\overline{X_2} = \frac{24}{5} - (-0.439\,77) \times \frac{59}{5} = \mathbf{9.9893}$$

and

$$a' = \frac{Cov(X_1, X_2)}{Var\,(X_1)} = \frac{-140.2}{78.8} = \textbf{-1.7792},$$

$$b' = \overline{X_2} - a\overline{X_1} = \frac{59}{5} - (-1.779\,) \times \frac{24}{5} = \textbf{20.339}$$

thus

$$X_1 = -0.43977X_2 + 9.9893 + \varepsilon \qquad \text{and} \qquad X_2 = -1.7792X_1 + 20.339 + \varepsilon'$$

The coefficient of determination in both cases is

$$R^2 = \frac{V_E\,(X_1)}{Var\,(X_1)} = \frac{V_E\,(X_2)}{Var\,(X_2)} = \rho\,(X_1, X_2)^2 = \textbf{0.78245}$$

Exercice 3. **I**– A diagnostic test for a disease is such that it (correctly) detects the disease in 90% of the individuals who actually have the disease. Also, if a person does not have the disease, the test will report that he or she does not have it with probability .9. Only 1% of the population has the disease in question. If a person is chosen at random from the population and the diagnostic test indicates that she has the disease

1. what is the conditional probability that she does, in fact, have the disease?

Set H : "a person has the disease" $\Longrightarrow P\,(H) = 1\% = 0.01$, and E : "the test shows a positive result". Thus $P\,(E|H) = 90\% = 0.9$ and $P\,(E|H^c) = 1\% = 0.1$.
We want to find $P\,(H|E)$, using bayes' theorem, we have

$$P\,(H|E) = \frac{P\,(E|H)\,P\,(H)}{P\,(E|H)\,P\,(H) + P\,(E|H^c)\,P\,(H^c)} = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} = \textbf{0.083333}$$

2. Are you surprised by the answer? Would you call this diagnostic test reliable?

surprisingly, the test does not appear to be reliable, because even with a positive result, the probability of having the disease is negligible

II– X and Y are independent, discrete random variables whose probability functions are given in the tables below:

x	1	2	3
$P\,(X = x)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

y	1	2	3
$P\,(Y = y)$	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

1. Compute the probability $P(X + Y = 4)$.

$$P(X+Y=4) = P\,(X=1, Y=3) + P\,(X=2, Y=2) + P\,(X=3, Y=1)$$
$$\overset{X \text{ and } Y \text{ are indep.}}{=} P\,(X=1)\,P\,(Y=3) + P\,(X=2)\,P\,(Y=2) + P\,(X=3)\,P\,(Y=1)$$
$$= \frac{1}{3}\frac{1}{6} + \frac{1}{2}\frac{1}{6} + \frac{1}{6}\frac{2}{3} = \textbf{0.25}$$

2. Compute the conditional probability $P(X \leq 2|X + Y = 4)$

$$P(X \leq 2|X+Y=4) = \frac{P(X \leq 2, X+Y=4)}{P(X+Y=4)} = \frac{P\,(X=1, Y=3) + P\,(X=2, Y=2)}{P(X+Y=4)}$$
$$= \frac{\frac{1}{3}\frac{1}{6} + \frac{1}{2}\frac{1}{6}}{0.25} = \textbf{0.556}$$

3. Give the expectation of $g(X,Y) = X^2Y$.

Since X and Y are independent, we have $E(g(X,Y)) = E(X^2Y) = E(X^2)E(Y)$ with

$$E\left(X^2\right) = 1^2 \times \frac{1}{3} + 2^2 \times \frac{1}{2} + 3^2 \times \frac{1}{6} = \mathbf{3.8333}$$
$$E(Y) = 1 \times \frac{2}{3} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} = \mathbf{1.5}$$

Therefore, $E(g(X,Y)) = \mathbf{5.75}$

4. Give the CDF of the $\min(X,Y)$

Set $Z = \min(X,Y)$, then, the CDF of Z is given by

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = 1 - P(Z > z) \\ &= 1 - P(\min(X,Y) > z) \\ &= 1 - P(X > z \text{ and } Y > z) \\ &= 1 - P(X > z)P(Y > z) \\ &= 1 - (1 - P(X \leq z))(1 - P(Y \leq z)) \\ &= 1 - (1 - F_X(z))(1 - F_Y(z)) \end{aligned}$$

Bibliography

[1] J.-R. Barra. *Notions fondamentales de statistiques mathématiques.* Dunod, 1971.

[2] P. Bickel and K. Doksum. *Mathematical statistics.* Prentice Hall, 1977.

[3] A. Borovkov. *Mathematical statistics.* Gordon and Breach Science Pub., 1999.

[4] N. Bouleau. *Probabilités de l'ingénieur : Variables aléatoires et simulation.* Actualités Scientifiques et Industrielles, 1986.

[5] P. Brémaud. *Introduction aux probabilités.* Springer Verlag, 1984.

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

[7] W. Feller. *An introduction to probability theory and its applications*, vol. 1. Wiley and Sons, 3rd ed., 1968.

[8] G. Grimmett and D. Stirzaker. *Probability and random processes.* Oxford University Press, 2nd ed., 1992.

[9] B. Jourdain. *Probabilités et statistique.* Ellipses, 2009.

[10] A. Kessira. *Cours de statistique et introduction aux probabilités.* École Nationale Supérieure de Mathématiques, 2022–2025.

[11] A. Monfort. *Cours de statistique mathématique.* Economica, 1982.

[12] A. Monfort. *Introduction à la statistique.* École Polytechnique, 1988.

[13] J. Neveu. *Bases mathématiques du calcul des probabilités.* Masson, 1964.

[14] J. Neveu. *Introduction aux probabilités.* École Polytechnique, 1988.

[15] W. Rudin. *Real and complex analysis*. McGraw-Hill, 3rd ed., 1986.

[16] G. Saporta. *Probabilités, analyse des données et statistique*. Technip, 2ème édition, 2006.

[17] L. Schwartz. *Analyse III : calcul intégral*. Hermann, 1993.