الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة عبد الحميد بن باديس مستغانم
Abdelhamid Ibn Badis Mostaganem University
كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم العلوم والتقنيات
Department of Science and Techniques

COURSE HANDOUT

PROBABILITY AND STATISTICS COURSE AND CORRECTED EXERCISES

Degree : 1st year Engineering (ST)

Experts :

- **Pr BELARBI (Born HAMANI) Samira**
  Faculty of Exact Sciences and Informatics – UMAB Mostaganem
- **Pr SAHRAOUI Rahma**
  Higher School of Agronomy of Mostaganem

PRESENTED BY : Dr. BERRIGHI NACERA

Academic Year 2025 - 2026

Contents

Part I

Statistics

Chapter 1

Basic Definitions

In order to understand statistics, it is essential to first learn some basic definitions. In this chapter, we introduce fundamental terms commonly used in statistics.

1.1 Population

Definition 1.1.1 *A population is the collection of all individuals (person , things or objects) under consideration in the study.*

Example 1.1.1 *All algerian citizens who are currently registered to vote.*

Example 1.1.2 *All patients treated at a particular hospital last year.*

In real-world situations, it is often impossible to gather information about an entire population. The main goal of statistics is to collect and analyze a subset of the population, known as a sample, in order to obtain information about specific characteristics of interest.

1.2 Sample

Definition 1.2.1 *A sample is a subset of individuals selected from the population.*

Remark 1.2.1 *In the best case the sample represent the population.*

Example 1.2.1 *The registered voters selected to participate in a recent survey concerning their intention to vote in the next election.*

Example 1.2.2 *The patients selected to fill out a patient-satisfaction questionnaire.*

1.3 Variable

In statistic what we examine or what we stady is variable.

Definition 1.3.1 *A variable is a characteristic under study that takes on different values for different individuals(elements).*

Examples:

1. Gender
2. The number of varieties of a brand of cereal
3. Weight and height
4. Number of students, age, Marital status and number of children in family

1.4 Modalitie

Definition 1.4.1 *A List variable can contain as many possible values as you desire. These possible values are called "modalities"*

Example 1.4.1 *"Marital status" whose possible modalities are single -divorced-married and widowed.*

Example 1.4.2 *"Age" for example, it could be recorded using methods corresponding to annual or five-year age classes depending on the desired analysis.*

1.5 Types of Variables

1.5.1 Qualitative Variables

1. Qualitative Variables (or Categorical Variables)

Definition 1.5.1 *Qualitative Variables are nonnumeric variables and can't be measured.*

Remark 1.5.1 *Qualitative variables described by a word or phrase.*

Example 1.5.1 *gender(male and female), religious affiliation, state of birth, Marital status and colour.*

1.5.2 Quantitative Variables

2.**Quantitative Variables (or Numerical Variables)**

Definition 1.5.2 *Quantitative Variables are numerical variables and can be measured.*

Remark 1.5.2 *Quantitative Variables described by a number.*

Example 1.5.2 *The number of previous presidential elections in which a citizen voted*

Example 1.5.3 *balance in checking account, number of children in family, the marks for a maths test, the number of universities in a country, the heights of a group of students, the weight of a group , ages, number of calls arriving at a telephone exchange in 5 seconds.*

Quantitative variables can be:

i) Discrete:

Definition 1.5.3 *The variable can only take one of a finite or countable number of values.*

Example 1.5.4 *the number of bedrooms in your house, number of children in family.*

ii) Continuous:

Definition 1.5.4 *The variable is a measurement which can take any value in an interval of the real line.*

Example 1.5.5 *weight, balance in checking account.*

1.6 Data

Definition 1.6.1 *Data are the different values associated with a variable. They may be numbers or they may be words.*

Remark 1.6.1 *The letters like x or y generally are used to represent data values.*

Remark 1.6.2 *A single value is called a datum.*

Notation: 1)Let the symbol x_i (read "x sub i") denote any of the N values $x_1, x_2, x_3, ..., x_N$ assumed by a variable x. The letter i in x_i, which can stand for any of the numbers $1, 2, 3.....N$ is called a subscript, or index.
2)The symbol $\sum_{i=1}^{N}$ is used to denote the sum of all the x_i's from $i = 1$ to $i = N$; by definition,

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + x_3 + ... + x_N$$

Example: The number of telephone calls per day to a person is recorded for 12 days. The resulting data set is:

2	0	3	4	1	0	5	3	0	0	1	2

Possible values for **the variable number of calls** are 0, 1, 2, 3, 4,5 ; these are isolated points on the number line, so we have a sample consisting of **discrete numerical data**.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
$x_1 = 2$	0	3	4	1	0	5	3	0	0	1	2

$$\begin{aligned}\sum_{i=1}^{12} x_i &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} \\ &= 2 + 0 + 3 + 4 + 1 + 0 + 5 + 3 + 0 + 0 + 1 + 2 = 21\end{aligned}$$

1.7 Exercises

Exercise1:
A study on the number of siblings (brothers and sisters) of first-year engineering students in the Department of Sciences and Techniques was conducted. The 20 students in a statistics class surveyed the number of siblings they each had. The results are presented in the following table.

3	5	4	1	1	2	2	4	2	1
2	0	3	3	7	1	2	5	3	3

1°Identify the following:
a) Individual(Statistical unit):
b) Population:
c) Sample:
d) Variable and modalities:
e)Type of variable:
2°Complete the table bellow:

Modalities(**number of siblings**)	the number of observations(No. of Students)
0	1
1	
2	
3	
4	
5	
6	
7	

Exercise2:
A study on the language most mastered by students in the department of Sciences and Techniques (Refer to the statistics class in Exercise1).

EN	FR	FR	EN	EN	FR	EN	EN	EN	FR
FR	FR	FR	EN	EN	FR	EN	FR	FR	FR

1°Identify the following:
a) Statistical unit:

b) Population:
c) Sample:
d) Variable:
e) Type of variable:
2°Complete the table :

Modalities	the number of observations (n_i)
EN	
FR	
Total	$\sum_i n_i = 20$

Exercise 3:
Refer to the data in Exercise1 (The information about the height of students). The students be classified according to height as given below.

class	No. of Students(n_i)
$[1.5; 1.6[$	3
$[1.6; 1.7[$	
$[1.7; 1.8[$	
$[1.8; 1.9[$	6
Total	$n = \sum n_i = 20$

1° Identify the variable and Type of variable.
2°Complete the table.

1.8 Solutions

Exercise1:
a) Individual(Statistical unit): The student.
b) Population: All students of first year engineering in the department of Sciences and Techniques .
c) Sample: 20 students and a sample size is 20.
d) Variable: The number of siblings (brothers and sisters) with modalities: $0, 1, 2, 3, 4, 5, 7$.
e)Type of variable:Quantitative variable (Discrete).

2°

Modalities(**number of siblings**)(x_i)	the number of observations (n_i)
0	1
1	4
2	5
3	5
4	2
5	2
6	0
7	1
Total	$n = 20$

Exercise2:
a)Statistical unit (individual): the student.
b)Population: students from the Sciences and Techniques department.
c)A sample size of 20 students.
d)Variable (character) studied: the language most mastered.
e) Type of variable:Qualitative Variable (or Categorical Variable).

Modalities	the number of observations(n_i)
EN	9
FR	11
Total	$\sum_i n_i = 20$

Exercise 3:
1°Variable: The height of student. Type of variable: Quantitative variable (**Continuous**).
2°

class	No. of Students(n_i)
$[1.5; 1.6[$	3
$[1.6; 1.7[$	7
$[1.7; 1.8[$	4
$[1.8; 1.9[$	6
Total	$n = \sum n_i = 20$

Chapter 2

One-Variable Statistics

This chapter explains how to organize and describe a set of data by means of tables, graphs, and calculation of some numerical summary measures. The methods consisting mainly of one variable.

2.1 Frequency

The collected data, also called raw data (or a statistical series), are initially unorganized and must be arranged and presented in a clear and comprehensible manner to facilitate further statistical analysis. This is achieved by constructing a frequency table, also known as a frequency distribution.

2.1.1 Frequency

Definition 2.1.1 *The **frequency** (or **absolute frequency**) of a value is the number of observations taking that value.*

2.1.2 Frequency Distributions

Definition 2.1.2 *A frequency distribution is the organization of raw data in table form, using classes and frequencies.*

Remark 2.1.1 *If the sample is composed of n individuals who form the population and the statistical series contains k different values $x = (x_1, x_2, ..., x_k)$, such that*

the frequency *of each value* x_i *where* $1 \leq i \leq k$ *is* ***denoted by*** n_i*, then* ***the size of sample*** $n = \sum_{i=1}^{k} n_i$.

Examples:1) Dataset of Exercise 1 Chapter 1:
1°Arrange data in ascending form:

0	1	1	1	1	2	2	2	2	2
3	3	3	3	3	4	4	5	5	7

2°

Modalities(**number of siblings**)(x_i)	**Frequency**(n_i)
0	1
1	4
2	5
3	5
4	2
5	2
6	0
7	1
Total	$\sum_i n_i = 20$

2) Dataset of Exercise 2 Chapter 1:

EN	FR	FR	EN	EN	FR	EN	EN	EN	FR
FR	FR	FR	EN	EN	FR	EN	FR	FR	FR

Modalities	**Frequency**(n_i) (n_i)
EN	9
FR	11
Total	$\sum_i n_i = 20$

2.1.3 The types of frequency distribution

There are three types of frequency distributions

1. Categorical frequency distribution:It is related to qualitative variable.

2. Ungrouped frequency distribution:It is related to quantitative variable.

3. Grouped frequency distribution:It is related to quantitative variable.

There are specific procedures for constructing each type.

Categorical frequency Distribution

The categorical frequency distribution is used for data that can be placed in specific categories.

Example 2.1.1 *Asocial worker collected the following data on marital status for* **25** *persons.*

(M=married, S=single, W=widowed, D=divorced):

M	S	D	W	D	S	S	M	M	M	W	D	S	M	W	M	S	W	D	S	S	D	D	W	D

We follow procedure to construct the frequency distribution.

Step 1: Make a table as shown.

marital status	Tally	Frequency
M		
S		
D		
W		

Step 2: Tally the data and place the result in column (2).
Step 3: Count the tally and place the result in column (3).

marital status	Tally	Frequency
M	////	5
S	//// //	7
D	//// //	7
W	////	6

Frequency table 1 Example 2.1.1

Ungrouped frequency distribution

Ungrouped frequency distribution are used for small set and when the range of values in the data set is smal and the sample size(N) is large.
Steps for constructing ungrouped frequency distribution:

1 First find the smallest and largest raw score in the collected data

2 Arrange the data in order of magnitude and count the frequency.

3 To facilitate counting one may include a column of tallies.

Example 2.1.2 *Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:*

4	3	6	8	5	3	2	2	4	5
5	5	3	6	3	5	5	2	4	3

- Let us organize it into a frequency distribution table:

1) First of all arrange the raw scores in ascending order:

2	2	2	3	3	3	3	3	4	4	4	5	5	5	5	5	5	6	6	8

2)

class(x_i)	Frequency(n_i)
2	3
3	5
4	3
5	6
6	2
8	1

Frequency table 2 Example 2.1.2

Grouped frequency distribution

When the range of the data is large, the data must be grouped in to classes that are more than one unit in width. in what is called a grouped frequency distribution.
Steps for constructing Grouped frequency Distribution :

1. Find the highest value (H) and lowest value (L) .
2. Compute the Range(R) : $R = H - L$
3. Select the number of classes(k) desired, usually between 5 and 20 .
4. Determine **the class width** $(W)$by dividing the range by the number of classes$\left(W = \dfrac{\text{Range}}{\text{Number of class}} = \dfrac{R}{k}\right)$ and rounding up. Generally, the class width should be the same size for all classes.
5. Select a starting point **less than or equal to the smallest (minimum) value**. The starting point is called the **lower limit** of the **first class**(lowest class interval). Add the width **to this lower limit** to get the upper class-limit.
6. Find the class-limits for the remaining classes.
7. Tally the items into the classes.
8. Count the number of items in each class (find the frequencies).

Example 2.1.3 *Twenty-eight* (28)*students were asked how many hours they worked per week. Their responses, in hours, are as follows:*

15	26	13	33	22	14	15	27	32	23	5	26	25	14
34	13	15	22	15	28	10	18	21	24	20	18	34	20

Construct a grouped frequency distribution using 5 classes.

- First of all arrange the raw scores in ascending order:

5	10	13	13	14	14	15	15	15	15	18	18	20	20
21	22	22	23	24	25	26	26	27	28	32	33	34	34

1. $(H = 34)$ and $(L = 5)$.

2. Compute the Range(R) : $R = H - L = 34 - 5 = 29$
3. Select the number of classes desired(**5 classes)**.
4. Determine **the class width** ; $W = \frac{29}{5} = 5.8 \approx 6$ (rounding up) .
5. The **lower limit** of the **first class** = **5**. Add the width **to this lower limit** to get the upper class-limit = 11 then the **first class is** $[5; 11[$.

Class	Frequency
$[5; 11[$	2
$[11; 17[$	8
$[17; 23[$	7
$[23; 29[$	7
$[29; 35[$	4
Total	$n = 28$

Frequency table 3 Example 2.1.3

Remark 2.1.2 *Class mark (Mid points): it is the average of the lower and upper class limits .*

- The midpoint is the numeric location of the center of the class.
- Midpoints are necessary for graphing.

Example:

Class	Frequency	Midpoints(c_i)
$[5; 11[$	2	$\frac{5+11}{2} = 8$
$[11; 17[$	8	$\frac{11+17}{2} = 14$
$[17; 23[$	7	20
$[23; 29[$	7	26
$[29; 35[$	4	32
Total	$n = 28$	/

Frequency table Example 3

Remark: Usually the **formulas** to determine the number of classes **are** given by:
i) Sturges

$$k = \log_2(n + 1)$$

ii) Brooks-Carruthers.

$$k = 5\log_{10}(n)$$

iii) Huntsberger

$$k = 1 + 3.332\log_{10}(n)$$

Where n is the total number of observations.
In example 3, we use the formula below to estimate number of intervals.

$$k = \log_2(n+1) = \log_2(29) \approx 5$$

2.2 Relative frequency

Definition 2.2.1 *The relative frequency of a class is the frequency of the class divided by the total frequency of all classes and is generally expressed as a percentage.*

$$\textbf{Relative frequency} = \frac{\text{Frequency}}{\text{Total number of observations}}$$

Let n_i denote the frequency of the class i and let n be sum of all frequencies(total number of values). Then the **relative frequency** f_i for the class i is defined as the ratio:

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

The sum of the relative frequencies of all classes is clearly 1.

Remark 2.2.1 *Relative frequency determines the proportion of observation in the particular class relative to the total observations.*

2.3 Percentage

Definition 2.3.1 *Find the percentage of values in each class by using the formula:*

$$\% = \frac{n_i}{n} \times\ 100\%$$

where n_i is frequency of the class i (or value x_i) and n is total number of values.

Remark 2.3.1 *The percent frequency of a class is the relative frequency multiplied by 100.*

2.4 Percentage-Relative frequency Distribution

Definition 2.4.1 *A relative frequency distribution is a tabular summary of a set of data showing the relative frequency for each class.*

Example 2.4.1 ***(Example 2.1.2)***

class(or Modalities) x_i	*Frequency*(n_i)	*Relative frequency* $\left(f_i = \frac{n_i}{n}\right)$	*Percent*$\left(\frac{n_i}{n} \times 100\%\right)$
2	3	$\frac{3}{20} = 0.15$	$\frac{3}{20} \times 100 = 15$
3	5	$\frac{5}{20} = 0.25$	25
4	3	$\frac{3}{20}$	15
5	6	$\frac{6}{20}$	30
6	2	$\frac{2}{20}$	10
8	1	$\frac{1}{20} = 0.05$	5
Total	$n = 20$	*1*	*100%*

2.5 Cumulative Frequency

Sometimes investigator is interested to know the number of observations less than a particular value. This is possible by computing the cumulative frequency.

2.5.1 Cumulative Frequency

Definition 2.5.1 *Cumulative frequency is the number of observations* ***less than*** *or equal to a specific value.*

2.5.2 Cumulative Frequency Distribution

Definition 2.5.2 *Cumulative Frequency Distribution is the tabular arrangement of class interval together with their corresponding cumulative frequencies.*

Frequency distribution of **Example 2.1.2**:

class(or Modalities)x_i	Frequency(n_i)
2	3
3	5
4	3
5	6
6	2
8	1
Total	$n = 20$

While cumulative frequency distribution is presented in table below:

	Cumulative Frequency
Less than or equal to 2	3
Less than or equal to 3	$3 + 5 = 8$
Less than or equal to 4	$3 + 5 + 3 = 11$
Less than or equal to 5	17
Less than or equal to 6	19
Less than or equal to 8	20

Or

class(or Modalities)x_i	Frequency(n_i)	Cumulative Frequency
2	3	3
3	5	$3 + 5 = 8$
4	3	$3 + 5 + 3 = 11$
5	6	17
6	2	19
8	1	20
Total	$n = 20$	/

Example 2.5.1 ***(Exercise 3 Chapter 1)***

Frequency distribution is:

class	No. of Students(Frequency)(n_i)
$[1.5; 1.6[$	3
$[1.6; 1.7[$	7
$[1.7; 1.8[$	4
$[1.8; 1.9[$	6
Total	$n = \sum n_i = 20$

While cumulative frequency distribution is presented in table below:

	Cumulative Frequency
Less than 1.6	3
Less than 1.7	10
Less than 1.8	14
Less than 1.9	20

Or

class	No. of Students(Frequency)(n_i)	Cumulative Frequency
$[1.5; 1.6[$	3	3
$[1.6; 1.7[$	7	10
$[1.7; 1.8[$	4	14
$[1.8; 1.9[$	6	20
Total	$n = \sum n_i = 20$	/

2.6 Cumulative relative frequency

2.6.1 Cumulative relative frequency

Definition 2.6.1 *The cumulative relative frequency for the class i is defined by:*

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

2.6.2 Cumulative Relative Frequency Distribution

Definition 2.6.2 *Cumulative Relative Frequency Distribution is the tabular arrangement of class interval together with their corresponding cumulative relative frequencies.*

Example 2.6.1 ***(Example 2.1.2)***

	Cumulative Relative Frequency
Less than or equal to 2	0.15
Less than or equal to 3	$0.15 + 0.25 = 0.4$
Less than or equal to 4	0.55
Less than or equal to 5	0.85
Less than or equal to 6	0.95
Less than or equal to 8	1

2.7 Graphical Representation

Once the data have been organized into a frequency distribution, they can be presented in graphical form. Statistical graphs are useful both for describing and analyzing the data set. In this section, we introduce several methods for graphically representing both qualitative and quantitative data.

2.7.1 Graphical Representation of categorical frequency distribution

We suppose that the variable we are stady is categorical(qualitative).
We can be represented by using:

1. Circle graph (Circle diagram, Pie diagram or Pie chart)
2. Bar diagram(Bar chart)

Circle graph (Pie chart)

Definition 2.7.1 *A pie chart is a circle that is divided in to sections or wedges according to the percentage of frequencies in each category of the distribution. The angle of the sector is obtained using:*

$$\textit{Angle of sector} = \frac{\textit{value of part}}{\textit{the whole quantity}} \times 360$$

$$\theta_i = \frac{n_i}{n} \times 360 = \frac{n_i \times 360}{n}$$

Using the frequency distribution given in Example 2.1.1, construct a pie chart.

Marital status	Frequency(n_i)	$\theta_i = \frac{n_i}{n} \times 360$
M	5	$\theta_1 = \frac{5}{25} \times 360 = 72°$
S	7	100.8°
D	7	100.8°
W	6	86.4°
Sum	$n = 25$	360°

Table1

Circle graph is shown in Fig.1 based on data presented in Table 1

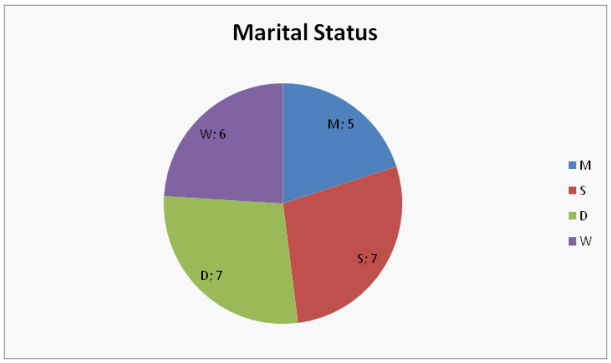


Fig.1

Bar diagram (Bar chart or bar graph)

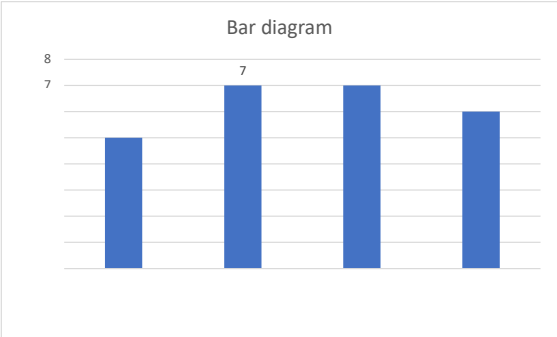
Definition 2.7.2 *A bar diagram consists of bars corresponding to each of the possible values, whose heights are equal to the frequencies (a frequency, relative frequency, or percent)*

Remark 2.7.1

- *The bars are thick lines having the same breadth.*
- Bars can be drawn either vertically or horizontally.

Example 2.7.1 *Using the frequency distribution given in Example 2.1.1, construct*

a bar diagram.



2.7.2 Graphical representation of ungrouped frequency distribution

Now suppose we are interested in a quantitative variable. It can be represented using:

1. Line graph
2. Bar diagram (bar graph)
3. Pie diagram(Circle graph)

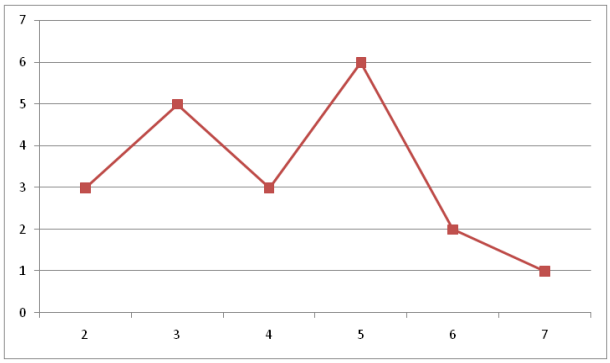
Line graph

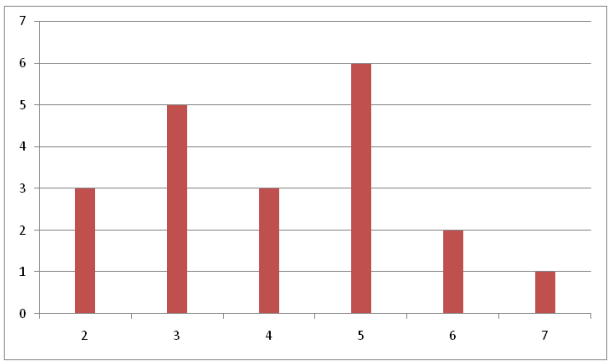
Most common graphical representation

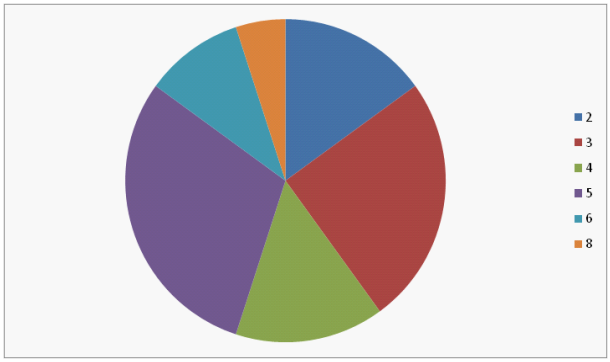
1) By plotting X axis on horizontally while Y axis vertically.
2) Find out the intersecting point or origin and join all intersections.

Example 2.7.2 *For the data of Example 2.1.2, construct a line graph , bar diagram*

and pie diagram.







2.7.3 Graphical representation of grouped frequency distribution

Can be represented by using:

1. Histogram

2. Frequency polygon

3. The cumulative frequency graph (or ogive)

Histogram

Histogram is one of the most popular method for presenting continuous frequency distribution in a form of graph.

Definition 2.7.3 *A histogram is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis. In a histogram, the bars are drawn adjacent to each other without any gaps.*

Example: Using the frequency distribution given in Example 2.1.3, construct a histogram.

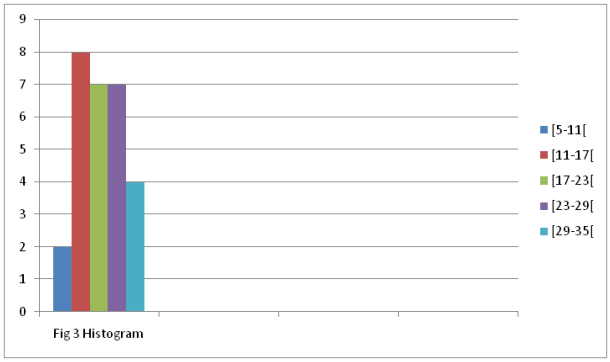


Fig3.1: Histogram

Frequency Polygon

Another way to represent the same data set is by using a frequency polygon.

Definition 2.7.4 *The frequency polygon is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.*

Example 2.7.3 *Using the frequency distribution given in Example 2.1.3, construct a frequency polygon Fig 3.2.*

class	*Midpoints*(c_i)	*Frequency*
$[5;11[$	*8*	*2*
$[11;17[$	*14*	*8*
$[17;23[$	*20*	*7*
$[23;29[$	*26*	*7*
$[29;35[$	*32*	*4*

Table 3

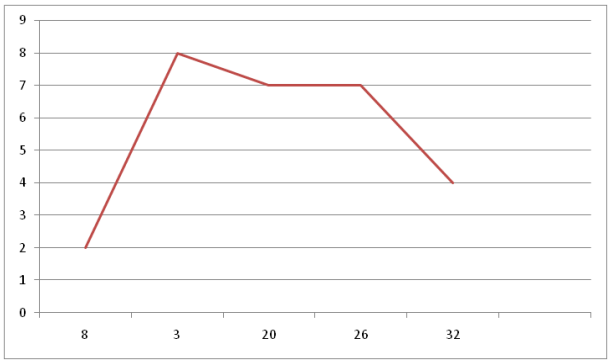


Fig 3.2:polygon

The cumulative frequency graph (or ogive)

A third type of graph that can be used represents the cumulative frequencies of the classes. This type of graph is called a cumulative frequency graph (or curve), also known as an ogive. The cumulative frequency is the sum of the frequencies accumulated up to the upper boundary of each class in the distribution.

Definition 2.7.5 *The ogive is a graph that represents the cumulative frequencies for the classes in a frequency distribution.*

Example 2.7.4 *Construct an ogive for the frequency distribution described in Example 2.1.3" Less than c.f".*

1. Find the cumulative frequency for each class:

class Interval	class boundary	Frequency	Cumulative Frequency
	5	0	0
$[5;11[$	11	2	2
$[11;17[$	17	8	10
$[17;23[$	23	7	17
$[23;29[$	29	7	24
$[29;35[$	35	4	28

Table 3

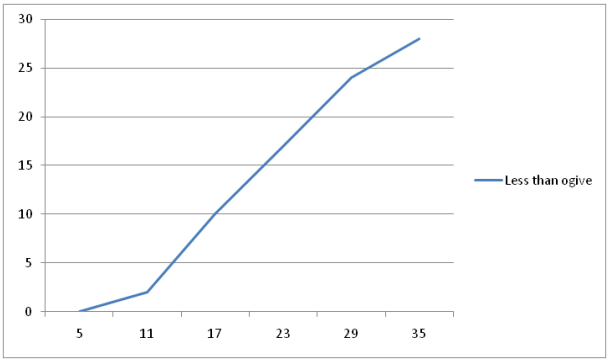


Fig 3.3:"Less than" cf type ogive

Example 2.7.5 *Example of 'Less than' and 'more than' cumulative frequencies based on data reported in table 4.*

Table 4

Class Interval	Class boundary	Frequency	Less than c.f.	More than c.f.
	5	0	0	28
$[5;11[$	11	2	2	26
$[11;17[$	17	8	10	18
$[17;23[$	23	7	17	11
$[23;29[$	29	7	24	4
$[29;35[$	35	4	28	0

The ogives for the cumulative frequency distributions given in above table are drawn in Fig. 4

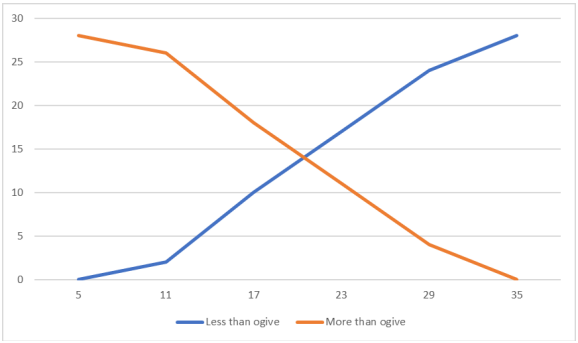


Fig.4:"Less than" and "more than" type ogives

Remark 2.7.2 *A graphic representation of relative-frequency distributions can be obtained from the histogram or the frequency polygon by simply changing the vertical scale from frequency to relative frequency, while keeping the diagram otherwise unchanged. The resulting graphs are called relative-frequency histograms (or percentage histograms) and relative-frequency polygons (or percentage polygons), respectively.*

2.8 Measures of Location

When given a set of raw data, one of the most useful ways to summarize it is by finding an average. An average is a measure of the center of the data set. There are three common measures used to describe the center of a set of numbers: the mean, the median, and the mode.
In this section, we assume that we have collected n numerical values, representing the values of a quantitative variable. Given the dataset of n numbers: $x_1, x_2, x_3, ..., x_n$.

2.8.1 The Mean

To calculate the arithmetic mean(the mean) of a set of data we must first add up (sum) all of the data values (x) and then divide the result by the number of values (n).
The notation $\overline{x}$ will be used to represent a mean.
We obtain the following formula for the mean $(\overline{x})$.

Mean of raw data

Of a set of n numbers $x_1, x_2, x_3, ..., x_n$, is denoted by $\overline{x}$ (read "x bar") and is defined as

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Or we are ready to formally define the mean:

$$\overline{x} = \frac{1}{n}\sum x_i$$

- The summation is over i, but often for shorthand it is simply written as:

$$\overline{x} = \frac{1}{n}\sum x$$

Example 2.8.1 *Obtain the mean of the following data set :*

- $(2, 2, 6, 7, 8, 8, 8, 9, 10, 10, 11, 18)$

  Mean is: $\overline{x} = \dfrac{2+2+6+7+8+8+8+9+10+10+11+18}{12} = 8,25$

- $(1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4\ )$

  $\overline{x} = \dfrac{1+1+1+2+2+3+4+4+4+4+4}{11} = \dfrac{3(1)+2(2)+1(3)+5(4)}{11} = 2,72$

- $(1, 2, 4, 4, 4, 5, 5, 6, 6, 6, 9)$

  $$\text{Mean} = \frac{1+2+4+4+4+5+5+6+6+6+9}{11} = \frac{52}{11} = 4.72$$

Mean of ungrouped data

If the numbers $x_1, x_2, x_3, ..., x_K$ occur with frequencies $n_1, n_2, ..., n_K$ respectively, then their arithmetic mean is given by:

$$\overline{x} = \frac{x_1 \times n_1 + x_2 \times n_2 + x_3 \times n_3 + ... + x_k \times n_k}{n_1 + n_2 + ... + n_K} = \frac{\sum_{i=1}^{k} x_i \times n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i \times n_i$$

where k is the number of classes and $\quad n = n_1 + n_2 + ... + n_K = \sum_{i=1}^{k} n_i$

Example 2.8.2 *Obtain the mean of the following number*$(1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4\ )$

Solution:

- Make a table:

x_i	frequency(n_i)	$x_i \times n_i$
1	3	3
2	2	4
3	1	3
4	5	20
Total	$n = 11$	30

- The mean is:

$$\overline{x} = \frac{\sum_{i=1}^{4} x_i \times n_i}{n} = \frac{30}{11} = 2.72$$

Mean of grouped data

Let the grouped data have k classes.
For grouped data, mean can be computed based on following procedure:

1) Make a table as show

$\overset{a}{\text{Class } i}$ $[\alpha_i,\alpha_{i+1}[$	$\overset{b}{\text{Frequency}}$ (n_i)	$\overset{c}{\text{Midpoint}}$ (c_i)	$\overset{d}{n_i \times c_i}$
$[\alpha_1, \alpha_2[$	n_1	$c_1 = \frac{\alpha_1+\alpha_2}{2}$	$n_1 \times c_1$
.....	.....	.....	.....
.....	.....	.....	.....
$[\alpha_k, \alpha_{k+1}[$	n_k	c_k	$n_k \times c_k$

2)Column (c) find midpoints of each class .
3)Column (d) multiply the frequency by the midpoints for each class.
4)Find the sum of column (d) .
5)Divide the sum obtained in column (d) by the sum of the frequencies obtained in column (b) .
The formula for the mean is

$$\overline{x} = \frac{c_1 \times n_1 + c_2 \times n_2 + c_3 \times n_3 + ... + c_k \times n_k}{n_1 + n_2 + ... + n_k} = \frac{\sum_{i=1}^{k} c_i \times n_i}{n} = \frac{1}{n}\sum_{i=1}^{k} c_i \times n_i$$

where c_i is the **midpoint** of the class i and n_i is the frequency of the class i.

Example 2.8.3 *Find the mean for following grouped data. This data represents the compressive strength for 20 randomly selected concrete cylinder samples.*

Class	Frequency(n_i)
$[5.5 - 10.5[$	1
$[10.5 - 15.5[$	2
$[15.5 - 20.5[$	3
$[20.5 - 25.5[$	5
$[25.5 - 30.5[$	4
$[30.5 - 35.5[$	3
$[35.5 - 40.5[$	2
Summation	20

Class	Frequency(n_i)	**midpoint**(c_i)	$c_i \times n_i$
$[5.5 - 10.5[$	1	8	8
$[10.5 - 15.5[$	2	13	26
$[15.5 - 20.5[$	3	18	54
$[20.5 - 25.5[$	5	23	115
$[25.5 - 30.5[$	4	28	112
$[30.5 - 35.5[$	3	33	99
$[35.5 - 40.5[$	2	38	76
Summation	20	/	490

get the mean

$$\overline{x} = \frac{490}{20} = 24.5$$

2.8.2 The Median

The median is the value located at the center of an ordered data set, where the data have been arranged in ascending (or descending) order.

Determining the Median of raw data or ungrouped data

The median is the middle number. It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not.

Example 2.8.4 *The set:* $\underbrace{1, 2, 4, 4, 4}_{\substack{\textit{Numbers} \\ \textit{below Median}}}$, $\underset{\substack{\uparrow \\ Median}}{\boxed{5}}$, $\underbrace{5, 6, 6, 6, 9}_{\substack{\textit{Numbers} \\ \textit{above Median}}}$

Then, The median = 5

Example 2.8.5 *The set:* $\underbrace{1; 1; 1; 2}$; $\boxed{2; 3}$; $\underbrace{4; 4; 4; 4}$. *Then the median* $= \frac{2+3}{2} = 2.5$

Remark 2.8.1 *You can quickly find the location of the median by using the expression* $\frac{n+1}{2}$. *The letter* n *is the total number of data values in the sample.*

1) If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest)

$$Median = x_{(\frac{n+1}{2})}$$

.

2) If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered.

$$Median = \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2})+1}}{2}$$

Example: Find the median of the following data:

class(x_i)	Frequency(n_i)
2	3
3	5
4	3
5	6
6	2
8	1
Total	$n = 20$

- Find the cumulative frequency table:

class(x_i)	Frequency(n_i)	Cumulative Frequency
2	3	3
3	5	8
4	3	11
5	6	17
6	2	19
8	1	20
Total	$n = 20$	/

- $n = 20$ is an even number, then

$$Median = \frac{x_{10}+x_{11}}{2} = \frac{4+4}{2} = 4$$

The median class for grouped data

Definition 2.8.1 *The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to* $\frac{n}{2}$.

Example 4:

Class	Frequency
$[40-45[$	7
$[45-50[$	10
$[50-55[$	22
$[55-60[$	15
$[60-65[$	12
$[65-70[$	6
$[70-75[$	3
Summation	75

Solutions:
1) First find the less than cumulative frequency.
2)Identify the median class.

Class	Frequency	Cumu.Freq(less than type)
$[40-45[$	7	7
$[45-50[$	10	17
$[50-55[$	22	39
$[55-60[$	15	54
$[60-65[$	12	66
$[65-70[$	6	72
$[70-75[$	3	75
Summation	75	/

We have $\frac{n}{2} = \frac{75}{2} = 37,5$ and 39 is the first cumulative frequency to be greater than or equal to $37,5$ then the median class is $[50-55[$.

Determining median for grouped data

If data are given in the shape of continuous frequency distribution and $[a,b[$is the median class.

The median is defined as

$$\boxed{\text{Median}=a+\frac{b-a}{n_{med}}\left(\frac{n}{2}-c\right)=L_{med}+w\left(\frac{\frac{n}{2}-c}{n_{med}}\right)}$$

where
$a = L_{med}$ =lower class boundary of the median class (i.e., the class containing the median)
n = number of items in the data (i.e., total frequency)
c = the cumulative frequency less than type **preceeding the median class**.
n_{med} =the frequency of the median class
$b - a = w$ = size of the median class interval

Example 2.8.6 *Find the median of example 4.*

$[50, 55[$ is the median class

Class	Frequency	Cumu.Freq(less than type)
$[40-45[$	7	7
$[45-50[$	10	$\boxed{\mathbf{17}=c}$
$\boxed{[\mathbf{50-55}[}$	$\boxed{\mathbf{22}=\mathbf{n}_{med}}$	39
$[55-60[$	15	54
$[60-65[$	12	66
$[65-70[$	6	72
$[70-75[$	3	75
Summation	$\boxed{n=75}$	/

$$\begin{aligned}\text{Median} &= a+\frac{b-a}{n_{med}}\left(\frac{n}{2}-c\right)\\ &= 50+\frac{5}{22}\left(\frac{75}{2}-17\right)\\ &= 54.65\end{aligned}$$

2.8.3 The Mode

Definition 2.8.2 *The mode of a set of numbers is that* ***value*** *which* ***occurs*** *with the* ***greatest frequency****. If all the data values are different, then by definition, the data set has no mode.*

Remark 2.8.2 *1. A data set that has only one value that occurs with the greatest frequency is said to be unimodal.*

2. If a data set has two values that occur with the same greatest frequency, both values are the mode and the data set is said to be bimodal.

3. If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be multimodal.

4. In case of discrete distribution the value having the maximum frequency is the model value.

Example 2.8.7 *The set 2, 2, 6, 7, 8, 8, 8, 9, 10, 10, 11, 18 has mode* $= 8$.

Example 2.8.8 *The set 3, 6, 8, 10, 12, 13, 15* ***has no mode****.*

Example 2.8.9 *The set 1, 2, 4, 4, 4, 5, 5, 6, 6, 6, 9 has* ***two modes, 4 and 6,*** *and is* ***called bimodal.***

Example 2.8.10 *The data set is: {red, red, red, green, green, yellow, purple, black, and blue}, then the mode is* ***red****.*

Example 2.8.11

x_i	*frequency*(n_i)
2	*2*
3	*1*
7	*3*
8	*1*

Mode $= 7$

Mode class for grouped Data

The class interval which contains the most values is known as **the modal class**.

Example 2.8.12

class	$[0;5[$	$[5;10[$	$[10;15[$	$[15;20[$
Frequency	11	28	15	0

The modal class is $[5; 10[$, since it has the largest frequency.

Remark 2.8.3 *Remember, mean, median, and mode are all examples of averages. However since the data is qualitative, you cannot find the mean and the median. The only average you can find is the mode(example2.8.10).*

2.8.4 Quartiles, Deciles, and Percentiles

If a set of data is arranged in order of magnitude, the middle value (or arithmetic mean of the two middle values) that divides the set into two equal parts is the median. By extending this idea, we can think of those values which divide the set into four equal parts. These values, denoted by $Q1, Q2$, and $Q3$, are called the first, second, and third quartiles, respectively, the value Q_2 being equal to the median. Similarly, the values that divide the data into 10 equal parts are called deciles and are denoted by $D_1, D_2, ..., D_9$ while the values dividing the data into 100 equal parts are called percentiles and are denoted by $P_1, P_2, ..., P_{99}$ The fifth decile and the 50th percentile correspond to the median. The 25th and 75th percentiles correspond to the first and third quartiles, respectively.

$$\underbrace{x_1....}_{25\%} Q_1 \quad \underbrace{.....}_{25\%} Q_2 \quad \underbrace{.....}_{25\%} Q_3 \quad \underbrace{.....}_{25\%} x_n$$

Q_2=Median

$\underbrace{x_1....}$ D_1	$\underbrace{.....}$ D_2	$\underbrace{.....}$ D_3	$\underbrace{.....}$ D_4	$\underbrace{.....}$ D_5	$\underbrace{.....}$ D_6	$\underbrace{.....}$ D_7	$\underbrace{.....}$ D_8	$\underbrace{.....}$ D_9	$\underbrace{.....}$ x_n
10%	10%	10%	10%	10%	10%	10%	10%	10%	10%

D_5=Median

$\underbrace{x_1....}$ P_1	$\underbrace{.....}$ P_2	$\underbrace{.....}$ P_3	.....	.....	$\underbrace{.....}$ P_{98}	$\underbrace{.....}$ P_{99}	$\underbrace{.....}$ x_n
1%	1%	1%	.....	.....	1%	1%	1%

Note: k^{th} percentile, P_k = value at $\left(\frac{kN}{100}\right)$position

$$P_k = \text{the } \left(\frac{kN}{100}\right)^{th} \text{ value}$$

Percentile P_{25} = Quartile Q_1

Percentile P_{50}= Quartile Q_2= Median
Percentile P_{75} = Quartile Q_3

Procedure for finding P_k: 1. Arrange data in ascending form.
2. Compute: $A = \dfrac{kn}{100}$
3.

- If A is an integer:

P_k is halfway between the value of the data in the A^{th} position and the value of the next data.

- If A is a fraction:

$d(P_k) = B$, the next larger integer. Then P_k is the value of the data in the B^{th} position.

Quantiles

Definition 2.8.3 *Collectively, quartiles, deciles, percentiles, and other values obtained by equal subdivisions of the data are called* ***quantiles****.*

Example 2.8.13 *Find the quartiles Q_1, Q_2, and Q_3 of the following data:*

20	30	25	23	22	32	36	18

- Arrange data in ascending form:

18	20	22	23	25	30	32	36

and $n = 8$ even number

$$q_1 = \frac{N}{4} = 2 \Rightarrow Q_1 = \frac{20 + 22}{2} = 21$$
$$q_2 = \frac{2N}{4} = 4 \Rightarrow Q_2 = \frac{23 + 25}{2} = 24 = Median$$
$$q_3 = \frac{3N}{4} = 6 \Rightarrow Q_3 = \frac{30 + 32}{2} = 31$$

Example 2.8.14 *Find the quartiles Q_1, Q_2, and Q_3 of the following data:*

20	30	25	23	22	32	36

Solution:

- Arrange data in ascending form:

Ascending arrangement :

20	22	23	25	30	32	36

and $n = 7$ is odd number

$$q_1 = \frac{7}{4} = 1.75 \Rightarrow Q_1 = 22$$

$$q_2 = \frac{2 \times 7}{4} = 3.5 \Rightarrow Q_2 = 25 = Median$$

$$q_3 = \frac{3 \times 7}{4} = 5.25 \Rightarrow Q_3 = 32$$

Example 2.8.15 *Find the 80^{th} percentile and 25^{th} percentile of the following data:*

18	*20*	*22*	*23*	*25*	*30*	*32*	*36*

Solution:
a)We need to find the 80th percentile. $p = 8 \times\ 0.80 = 6.4$. We round up to the next largest integer (7), and the 80th percentile is the 7^{th} ($P_{80} = x_7 = 32$).
b) We need to find the 25th percentile.$p = 8 \times\ 0.25 = 2$. The 25th percentile is the average of the 2^{nd} and 3^{rd} ranked values: $P_{25} = \frac{x_2+x_3}{2} = \frac{20+22}{2} = 21$
Remarks:
1)The quartile class (class containing Q_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to$\frac{in}{4}$.
2)The decile class (class containing D_i)is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{in}{10}$
3)The percentile class (class containing P_i)is the class with the smallest cumulative frequency (less than type) greater than or equal to$\frac{in}{100}$.
4)For grouped data: we have the following formula
$i = 1, 2, 3$

$$Q_i = L_{Q_i} + w\left(\frac{\frac{in}{4} - c}{n_{Q_i}}\right)$$

where
L_{Q_i} =lower class boundary of the quartile class(i.e., the class containing the quartile)
N = number of items in the data (i.e., total frequency)
c = the cumulative frequency less than type preceeding the quartile class.
n_{Q_i} =the frequency of the quartile class
w = size of the quartile class interval

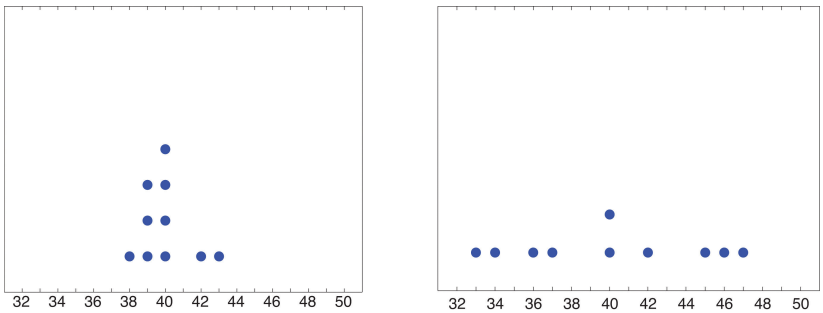
Exercise 1: Table* below :Two Data Sets

Set I	40	38	42	40	39	39	43	40	39	40

Set II	46	37	40	33	42	36	40	47	34	45

Find the mode, median, mean
Solution:

- The two sets of ten measurements each center at the same value: they both have mean, median, and mode 40. Nevertheless a glance at the figure shows that they are markedly different.
- In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.
- Look at the two data sets in Table * "Two Data Sets" and the graphical representation of each, called a dot plot, in Figure "Dot Plots of Data Sets"



(a) Set I (b) Set II

Figure : Dot Plots of Data Sets

2.9 Measures of Dispersion

The mean is the value usually used to indicate the centre of a distribution. If we are dealing with quantity variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the average.
For the spread or variability of a data set, three measures are commonly used:
• Range.
• Variance and Standard deviation.
Each measure will be discussed in this section.

2.9.1 The Range

One very simple measure of dispersion is the range.

Definition 2.9.1 *The range (R) is the largest value minus the smallest value.*

$$Range = largest\ value\text{-}smallest\ value.$$

The symbol R is used for the range.

Example 2.9.1 *Find the range of each data set in Table * "Two Data Sets".*

1. For Data Set I the maximum is 43 and the minimum is 38, so the range is $R = 43 - 38 = 5$.

2. For Data Set II the maximum is 47 and the minimum is 33, so the range is $R = 47 - 33 = 14$

Remark 2.9.1 *The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite.*

2.9.2 The variance

The variance of raw data

- For raw data $(x_1, x_2, x_3, ..., x_n)$

The variance is defined by

$$V = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{n} = \frac{1}{n} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

The variance of ungrouped data

If the numbers $x_1, x_2, x_3, ..., x_K$ occur with frequencies $n_1, n_2, ..., n_K$, respectively The variance is defined by

$$V = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2$$

where f_i is relative frequency.

- We usually use the following short cut formula:

1)for raw data

$$V = \frac{\sum_{i=1}^{N} x_i^2 - n\overline{x}^2}{n} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - (\overline{x})^2 = \bar{x^2} - (\bar{x})^2$$

2)For frequency distribution

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

2.9.3 Standard Deviation

Definition 2.9.2 *We define standard deviation* $(sd = \sigma)$ *as the square root of the variance.*

$$\begin{aligned} \textit{standard deviation} &= \sqrt{\textit{variance}}. \\ \sigma &= \sqrt{V} \end{aligned}$$

Remark 2.9.2 $\sigma^2 = V$

Example 2.9.2 *Find the variance and standard deviation of the following sample data:*

5	17	12	10

- Find $\overline{x}$ the mean for the data: $\overline{x} = 11$

x_i	5	10	12	17	Total
$(x_i - \overline{x})^2$	36	1	1	36	74

The variance

$$V = \frac{\sum_{i=1}^{4} (x_i - \overline{x})^2}{4} = \frac{74}{4} = 18.5$$

and standard deviation

$$\sigma = \sqrt{V} = \sqrt{18.5} = 4.3$$

Exercise 2: Two people work in a factory making parts for concrete mixer. The table shows how complete parts they make in 6 day .

Worker A	Worker B
10	35
60	45
50	30
30	35
40	40
20	25

1) Find the mean of esche worker.

2) Find the variance and standard deviation for the data set for workerA and workerB .

Solution:

1)The mean for worker A

$\overline{x_1} = 35$ Parts

The mean for worker B

$\overline{x_2} = 35$ Parts

2) worker A: the mean for the data: $\overline{x} = 35$

- Construct the following table:

Values(x_i)	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
10	-25	625
60	$+25$	625
50	$+15$	225
30	-5	25
40	$+5$	25
20	-15	225
Total	/	1750

i) The variance

$$V = \frac{\sum_{i=1}^{6} (x_i - \overline{x})^2}{6} = \frac{1750}{6} = 291.7$$

ii) Take the square root to get the standad deviation:

$$\sigma = \sqrt{V} = \sqrt{\frac{\sum_{i=1}^{6} (x_i - \overline{x})^2}{6}} = \sqrt{291.7} = 17.1$$

- Worker B:the mean for the data: $\overline{x} = 35$

i) Construct the following table:

Values(x_i)	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
35	0	0
45	+10	100
30	−5	25
35	0	0
40	+5	25
25	+10	100
Total	/	250

get the variance

$$V = \frac{\sum_{i=1}^{6} (x_i - \overline{x})^2}{6} = \frac{250}{6} = 41.7$$

ii) Take the square root to get the standad deviation:

$$\sigma = \sqrt{V} = \sqrt{\frac{\sum_{i=1}^{6} (x_i - \overline{x})^2}{6}} = \sqrt{41.7} = 6.5$$

Hence, the standad deviation is 6.5.

Uses of the Variance and Standard Deviation:

• Variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.

2.9.4 Coefficient of Variation

Definition 2.9.3 *The coefficient of variation (C.V)is defined as the ratio of standard deviation to the mean usually expressed as percents.*

$$CV = \frac{\sigma}{\overline{x}} \text{ or } \frac{\sigma \times 100}{\overline{x}}$$

- The distribution having less $C.V$ is said to be less variable or more consistent.

σ =standard deviation

Example 2.9.3 *Two groups of people were trained to perform a certain task and tested to find out which group is faster to learn the task. For the two groups the following information was given:*

Value	*Group one*	*Group two*
Mean	$10.4min$	$11.9min$
Stan.dev.	$1.2min$	$1.3min$

Which group is more consistent in its performance?
Solution : Use coefficient of variation:

$$CV_1 = \frac{\sigma_1 \times 100}{\overline{x}_1} = \frac{1.2 \times 100}{10.4} = 11.54\%$$

$$CV_2 = \frac{\sigma_2 \times 100}{\overline{x}_2} = \frac{1.3 \times 100}{11.9} = 10.92\%$$

Since $C.V_2 < C.V_1$, group 2 is more consistent.

2.10 Measures of shape

In statistics moments are certain constant values in a given distribution which help us to ascertain the nature and form of distribution. They provide the only measures of skewness and kurtosis.

2.10.1 Moments

Definition 2.10.1 *1)If X is a variable that assume the values $x_1, x_2, \ldots.., x_n$ with mean $\overline{x}$ and $r \in \mathbb{N}$.*

The r^{th} moment is defined as:

$$m_r = \frac{x_1^r + x_2^r + x_3^r + ... + x_n^r}{n} = \frac{\sum_{i=1}^{n} x_i^r}{n}$$

2) For the case of frequency distribution this is expressed as:

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

or

$$m_r = \frac{\sum_{i=1}^{k} n_i c_i^r}{n}$$

c_i is mid values in **case of class intervals.**

Remark 2.10.1 *If $r = 1$,it is the simple arithmetic mean, this is called the first moment.*

Example 2.10.1 *The first moment of the values $1, 3, 6, 10$ is*

$$m_1 = \frac{1+3+6+10}{4} = 5$$

Example 2.10.2 *The second moment of the values $1, 3, 6, 10$ is*

$$\frac{1^2+3^2+6^2+10^2}{4} = \frac{1+9+36+100}{4} = \frac{146}{4} = 36.5$$

Example 2.10.3 *The third moment of the values $1, 3, 6, 10$ is*

$$\frac{1^3+3^3+6^3+10^3}{4} = \frac{1+27+216+1000}{4} = 311$$

2.10.2 Central moment

Definition 2.10.2 *The r^{th} moment about the mean (the r^{th} central moment) denoted by μ_r and defined as:*

$$\mu_r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^r}{n}$$

- For the case of frequency distribution:

If $x_1, x_2, x_3, ..., x_K$ are k values (or mid values (c_i) in **case of class intervals**) of a variable x with their corresponding frequencies $n_1, n_2, ..., n_K$, respectively,then their the r^{th} central moment is expressed as:

$$\mu_r = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^r}{n} \text{ or } \mu_r = \frac{\sum_{i=1}^{k} n_i (c_i - \overline{x})^r}{n}$$

Remark 2.10.2 *If $r = 2$, it is variance, this is called the second central moment.*

The second moment μ_2 is equal to the variance (V)

Example 2.10.4 *For the following distribution, calculate first four moments about mean.*

Class x	2	3	4	5	6
Frequency(n_i)	1	3	7	3	1

Calculation of first four central moments:

i) First we construct the following frequency distribution for calculation of central moments:

x	n_i	$d = (x - \overline{x})$	$n_i d$	$n_i d^2$	$n_i d^3$	$n_i d^4$
2	1	-2	-2	4	-8	16
3	3	-1	-3	3	-3	3
4	7	0	0	0	0	0
5	3	1	3	3	3	3
6	1	2	2	4	8	16
totals	$n = 15$	/	0	14	0	38

We therefore have,

- First central moment:

$$\mu_1 = \frac{\sum_i n_i (x_i - \overline{x})}{n} = \frac{\sum_i n_i d}{n} = \frac{0}{15}$$

- Second central moment: The second moment μ_2 is equal to the variance (V)

$$\mu_2 = \frac{\sum_i n_i \left(x_i - \overline{x}\right)^2}{n} = \frac{\sum_i n_i d^2}{n} = \frac{14}{15} = 0.93 = V$$

- Third central moment:

$$\mu_3 = \frac{\sum_i n_i \left(x_i - \overline{x}\right)^3}{n} = \frac{\sum_i n_i d^3}{n} = \frac{0}{15} = 0$$

- Fourth central moment:

$$\mu_4 = \frac{\sum_i n_i \left(x_i - \overline{x}\right)^4}{n} = \frac{\sum_i n_i d^4}{n} = \frac{38}{15} = 2.53$$

2.10.3 Applications of Moments

As mentioned above, the first moment is the mean and the second moment about the mean is the sample variance. Karl Pearson(British mathematician(1857 -1936)) introduced the use of the third moment about the mean in calculating skewness and the fourth moment about the mean in the calculation of kurtosis.

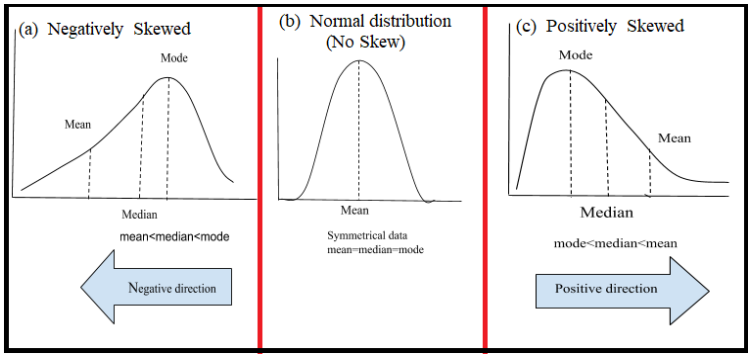
The moment coefficient of skewness

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Where σ is the standard deviation.
The shape of the curve is determined by the value of γ_1.

1. If $\gamma_1 > 0$ then the distribution is positively skewed .
2. If $\gamma_1 = 0$ then the distribution is symmetric.

3. If $\gamma_1 < 0$ then the distributionis negatively skewed



Example:From the above example

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = 0$$

Since $\gamma_1 = 0$ that means the distributions is symmetrical.

Remark 2.10.3 *In a distribution with zero skew, the mean and median are equal.*

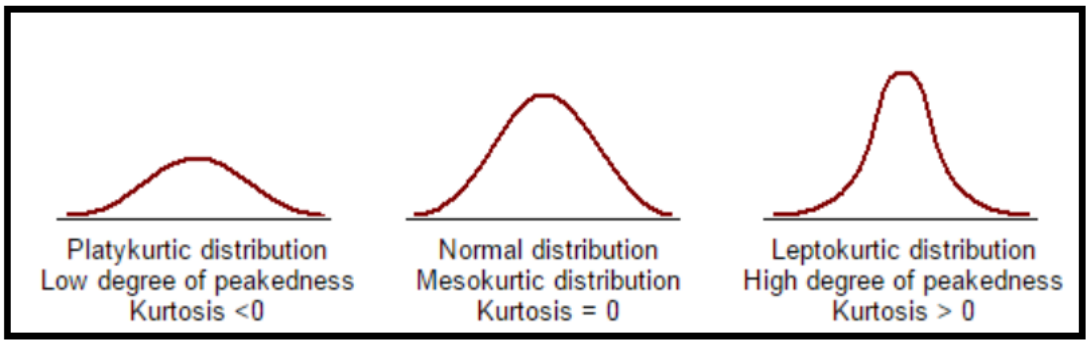
Measures of kurtosis

The moment coefficient of kurtosis:
Denoted by γ_2 and given by:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

- Positive kurtosis($\gamma_2 > 0$) . A distribution has fatter tails than the normal distribution.(more peakedness than normal curve)
- Negative kurtosis($\gamma_2 < 0$) . A distribution has thinner tails than the normal distribution.(flatter than normal curve)
- $\gamma_2 = 0$ same peakedness as normal curve(a mesokurtic curve).

The different types of Kurtosis:



Example 2.10.5 *From the above example 2.10.4.*

Second central moment: The second moment μ_2 is equal to the variance (V)

$$\mu_2 = 0.93 = V = \sigma^2$$

and the fourth central moment: $\mu_4 = 2.53$.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{V^2} - 3 = -0,06$$

Since $\gamma_2 < 0$(Negative kurtosis). Thus, flatter than a normal distribution curve.

2.11 Exercises

Exercise N°1:
Twenty-five army inductees were given a blood test to determine their blood type. The data set is :

A	B	B	AB	O	O	O	B	AB	B	O	B	B
O	O	O	AB	O	AB	A	O	A	B	A	A	

1) Identify this data set.
2)Construct a table that gives the frequency distribution of this data.
3)Find the percentage of values in each class.
4) Analyze the data and interpret the results in a graphical form by;
(a) A bar diagram.

(b) A pie chart.

Exercise N°2:

In a season a football team scored a total of 50 matches. the table below gives a summary of the number of goals per match.

Goals per Match	0	1	2	3	4	5
Frequency	4	6	n_3	8	2	1

1)Identify the variable, modalities and type of variable of this Data set.
2)In how many matches did they score 2 goals?
3)Make a frequency table displaying relative frequencies and percentages.
4) i) Construct a bar diagram .
ii) Construct a pie diagram.

Exercise N°3:

The table below gives data on the heights(in cm) of 51 children.

Class	$[140-150[$	$[150-160[$	$[160-170[$	$[170-180[$
Frequency	6	16	21	8

1) Represent the data in a histogram and construct a frequency polygon.
2) Make a cumulative frequency table. Then draw a cumulative frequency graph.

Exercise N°4:

The following data represent the mark of 20 students.

80	76	88	85	80
70	60	62	70	85
65	60	63	74	75
76	70	70	80	85

1) Construct a frequency distribution, which is grouped (use classes $[60-70[$,$[70-80[$ and so on).
2)Find the relative frequency table. Then construct a relative frequency histogram.
3)Find the midpoints of each class with. Then construct a relative frequency polygon.

Exercise N°5:

Scores of 30 students are given below:

3	30	14	30	27	11	25	16	18	33	18	29	29	20	20
20	10	25	14	18	9	49	35	14	39	29	22	29	15	25

1)Determine k number of Classes (use $k = 1 + 3.332 \log_{10}(n)$).
2)Find the class width. Then make a frequency table with (6) classes showing the class interval, class boundaries, frequencies and cumulative frequencies("less than" and "more than" type).
3)Draw a histogram to represent these data.
4)Draw an ogive for the cumulative frequency distributions("less than" and "more than" cf type ogives).

Exercise N°6:
1)The birth weights in pounds of five babies born in a hospital on a certain day are: 9.2, 6.4, 10.5, 8.1, and 7.8. Obtain the mean.
2)Find the median of the birth-weight data given.

Exercise N°7:
The ordered weights of the 8 boxes of cereal are:

684	684	684	686	686	691	691	691

1)Calculate the mean, median and mode.
2) Find the 80^{th} percentile and the 25^{th} percentile.

Exercise N°8:For the table below:

Class x	0	1	2	3	4	5
Frequency	8	23	42	18	6	3

1)Construct a bar diagram.
2)Make a cumulative frequency table.
3)Find the mode and calculate the median and the mean.
4)Find the first quartile and the third quartile.
5)Find the second moment (m_2) .Calculate the variance and the standard deviation.

Exercise N°9:
1) Calculate mean, variance and standard deviation of the following frequency distribution:

Marks Obtained	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$[40-50[$	$[50-60[$	$[60-70[$
Frequency	6	12	22	24	16	12	8

2)Compute the Moment coefficient of skewness (γ_1)from the data.

Exercise N°10:
The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the distribution.

2.12 Solutions

Exercise N°1:
1) Identify the data:

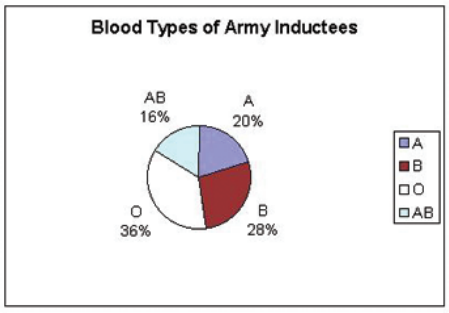
- Individual(Statistical unit): army inductee
- Sample: Twenty-five army inductees.
- Variable: blood type.
- Modalities:$\{A, B, AB, O\}$
- Type of variable:Qualitative variable (Categorical Variable).

2) and 3)
The frequency distribution and percentage are summarized by the following table:

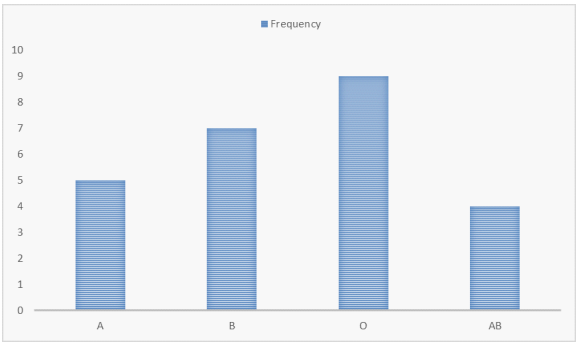
Modalities "blood type"	Frequency	percent
A	5	20
B	7	28
O	9	36
AB	4	16
Total	$n = 25$	100

4)For the sample, more people have type O blood than any other type.

class	Frequency(n_i)	$\theta_i = \frac{n_i}{n} \times 360$
A	5	72°
B	7	100.8°
O	9	129.6°
AB	4	57.6°
Total	$n = 25$	360°



A pie chart



A bar diagram

Exercise N°2:
1)

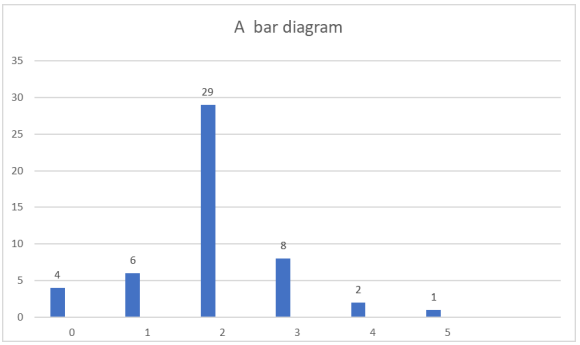
- The variable: number of goals per match.
- Modalities:$\{0, 1, 2, 3, 4\}$
- Type of variable:Quantitative variable (Discrete).

2)We have :$4 + 6 + n_3 + 8 + 2 + 1 = 50$. Then $n_3 = 29$.
Therefore, the number of matches scored 2 goals is : 29

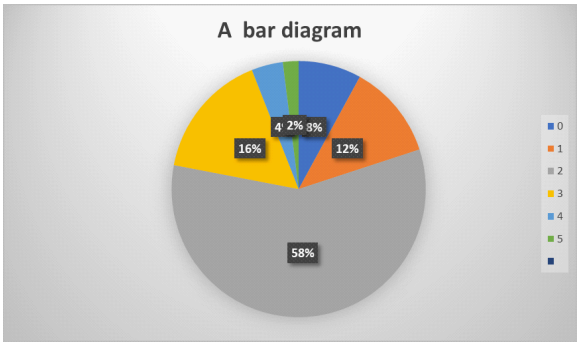
3)The relative frequency and percentage distribution are summarized by the following table:

Goals per Match	Frequency(n_i)	Relative Frequency$\left(f_i = \frac{n_i}{n}\right)$	percentage($f_i \times 100$)	$\theta_i = \frac{n_i}{n} \times 360$
0	4	$\frac{4}{50} = 0.08$	8	28.8
1	6	$\frac{6}{50} = 0.12$	12	43.2
2	29	$\frac{29}{50} = 0.58$	58	208.8
3	8	$\frac{8}{50} = 0.16$	16	57.6
4	2	$\frac{2}{50} = 0.04$	4	14.4
5	1	$\frac{1}{50} = 0.02$	2	7.2
Totals	$n = 50$	1	100	360

4) i) Construct a bar diagram .

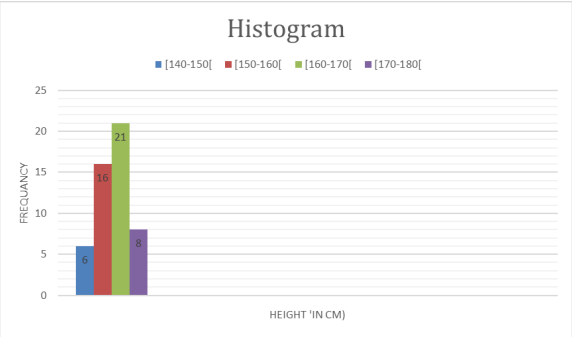


ii) Construct a pie diagram.



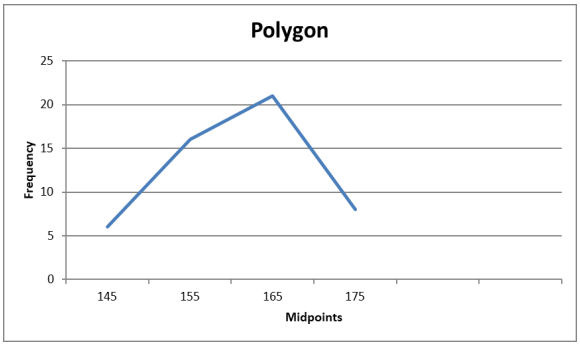
Exercise N°3:

1)



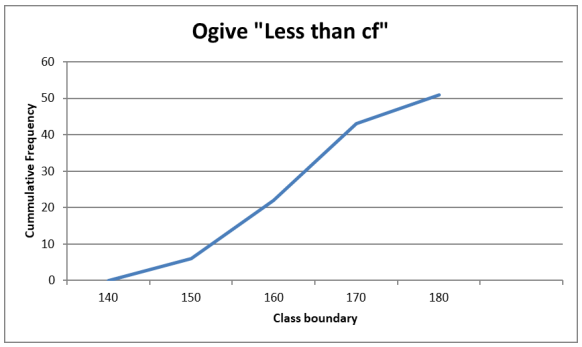
and

Midpoints	145	155	165	175
Frequency	6	16	21	8



2)

Class (height(in cm))	$[140-150[$	$[150-160[$	$[160-170[$	$[170-180[$
Frequency	6	16	21	8
Cumulative frequency	6	22	43	51



Exercise N°4:

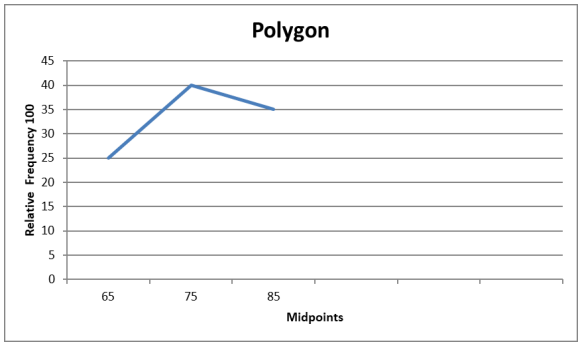
1))The frequency distribution is summarized by the following table:

Class (mark)	$[60-70[$	$[70-80[$	$[80-90[$	
Frequency(n_i)	5	8	7	$n=20$

2)The relative frequency distribution is summarized by the following table:

Class (mark)	$[60-70[$	$[70-80[$	$[80-90[$	
Frequency(n_i)	5	8	7	$\sum n_i = 20$
Relative frequency(f_i)	$\frac{5}{20}=0.25$	$\frac{8}{20}=0.4$	$\frac{7}{20}=0.35$	$\sum f_i = 1$

Class (mark)	$[60-70[$	$[70-80[$	$[80-90[$
Midpoint(c_i)	$\frac{60+70}{2}=65$	75	85
Relative frequency(f_i)	0.25	0.4	0.35



Exercise N°5:

1) The number of classes k can be calculated as under :

$$k = 1 + 3.332\log_{10}(n) = 1 + 3.322 \times \ 1.4771 = 1 + 4.9069262 \simeq 6$$

2) The width of a class interval $=W = \frac{49-3}{6} = 7.66 = 8$ (rounding up) .

- Arrange the raw scores in ascending order:

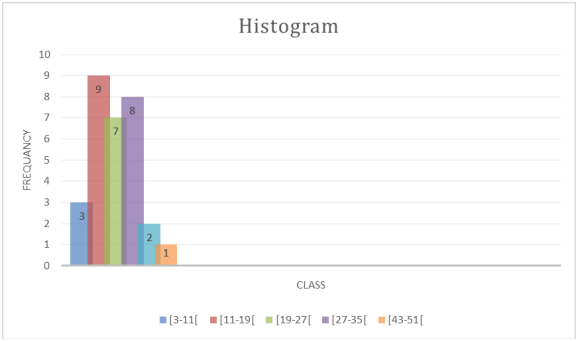
3	9	10	11	14	14	14	15	16	18	18	18	20	20	20
22	25	25	25	27	29	29	29	29	30	30	33	35	39	49

TABLE 1

Class Interval	Frequency(n_i)
$[3;11[$	3
$[11;19[$	9
$[19;27[$	7
$[27;35[$	8
$[35;43[$	2
$[43;51[$	1
Total	$n = 30$

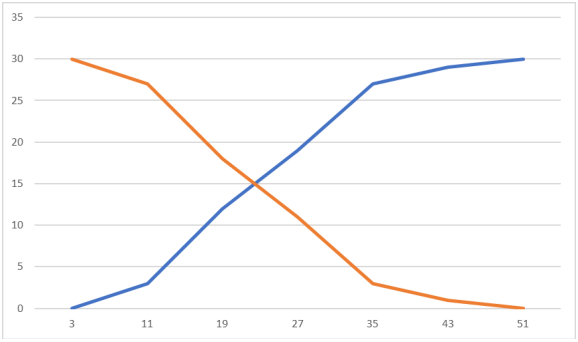
Class Interval	Class boundary	Frequency	Less than c.f.	More than c.f.
	3	0	0	30
$[3;11[$	11	3	3	27
$[11;19[$	19	9	12	18
$[19;27[$	27	7	19	11
$[27;35[$	35	8	27	3
$[35;43[$	43	2	29	1
$[43;51[$	51	1	30	0
	/	$n = 30$		

3) Histogram



4)

Fig.:"Less than" and "more than" type ogives:



Exercise N°6:
1) The mean birth weight for these data is

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4}{n} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42}{5} = 8.4 \text{ pounds}$$

2) The measurements, ordered from smallest to largest, are

$$6.4 \quad 7.8 \quad 8.1 \quad 9.2 \quad 10.5$$

The middle value is 8.1, and the median is therefore 8.1 pounds.
Exercise N°7:
1)Calculate the mean:

$$\overline{x} = \frac{684 + 684 + 684 + 686 + 686 + 691 + 691 + 691}{8} = 687,125$$

i) Arrange the raw in ascending order:

684	684	684	686	686	691	691	691

The median = 686 and mode = 684 and 691 .
2)
a)We need to find the 80^{th} percentile. $p = 8 \times \ 0.80 = 6.4$.

- We round up to the next largest integer (7), and the 80^{th} percentile is the 7^{th}:

$$P_{80} = x_7 = 691$$

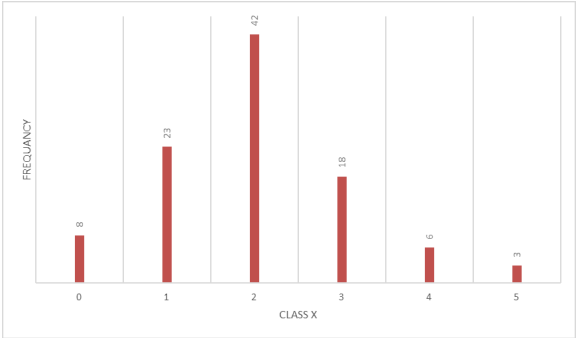
.

b) We need to find the 25^{th} percentile. $p = 8 \times \ 0.25 = 2$.

- The 25^{th} percentile is the average of the 2^{nd} and 3^{rd} ranked values. Then,

$$P_{25} = \frac{684 + 684}{2} = 684$$

Exercise N°8:
1)Construct a bar diagram.



2)

Class x	0	1	2	3	4	5
Frequency(n_i)	8	23	42	18	6	3
cumulative frequency	8	31	73	91	97	100

3) The mode = the median = the mean = 2.
4) The first quartile $Q_1 = 1$ and the third quartile: $Q_3 = 3$
5)Find the second moment (m_2)

$$m_2 = \frac{1}{n}\sum_i n_i x_i^2 = 5.24 = \bar{x^2}$$

Calculate the variance:

$$V = \bar{x^2} - (\bar{x})^2 = 5.24 - 4 = 1.24$$

and the standard deviation:

$$\sigma = \sqrt{V} = \sqrt{1.24} = 1.11$$

Exercise N°9:
1) Calculate mean: $\overline{x} = 35$, variance:$V = \sigma^2 = 260$ and standard deviation:$\sigma = \sqrt{260} = 16.12$
2)Compute the Moment coefficient of skewness (γ_1)from the data:

- First we construct the following frequency distribution for calculation of central moments:

Class Interval	Frequency(n_i)	Midvalues $x = c_i$	$d = (x - \overline{x})$	$n_i d^3$
$[0-10[$	6	5	-30	-162000
$[10-20[$	12	15	-20	$-96\ 000$
$[20-30[$	22	25	-10	-22000
$[30-40[$	24	35	0	0
$[40-50[$	16	45	10	16000
$[50-60[$	12	55	20	96000
$[60-70[$	8	65	30	216000
totals	$n = 100$	/	/	48000

$\mu_3 = 480$ and $\sigma = \sqrt{260} = 16.12$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{480}{\left(\sqrt{260}\right)^3} = 0.11$$

Since $\gamma_1 > 0$, then the distribution is positively skewed.
Exercise N°10:
We have: $\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 0.7$,$\mu_4 = 18.75$ and $\sigma = \sqrt{\mu_2} = \sqrt{2.5}$

- To examine skewness, we compute γ_1.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{0.7}{\left(\sqrt{2.5}\right)^3} = 0.17$$

Since $\gamma_1 > 0$ then the distribution is positively skewed .

- Kurtosis is given by the coefficient

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{18.75}{\left(\sqrt{2.5}\right)^4} - 3 = 0$$

Hence, the curve is mesokutic(normal curve).

Chapter 3

Two-Variable Statistics

3.1 Introduction

In this chapter, we will focus on the description of bivariate data (data involving two variables). Specifically, we will discuss descriptive methods used to explore the joint distribution of pairs of values from two variables. For example, we might consider the height and weight recorded for each individual in a study.

3.2 Contingency Table

In statistics, a contingency table (also known as a cross-tabulation, crosstab, or two-way frequency table) is a type of table in matrix format that displays data for one variable in the rows and data for another variable in the columns. By organizing the data in rows and columns, the relationship between the two variables can be easily observed by examining the table cells where the two datasets intersect. Contingency tables are frequently used in engineering and scientific research.
Let $X = \{x_1; ..., x_k\}$ and $Y = \{y_1; ...; y_l\}$ are two datasets.
For the technical details that follow, we assume a contingency table of counts with k rows and l columns as in the table below.
Let n_{ij} be the observed count for the ithrow ($i = 1$ to k) and j th column ($j = 1$ to

l).

$X\backslash Y$	Column1 y_1	...	Column j y_j	.....	Column l y_l	Total
Row1(x_1)	n_{11}	...	n_{1j}	.....	n_{1l}	$n_{1.}$
..			....		...	...
Row i (x_i)	n_{i1}		n_{ij}		n_{il}	$n_{i.}$
...			....		...	...
Row k (x_k)	n_{k1}		n_{kj}	....	n_{kl}	$n_{k.}$
Total	$n_{.1}$	....	$n_{.j}$	...		N

Let the row and column marginal totals be designated as $n_{i\ .}$ and $n_{.\ j}$, respectively, where

$$n_{i\ .} = \sum_{j=1}^{l} n_{ij}$$

and

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

Let the total number of counts in the table be N, where

$$\begin{aligned} N &= \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} \\ &= \sum_{i=1}^{k} n_{i\ .} \\ &= \sum_{j=1}^{l} n_{.\ j} \end{aligned}$$

Example 3.2.1 *A survey is made among 200 students in a middle school. They are asked, how they travel to school. The table given below shows the results of the survey:(Y=transport)and (X=Gender)*

$X\backslash Y$	*car*	*bus*	*other transport*	*Total*
Girl	22	38	40	100
Boy	25	34	41	100
Total	47	72	81	200

Example 3.2.2 *The table shows the results of a survey of* 100 *randomly-selected people entering an amusement park who were asked whether they were planning to ride the Monster Loop, a roller coaster.*

$\diagdown$	*Ages* $[8-15[$	*Ages* $[15-25[$	*Ages* $[25-35[$	*35 and Older*
Yes	19	23	8	14
No	8	11	12	5

Example 3.2.3 *This two-way table shows information about the students in years 8, 9 and 10.*

$\diagdown$	year 8	year 9	year 10
Boys	45	38	51
Girls	32	52	28

Example 3.2.4 *(Age* (x) *in years and blood pressure*(y) *in mm Hg)*

$x \diagdown y$	12	14	15	$n_{i.}$
30	5	2	1	
40	4	6	6	
50	0	3	5	
60	0	0	8	
$n_{.j}$				$N =$

3.3 Marginal and Conditional Distributions in Contingency Tables

Contingency tables are a fantastic way of finding marginal and conditional distributions. These two distributions are types of frequency distributions.

3.3.1 Frequency Distribution

These distributions represent the frequency distribution of one variable without regard for other variables. You can find these distributions in the margins of a contingency table.

A joint frequency

Definition 3.3.1 *Each entry in the table is called* ***a joint frequency****(or* ***conditional frequency****).*

Remark 3.3.1 *A joint Frequency* n_{ij} *is a frequency of* (x_i, y_j) $1 \leq i \leq k$ *and* $1 \leq j \leq l$.

Marginal frequency

Definition 3.3.2 *The sums of the rows and columns in a two-way table are called* ***marginal frequencies****.*

1. $n_{i\ .}$ is a marginal frequency of x_i, $\quad n_{i\ .} = \sum_{j=1}^{l} n_{ij}$.

2. $n_{.\ j}$ is a marginal frequency of y_j, $\quad n_{.\ j} = \sum_{i=1}^{k} n_{ij}$.

Remark 3.3.2 *For these distributions, you specify the value for one of the variables in the contingency table and then assess the distribution of frequencies for the other variable. In other words, you condition the frequency distribution for one variable by setting a value of the other variable. That might sound complicated, but it's easy using a contingency table. Just look across one row or down one column.*

Examples: Use the survey results in examples above to make a two-way table that shows the joint and marginal frequencies.

Example 3.3.1 *Refer to the two-way table from example* 3.2.4. *Make a table of joint and marginal frequencies.*

A table of joint and marginal frequencies

$x \diagdown y$	12	14	15	Total
30	$5 = n_{11}$	$2 = n_{12}$	$1 = n_{13}$	$n_{1.} = n_{11} + n_{12} + n_{13} = 8$
40	$4 = n_{21}$	6	6	16
50	$0 = n_{31}$	3	5	8
60	$0 = n_{41}$	0	8	8
Total	$n_{.1} = n_{11} + n_{21} + n_{31} + n_{41} = 9$	11	20	N=40

$$\begin{aligned} n_{\cdot 1} &= n_{11} + n_{21} + n_{31} + n_{41} \\ &= 5 + 4 + 0 + 0 \\ &= 9 \end{aligned}$$

.

A table of marginal frequencies of *variable* X

x_i	$n_{i\cdot}$
30	8
40	16
50	8
60	8
\	$\sum_{i=1}^{4} n_{i\cdot} = 40$

A table of marginal frequencies of *variable* Y

y_j	$n_{\cdot j}$
12	9
14	11
15	20
\	$\sum_{j=1}^{3} n_{\cdot j} = 40$

Example 3.3.2 *Refer to the two-way table from example* 3.2.1.

$X\backslash Y$	*car*	*bus*	*other transport*
Girl	22	38	40
Boy	25	34	41

Find and interpret the marginal frequencies for the two-way table above.
Solution:
Create a new column and a new row for the marginal frequencies. Then add the entries in each row and column.

$X\backslash Y$	car	bus	other transport	**Total**
Girl	22	38	40	**100**
Boy	25	34	41	**100**
Total	**47**	**72**	**81**	**N=200**

variable X

Modalities(x_i)	$n_i.$
Girl	100
Boy	100
\	$\sum_{i=1} n_{i.} =$ 200

variable Y

y_j	$n_{.j}$
car	47
bus	72
other transport	81
\	$\sum_{j=1}^{3} n_{.j} = 200$

3.3.2 Relative Frequency Distribution

Two-way tables can also dispay relative frequencies based on the total number of values or observations.

Joint Relative Frequency

Definition 3.3.3 *A joint Relative Frequency(f_{ij}) is the ratio of a joint frequency to the total number of observations*

$$f_{ij} = \frac{n_{ij}}{N}$$

Marginal relative frequency

Definition 3.3.4 *A marginal relative frequency is the sum of the joint relative frequencies in a row or a column.*

$$f_{i.} = \frac{n_{i.}}{N} \text{ and } f_{.j} = \frac{n_{.j}}{N}$$

Remark 3.3.3 *When finding relative frequencies in a two-way table, you can use the corresponding decimals or percents.*

Example 3.3.3 *Use the above table to find a relative frequency table.*

To find the relative frequencies, divide each frequency by the grand total

$X\backslash Y$	car	bus	other transport	Total
Girl	$\frac{22}{200}$	$\frac{38}{200}$	$\frac{40}{200}$	$\frac{100}{200}$
Boy	$\frac{25}{200}$	$\frac{34}{200}$	$\frac{41}{200}$	$\frac{100}{200}$
Total	$\frac{47}{200}$	$\frac{72}{200}$	$\frac{81}{200}$	$\frac{200}{200} = 1$

2)
(i) Find the marginal relative frequency of the students who prefer car.
(ii) Find the marginal relative frequency of boys.
(iii) Find the marginal relative frequency of girls.
(iv) Find the marginal relative frequency of the students who prefer bus.
Solution :
(i)Divide the total number of students who prefer car by the grand total. Express your answer as a decimal and as a percent.

$$\frac{47}{200} \approx 0.24$$

24% are students who prefer car.
(ii)Divide the total number of boys by the grand total. Express your answer as a decimal and as a percent.

$$\frac{100}{200} = 0.50$$

50% are boys
(iii) Divide the total number of girls by the grand total. Express your answer as a decimal and as a percent.

$$\frac{100}{200} = 0.50$$

50% are girls
(iv) Divide the total number of students who prefer bus by the grand total. Express your answer as a decimal and as a percent.

$$\frac{72}{200} = 0.36$$

36% are students who prefer bus.

3.3.3 Conditional Relative Frequency

A conditional relative frequency describes what portion of a group with a given characteristic also has another characteristic.

Definition 3.3.5 *A conditional relative frequency is the ratio of a joint relative frequency to the marginal relative frequency. You can find a conditional relative frequency using a row total or a column total of a two-way table.*

Remark 3.3.4 *A conditional relative frequency compares a frequency count to the marginal total that represents the condition of interest.*

Example 3.3.4 *Use the data from Example* 3.2.2. *Find the conditional relative frequency that a a person in data prefers bus, given that the person is a boy.*

$$\left(\frac{\textit{Number of boys who prefer bus}}{\textit{Total number of boys}} = \frac{34}{100}\right)$$

34% of the boys in the sample prefer bus.

3.4 Graphical Representation

In this section we restrict to bivariate data description for two quantitative variables. We will make a distinction between two types of variables. where exactly two measurements are made on each observation. The two measurements will be called x and y . Since x and y are obtained for each observation, the data for one observation is the pair (x, y).

A Notation

In general, there are N pairs of such bivariate data given by:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_N, y_N)$$

3.4.1 Scatter plot

Definition 3.4.1 *Scatter plot (scattergram or scatter diagram) is the graph that represents the bivariate data in x and y cartesian plane. A scatter graph of bivariate data is a two-dimensional graph with one variable on one axis, and the other variable on the other axis. We then plot a point (x_i, y_i) for each observation, where the abscissa x is the first data value in the observation and the ordinate y is the second. Generally, the points are not connected to each other.*

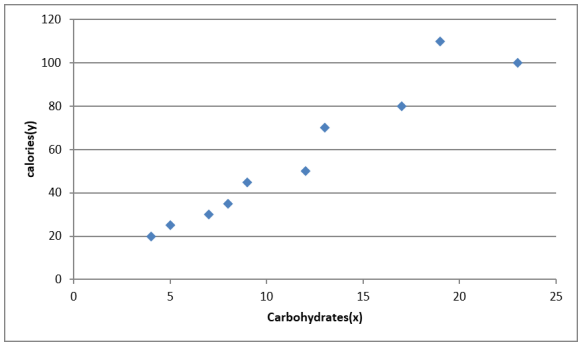
Remark 3.4.1 *A scatter plot is a visual image of the ways in which variables may or may not be related. the scatter plot gives an idea about the appearance, the sense, and the kind of relation between the tow variables.*

Given bivariate quantitative data, we make the scatterplot of this data as follows:

Example 3.4.1 *The table below shows the number of calories and the number of grams of carbohydrates in a half-cup serving of ten different canned or frozen vegetables*

$Carbohydrates(x_i)$	9	23	4	5	19	8	12	7	13	17
$calories(y_i)$	45	100	20	25	110	35	50	30	70	80

Draw a scatter plot on graph paper. Let the horizontal axis represent grams of carbohydrates and the vertical axis represent the number of calories.



A statistical measure of the strength of the relationship between two quantitative variables that takes these factors into account is the subject of the next section

3.5 The covariance

Covariance is a statistical measure that describes the relationship between a pair of random variables where change in one variable causes change in another variable.

Definition 3.5.1 *For a bivariate data set the covariance $Cov(x, y)$ between the variables x and y is:*

$$\begin{aligned} Cov(x,y) &= \overline{xy} - \ \bar{x}\ \bar{y} \\ &= \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j\right) - \bar{x}\ \bar{y} \\ &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \bar{x}\ \right)\left(y_j - \bar{y}\right) \end{aligned}$$

Where
x_i is the values of the x-variable
y_j is the values of the y-variable
$\overline{x}$ is the mean (average) of the x-variable
$\overline{y}$ is the mean (average) of the y-variable
N is the number of data points in the set.

Remark 3.5.1 *Covariance takes any value , where the negative value represents the negative relationship where as a positive value represents the positive relationship. It is used for the linear relationship between variables. It gives the direction of relationship between variables.*

Remark 3.5.2 *Let* $x = \{x_1; ....; x_N\}$ *and* $y = \{y_1; ....; y_N\}$ *are two datasets (two samples) with means* $\overline{x}$ *and* $\overline{y}$ *respectively, then the variables* x *and* y *is:*

$$Cov(x,y) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x}\ )(y_k - \bar{y}) \text{ or } Cov(x,y) = \overline{xy} - \ \bar{x}\ \bar{y} = \frac{1}{N}\sum_{k=1}^{N} x_k y_k - \ \bar{x}\ \bar{y}$$

3.5.1 Properties

Let X, Y be jointly distributed random variables. From the definition of covariance, we derive the following:
(1) The covariance generalizes variance: $Cov(X;X) = Var(X)$.
(2) The covariance is symmetric: $Cov(X;Y) = Cov(Y;X)$.
(3) For any fixed scalars $a; b; c; d \in \mathbb{R}$, $Cov(aX + b; cY + d) = acCov(X;Y)$.
(4) X and Y are independent $\Rightarrow Cov(X;Y) = 0$(the reverse implication does not always hold)
(5)$|Cov(X;Y)| \leq \sigma_X \sigma_Y$

3.6 The correlation coefficients

The scatter diagram provides a visual impression of the nature of the relation between the x and$y$ values in a bivariate data set. In a great many cases, the points appear to band around a straight line. Our visual impression of the closeness of the

scatter to a linear relation can be quantified by calculating a numerical measure, called the correlation coefficient.
The correlation coefficient, denoted by r, is a measure of strength of the linear relation between the x and y variables.

Definition 3.6.1 *The correlation coefficient (sometimes called Pearson's correlation coefficient) between two variables x and y is given by the formula:*

$$r = corr(x, y) = \frac{Cov(x, y)}{\sigma_x \times \sigma_y} = \frac{Cov(x, y)}{\sqrt{V(x) \times V(y)}}$$

where
$r = corr(x, y)$ is the correlation between the variables x and y.
$Cov(x, y)$ is the covariance between the variables x and y
σ_x is the standard deviation of the x-variable.
σ_y isthe standard deviation of the y-variable.

Remark 3.6.1 *A linear correlation coefficient r that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables.*

Example 3.6.1 *Consider the following data:*

x	2	2	3	4	5	6	7	7,6
y	14	26	31	29	44	40	54	50

(i) Compute $Cov(x; y)$.
(ii) Calculate the linear correlation coefficient $r = corr(x, y)$ between x and y.
Solution:
(i) Compute $Cov(x; y)$:

Determine : $\bar{x}, V(x), \bar{y}$ and $V(y)$

$$\begin{aligned}
\bar{x} &= \frac{\sum x}{n} = \frac{2+2+3+4+5+6+7+(7,6)}{8} = 4,575 \\
V(x) &= \overline{x^2} - \bar{x}^{\,2} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 \\
&= \frac{2^2+2^2+3^2+4^2+5^2+6^2+7^2+(7,6)^2}{8} - (4,575)^2 \approx 4,16 \\
\bar{y} &= \frac{\sum y}{N} = 36 \\
V(y) &= \overline{y^2} - \bar{y}^2 = \frac{\sum y^2}{N} - \left(\frac{\sum y}{N}\right)^2 = \frac{11626}{8} - (36)^2 = 157,25
\end{aligned}$$

Then,

$$\begin{aligned}
Cov(x,y) &= \overline{xy} - \bar{x}\,\bar{y} = \frac{\sum xy}{N} - \left(\frac{\sum x}{N}\right)\left(\frac{\sum y}{N}\right) \\
&= \frac{2\times 14 + 2\times 26 + 3\times 31 + 4\times 29 + 5\times 44 + 6\times 40 + 7\times 54 + 7,6\times 50}{8} - 4,575\times\ 36 \\
&= \frac{1503}{8} - 164,7 = 23,675
\end{aligned}$$

and the linear correlation coefficient of this data set is:

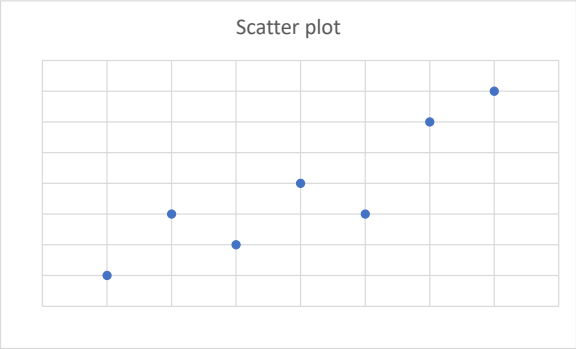
$$r = \frac{Cov(x,y)}{\sqrt{V(x)\times V(y)}} = \frac{23,675}{\sqrt{4,16\times 157,25}} = 0.92$$

Exercise 1 *For the following data:*

x	1	2	3	4	5	6	7
y	0.5	2.5	2	4	3.5	6	5.5

1) Display the scatter plot.
2)Calculate the correlation coefficient between x and y.

Solution



x	1	2	3	4	5	6	7	$\sum x = 28$
y	0.5	2.5	2	4	3.5	6	5.5	$\sum y = 24$
x^2	1	4	9	16	25	36	49	$\sum x^2 = 140$
y^2	0.25	6.25	4	16	12.25	36	30.25	$\sum y^2 = 105$
xy	0.5	5	6	16	17.5	36	38.5	$\sum xy = 119.5$

The correlation coefficient of this data set is:

$$r = \frac{Cov(x,y)}{\sqrt{V(x) \times V(y)}}$$

$$V(x) = \overline{x^2} - \bar{x}^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2 = \frac{140}{7} - \left(\frac{28}{7}\right)^2 = 4$$

$$V(y) = \overline{y^2} - \bar{y}^2 = \frac{\sum y^2}{N} - \left(\frac{\sum y}{N}\right)^2 = \frac{105}{7} - \left(\frac{24}{7}\right)^2 = 3.3$$

$$Cov(x,y) = \overline{xy} - \bar{x}\,\bar{y} = \frac{\sum xy}{N} - \left(\frac{\sum x}{N}\right)\left(\frac{\sum y}{N}\right) = \frac{119.5}{7} - \left(\frac{28}{7}\right)\left(\frac{24}{7}\right) = 3.39$$

Then,

$$r = \frac{Cov(x,y)}{\sqrt{V(x) \times V(y)}} = \frac{3.39}{\sqrt{4 \times 3.3}} = 0.93$$

3.6.1 Types of correlation

There are three types of correlation:
• Negative correlation
• Zero correlation
• Positive correlation

3.6.2 Properties

The correlation coefficient satisfies the following properties:
1) The correlation coefficient is a number in the interval $[-1, 1]$.$(-1 \leq r \leq 1)$
2)The sign of r indicates the direction of the correlation between x and y .
i) Positive values denote positive linear correlation.
ii) Negative values denote negative linear correlation.
3) If $r = 0$, there is no linear correlation.
4)If $r = 1$or $r = -1$, there is perfect correlation and the line on the scatter plot is increasing or decreasing respectively.

3.7 Regression line and Mayer Line

3.7.1 Regression line

Definition 3.7.1 *A regression line is a line that best represents all of the points in a scatterplot.*

Remark 3.7.1 *Such a regression line (a straight line) is called the "line of best fit".*

- Fitting a straight line to a set of paired observations $(x_1, y_1), (x_2, y_2), , ..., (x_{_N}, y_{_N})$.

The equation of **a regression line** relating these two variables is:

$$y = ax + b$$

where b is the intercept, and a is the slope.

3.7.2 Mayer Line Method

This method helps us calculate the line of best fit for a scatter plot.

Steps for finding the regression line equation: $y = ax + b$**:**

1. Organise coordinates in a table in order according to the x-coordinates.
2. Divide into two equal groups (if possible)
3. Find the mean of the x and y coordinates for each group. $M_1(\overline{x_1}, \overline{y_1})$ and $M_2(\overline{x_2}, \overline{y_2})$.(M_1 and M_2 are called the mean points)
4. Find the equation ($y = ax + b$) of the Regression line passing through the point M_1 and M_2.

Example 3.7.1 *Determine the equation of the straight line passing through the mean points (Mayer Line of Best Fit)*

x	6	7	10	13	14	15	18	19	23	25
y	23	26	39	44	48	55	50	65	68	72

Calculate the mean of the x-values and the mean of the y-values. From this you will create two new points called the mean points $M_1(10, 36)$ and $M_2(20, 62)$
Regression Line: $y = ax + b$

- Find slope a :

$$a = \frac{62 - 36}{20 - 10} = 2.6 \text{ and } b = 10$$

- Regression Line is: $y = 2.6x + 10$

Remark 3.7.2 *A regression line allows you to predict values that may not be in the original data set.*

Example: if $x = 28$, $y =$?

$$y = 2.6\,(28) + 10 = 82.8$$

Example 3.7.2 *Find the regression line equation (Mayer's Method):*

x	1	5	8	3	5
y	4	12	13	6	11

Step 1 - Make sure the data is in increasing order based on x values.

x	1	3	5	5	8
y	4	6	11	12	13

Step 2 - Divide into two equal groups
Group 1

x	1	3	5
y	4	6	11

Group 2

x	5	5	8
y	11	12	13

Step 3 - Find points $M_1$and M_2 by calculating the mean of each group:

- Find point M_1:

x-coordinate $= \dfrac{1+3+5}{3} = 3$
y-coordinate $= \dfrac{4+6+11}{3} = \dfrac{21}{3} = 7$
Then, $M_1(3, 7)$

- Find point M_2:

$x-$coordinate $= \dfrac{5+5+8}{3} = \frac{18}{3} = 6$
$y-$coordinate $= \dfrac{11+12+13}{3} = \frac{36}{3} = 12$
Then, $M_2(6, 12)$
Step 4 - Find the equation of the line ($y = ax + b$) that goes through points M_1 and M_2:

- Slope:

$$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{12-7}{6-3} = \frac{5}{3}$$

- Find the "b":

by substituting any one of your points in the equation:

$$y = \frac{5}{3}x + b$$

Use point $M_1(3, 7)$:

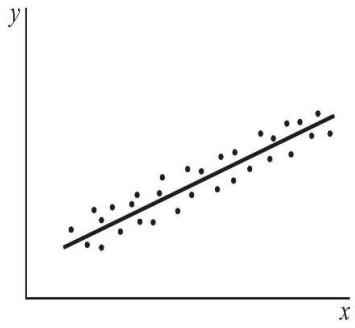
$$\begin{aligned} 7 &= (\frac{5}{3})(3) + b \\ b &= 2 \end{aligned}$$

The equation of the regression line is:

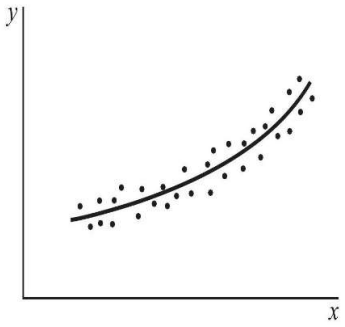
$$y = \frac{5}{3}x + 2$$

3.8 Regression curve and correlation

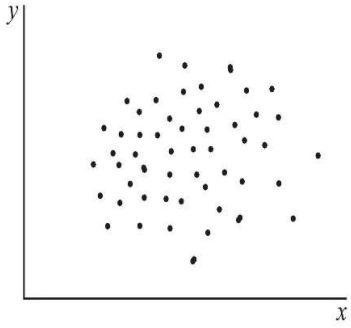
One of the primary goals of curve fitting is to estimate one variable (the dependent variable) based on the other (the independent variable). This estimation process is commonly known as regression. When y is estimated from x using an equation, the equation is referred to as the regression equation of y on x, and the corresponding curve is called the regression curve of y on x.
From a scatter diagram, it is often possible to visually identify a smooth curve that approximates the data. This curve is known as an approximating curve.



Linear relation ship



Non linear relation ship



No relation ship

The relationship between x and y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. This means that te relationship between x and y is in the form of a straight line and is termed linear regression.
If $(r = \pm 1)$we say that there is perfect linear correlation (perfect linear regression)
If $(r = 0)$ we say that there is bad regression.

3.9 Function Fitting

The general problem of finding equations of approximating curves that fit given sets of data is called curve fitting. In practice the type of equation is often suggested from the scatter diagram. For Linear relationship, we could use a straight line $(y = ax + b)$, while for Nonlinear relationship, we could try a parabola or quadratic curve $(y = c + ax + bx^2)$ where a, b and c are constants.

3.9.1 Least Squares

Linear regression models typically employ a least-squares method to find the line of best fit. This technique involves minimizing the sum of squared differences, where each square is calculated by squaring the distance between a data point and the regression line or the mean value of the dataset.
The least-squares line approximating the set of points $(x_1, y_1), ..., (x_n, y_n)$ has the equation: $y = ax + b$
The constant a is the slope of the line $(y = ax + b)$, is the fundamental constant in determining the line. It is also seen that the least-squares line passes through the point $(\overline{x}, \overline{y})$ which is called the centroid or center of gravity of the data.

Fitting a straight line- method of least squares

Let line of best fit be given by:

$$y = ax + b \qquad ((1))$$

Normal equations are given by:

$$\sum y = a \sum x + bN \qquad (2)$$

and

$$\sum xy = a \sum x^2 + b \sum x \qquad (3)$$

Calculating $\sum x; \sum y; \sum xy$, and $\sum x^2$
Values of a and $b$are obtained by solving equations (2) and (3).
The problem is to determine a and b so that the straight line $y = ax + b$ is the line of the best fit.

Example 3.9.1 *Consider*

x	3	5	6	8	9	11
y	2	3	4	6	5	8

We have

x	3	5	6	8	9	11	$\sum x = 42$
y	2	3	4	6	5	8	$\sum y = 28$
x^2	9	25	36	64	81	121	$\sum x^2 = 336$
xy	6	15	24	48	45	88	$\sum xy = 226$

Then,

$$\begin{cases} 42.a + 6.b = 28 \\ \text{and} \\ 336.a + 42.b = 226 \end{cases} \Rightarrow \begin{cases} a = \frac{5}{7} \\ \\ b = -\frac{1}{3} \end{cases}$$

Therefore, the line of the best fit is:

$$y = \frac{5}{7}x - \frac{1}{3}$$

3.9.2 Regression line

Regression line by the method of least squares

Regression line of y on x

The y on x regression line equation is represented as follows:

$$y = ax + b$$

where

$$a = \frac{Cov(x, y)}{V(x)} \text{ and } b = \overline{y} - a\overline{x}$$

Regression line of x on y

The line of regression of x on y is :

$$x = a'y + b'$$

where

$$a' = \frac{Cov(x, y)}{V(y)} \text{ and } b' = \overline{x} - a\overline{y}$$

Examples:(example 3.6.1)
1)The y on x regression line equation is :

$$y = ax + b$$

where

$$a = \frac{Cov(x,y)}{V(x)} = \frac{23,675}{4,16} = 5,69$$

and

$$b = \overline{y} - a\overline{x} = 36 - 5,69 \times 4,575 = 9,97$$

Then, The y on x regression line equation is :

$$y = 5,69x + 9,97$$

2) The line of regression of x on y is :

$$x = a'y + b'$$

where

$$\begin{aligned} a' &= \frac{Cov(x,y)}{V(y)} = \frac{23,675}{157,25} = 0,15 \\ \text{and } b' &= \overline{x} - a\overline{y} = 4,575 - 0,15 \times 36 = -0,825 \end{aligned}$$

Then, The y on x regression line equation is :

$$x = 0.15y - 0,825$$

Remark:
Equation of line of regression of y on x is given by :

$$(y - \overline{y}) = a(x - \overline{x})$$

where

$$a = \frac{Cov(x,y)}{V(x)}$$

Similarly line of regression of x on y is given by:

$$(x - \overline{x}) = a'(y - \overline{y})$$

where

$$a' = \frac{Cov(x,y)}{V(y)}$$

3.10 Exercises

Exercise N°1:
Eighty students were surveyed about playing an instrument. The results are shown in the two-way frequency table.

Gender\Play an Instrument	Yes	No	Total
Male	20	15	/
Female	28	17	
Total	/		

1)Complete the two-way relative frequency table for the data.
2)What percent of the students surveyed play an instrument? Identify what type of frequency each percent is.
3)What percent of students surveyed who are female?
4) Find the conditional relative frequency that a student surveyed is a female, given that she plays an instrument.
5) Is there an association between the sex of a student and whether the student plays an instrument? Explain.

Exercise N°2:
I)The table below shows the coordinates of a scatter plot:

x	1	5	6	6	7	7	10	12	15	16
y	15	14	13	11	12	9	7	6	6	2

1) Draw a scatter plot.
2) By the method of Mayer, find a regression line.
II) Consider the following data:

x	60	61	62	63	65
y	3.1	3.6	3.8	4	4.1

By the method of Mayer,find the regression line of y on x and deduce the approximated value of y when $x = 64$.

Exercise N°3:
Consider the following data:

x	3	5	6	8	9	11
y	2	3	4	6	5	8

1) Represent the data in a scatter plot.
2) Calculate the correlation coefficient between x and y.
3)Fit the least squares regression (regression line of y on x) for these data.
4) Estimate the value of y for $x = 4$, $x = 7$, $x = 16$

Exercise N°4

1) Find both regression lines of y on x for the following data:

x	5	4	4	6	7	3	5	6
y	3	5	3	6	6	3	4	2

2) Estimate the value of y when x is given to be 10(for regression lines of y on x).

3.11 Solutions

Exercise N°1:

Two-way frequency table:

Gender\Play an Instrument	Yes	No	Total
Male	20	15	35
Female	28	17	45
Total	48	32	80

1)Two-way relative frequency table:

Gender\Play an Instrument	Yes	No	Total
Male	$\frac{20}{80} = 0.25$	$\frac{15}{80} = 0.1875$	$\frac{35}{80} = 0.4375$
Female	$\frac{28}{80} = 0.35$	$\frac{17}{80} = 0.2125$	$\frac{45}{80} = 0.5625$
Total	$\frac{48}{80} = 0.60$	$\frac{32}{80} = 0.40$	$\frac{80}{80} = 1$

2) 60% of the students surveyed play an instrument. This is a marginal relative frequency.

3)Percent of students surveyed who are female: is 56.25%

4)The conditional relative frequency that a student surveyed is a female, given that she plays an instrument.

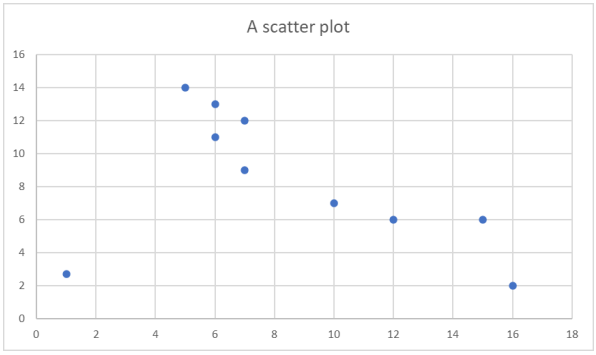
$$\frac{28}{48} \approx 0.583$$

This conditional relative frequency indicates that approximately 58.3% of females plays an instrument.

5) Females make up 58.3% of students who play an instrument, but they only make up 56.25% of the students surveyed. So, females are more likely to play an instrument than males.

Exercise N°2:

I)1) Draw a scatter plot.



2)

Step 1 - Make sure the data is in increasing order based on x values.

x	1	5	6	6	7	7	10	12	15	16
y	15	14	13	11	12	9	7	6	6	2

Step 2: Divide into 2 equal groups (which is possible for this data)

Group 1 :

x	1	5	6	6	7
y	15	14	13	11	12

Group 2 :

x	7	10	12	15	16
y	9	7	6	6	2

Step 3 : Find points $M_1$and M_2 by calculating the mean of each group:

- Find point M_1:

$$x\text{-coordinate} = \frac{1+5+6+6+7}{5} = \frac{25}{5} = 5$$

$$y\text{-coordinate} = \frac{15+14+13+11+12}{5} = \frac{65}{5} = 13$$

Then,

$$M_1(5, 13)$$

- Find point M_2:

$$x-\text{coordinate} = \frac{7+10+12+15+16}{5} = \frac{60}{5} = 12$$
$$y-\text{coordinate} = \frac{9+7+6+6+2}{5} = \frac{30}{5} = 6$$

Then,

$$M_2(12, 6)$$

Step 4: Find the equation of the line ($y = ax + b$) that goes through points M_1 and M_2:

- Find **slope a** :

$$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{6-13}{12-5} = \frac{-7}{7} = -1$$

- Find the "b":

by substituting any one of your points in the equation:

$$y = -1x + b$$

Use point $M_1(5, 13)$:

$$\begin{aligned} 13 &= (-1)(5) + b \\ 13 &= -5 + b \\ 18 &= b \end{aligned}$$

Thus, the equation of the regression line is:

$$y = -x + 18$$

II) 1°
Find the regression line of y on x (method of Mayer)
Consider the following Groups:

Group1

x	60	61	62
y	3.1	3.6	3.8

and Group 2

x	62	63	65
y	3.8	4	4.1

- For each group, calculate the mean of the x-values and the mean of the y-values. From this you will create two new points called the mean points $M_1(61, 3.5)$and $M_2(\frac{190}{3}, 4.1)$.
- Determine the equation $(y = ax + b)$ of the straight line passing through the mean points M_1 and M_2(Mayer Line of Best Fit).

Equation of regression Line is:

$$y = \frac{18}{70}x - \frac{853}{70}$$

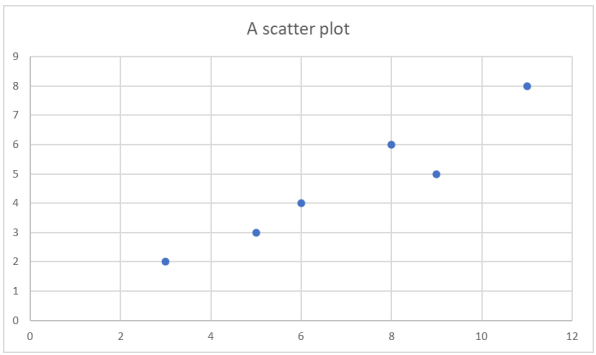
2°)Deduce the approximated value of y when $x = 64$:
Now that we have the regression Line Equation:

$$y = \frac{18}{70}x - \frac{853}{70}$$

If $x = 64$, we have:

$$y = \frac{18}{70}(64) - \frac{853}{70} = 4.27$$

Exercise N°3:
1)



2) Calculate the correlation coefficient between x and y.
The correlation coefficient is:

$$r = \frac{Cov(x, y)}{\sqrt{V(x) \times V(y)}}$$

- Calculating $\sum x, \sum y, \sum x^2, \sum y^2$ and $\sum xy$:

x	3	5	6	8	9	11	$\sum x = 42$
y	2	3	4	6	5	8	$\sum y = 28$
x^2	9	25	36	64	81	121	$\sum x^2 = 336$
y^2	4	9	16	36	25	64	$\sum y^2 = 154$
xy	6	15	24	48	45	88	$\sum xy = 226$

$$V(x) = \overline{x^2} - (\overline{x})^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2 = \frac{336}{6} - \left(\frac{42}{6}\right)^2 = 7$$

$$V(y) = \overline{y^2} - \bar{y}^2 = \frac{\sum y^2}{N} - \left(\frac{\sum y}{N}\right)^2 = \frac{154}{6} - \left(\frac{28}{6}\right)^2 = 3.88$$

$$Cov(x,y) = \overline{xy} - \bar{x}\,\bar{y} = \frac{\sum xy}{N} - \left(\frac{\sum x}{N}\right)\left(\frac{\sum y}{N}\right) = \frac{226}{6} - \left(7 \times \frac{28}{6}\right) = 5$$

Then,

$$r = \frac{Cov(x,y)}{\sqrt{V(x) \times V(y)}} = 0.96$$

Hence, coefficient of correlation between x and y is $r = 0.96$.
The value of $r = 0.96$ confirms positive relation noticed in the scatter plot of data.
3)The regression line of y on x:

- Let line of regression of y on x be represented by: $y = ax + b$

where

$$a = \frac{Cov(x,y)}{V(x)} = \frac{5}{7}$$

and

$$b = \overline{y} - a\overline{x} = \frac{28}{6} - \frac{5}{7} \times 7 = -\frac{1}{3}$$

Then, The y on x regression line equation is :

$$y = \frac{5}{7}x - \frac{1}{3}$$

4) Estimate the value of y for $x = 4$, $x = 7$, $x = 16$
For $x = 4$:

$$y = \frac{5}{7}(4) - \frac{1}{3} = \frac{20}{7} - \frac{1}{3} = \frac{53}{21}$$

Similarly,for $x = 7$, $x = 16$
Remark:

- **Method** 2 :find equation of line of regression of y on x

Equation of line of regression of y on x is given by :

$$(y - \overline{y}) = a(x - \overline{x})$$

where

$$a = \frac{Cov(x, y)}{V(x)} = \frac{5}{7}$$

Thus, line of regression of y on x is

$$\left(y - \frac{28}{6}\right) = \frac{5}{7}(x - 7)$$

$$y = \frac{5}{7}x - \frac{1}{3}$$

2) If $x = 10$,

$$y = \frac{5}{7}(10) - \frac{1}{3} = \frac{143}{21}$$

Part II

Probability

Chapter 4

Combinatory analysis

4.1 Introduction

Combinatorial analysis is the branch of mathematics that focuses on determining the number of distinct ways to arrange objects. Its main objective is to count the various ways of grouping the elements of a set E with n (where $n \in \mathbb{N}^*$) elements, according to specific rules.

4.2 Arrangements

In all that follows, we consider E a set with $Card\ E = n$ ($n \in \mathbb{N}^*$) and p a non-zero positive integer satisfying $1 \leq p \leq n$.

4.2.1 Arrangement

Definition 4.2.1 *An arrangement of p elements of E is called any ordered sequence of p among the n elements of E. The number of arrangements is denoted by: A_n^p.*

4.2.2 Types of arrangement

Arrangement can be classified in tow different categories:

1)Arrangement with repetition (where repetition is allowed)
2)Arrangement without repetition (when repetition is not allowed)

Arrangement with repetition

Definition 4.2.2 *When an element can be chosen or observed several times, the number of arrangements with repetition of p elements from a set of n elements is:*

$$A_n^p = n^p.$$

Example 4.2.1 *How many 3-letter words can be formed from the 26 letter alphabet?*

Remark 4.2.1 *By a word here we do not mean a word that is necessarily in the dictionary, just any combination of the alphabets. Examples of "words" would be* ***MNO, MON,ONM ,MMO****...etc*

$$26 \times 26 \times 26 = 26^3 = 17576.$$

Example 4.2.2 *How many* 3 *letter words with or without meaning can be formed out of the letters of the word "**mother**" when repetition of words is allowed?*

The number of objects, in this case, is 6, as the word "**mother**" has 6 alphabets and $r = 3$, as 3-letter word has to be chosen.
Thus, the arrangement will be:
Arrangement (when repetition is allowed)$= 6^3 = 216$

Arrangement without repetition

Definition 4.2.3 *When an element can be chosen or observed only once, the number of non-repetition arrangements of p elements from a set of n elements is:*

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

Example 4.2.3 *How many different* 4*-letter words can be formed from the letters of the word "**MOTHER**" .*

There are 6 different letters in the word "**MOTHER**", and we are to choose 4-letters to form a word, therefore, we have $A_6^4 = \frac{6!}{2!} = 360$ words.

Example 4.2.4 *How many different 7-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers?*

- Where repetition is allowed:

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 \times 10 = 175.760.000$$

- If repetition among letters or numbers is prohibited (No repetition among letters or numbers).

$$26 \times 25 \times 24 \times 10 \times 9 \times 8 \times 7 = 78.624.000$$

4.3 Permutations

Example: How many different **ordered arrangements** of **a,b,c**?
Solution: **abc; acb; bac; bca; cab; cba**
In the above example, each ordering is known as a permutation. By the principle of counting, there are

$$3 \times 2 \times 1 = 6$$

different permutations.

4.3.1 Permutation

Definition 4.3.1 *A permutation of n objects is an ordered sequence of those n objects.*

The number of permutations is denoted by:

$$P_n$$

Remark 4.3.1 *In a permutation, the order that we arrange the objects in is important.*

Number of ways to order n distinct elements:

Based on the reasoning mentioned earlier, if we have n distinct elements, there would be $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ ways of ordering them. We denote this quantity by $n!$ and in words by n factorial.

$$\begin{aligned} P_n &= n \times (n-1) \cdots \times 2 \times 1 \\ P_n &= n! \end{aligned}$$

Notation: We use $n!$ to denote the number of permutations of n objects, so

$$n! = n \times (n-1) \cdots \times 2 \times 1 = \prod_{j=1}^{n} j$$

By convention, we define $0! = 1$.

- We read $n!$ as "n factorial," so n factorial is

$$n! = n \times (n-1) \cdots \times 2 \times 1$$

Example 4.3.1 *How many different words can we form by using the letters of the word "**MOTHER**"?*

That all six(6) of the letters in the word "**MOTHER**" are different. We may use any of the 6 letters to be the first letter of the word, so for the first letter of the word we have 6 choices. On the second letter of the word we have 5 choices left,on the third letter we have 4 choices left, on the fourth letter we have 3 choices left, then we have 2 choices for the fifth letter, then only 1 choice for the last letter.
Using the counting principle we have :$6 \times 5 \times 4 \times 3 \times 2 \times 1 = 6! = 120$ different words that can be formed.
Formula: $P_6 = 6! = 720$

Example 4.3.2 *Consider arranging 3 balls, Red, Blue, Grey- 3 letters: R, B, G. How many ways can this be done?*

Permutations: RBG, RGB, BRG, BGR, GBN, GBR(6 possibilities)
Formula: $P_3 = 3! = 6$

Number of ways to order n elements (some of which are not distinct):

Definition 4.3.2 *When an element exists k at a times, the number of permutations with repetition of n elements from a set of n elements is equal to:*

$$P_n = \frac{n!}{k!}.$$

In case each element x_i of the set E exists k_i at a times such that:
$1 \leq i \leq m$ and $\sum_{i=1}^{m} k_i = k_1 + k_2 + ... + k_m = n$,

$$\boxed{E = \left\{ \underbrace{x_1 ... x_1}_{k_1 \ times} \underbrace{x_2 ... x_2}_{k_2 times} ..... \underbrace{x_m ... x_m}_{k_m \ times} \right\}},$$

the number of permutations is equal to:

$$P_n = \frac{n!}{k_1! \times k_2! \times ... \times k_m!}.$$

Example 4.3.3 *How many different letter arrangements can be formed from the letters PEPPER?*

$$\boxed{\{P, E, P, P, E, R\} = \left\{ \underbrace{P, P, P}_{3 \ times} \ \underbrace{E, E}_{2 \ times} \ \underbrace{R}_{1 \ times} \right\}}$$

- Assuming for the moment the three *P's* and two *E's* were distinguishable.

Then, we would have $6! = 720$ ways to permute them.
Since the three *P's* and two *E's* are actually indistinguishable, we remove these two sets of duplicates from the earlier permutations.
Therefore, the number of distinct letter arrangements is:

$$P_6 = \frac{6!}{2!3!} = 60$$

4.4 Combinations

4.4.1 Combination

Definition 4.4.1 *A combination of p objects among n objects is non ordered subset of p objects.*

Remark 4.4.1 *An arrangement of a set of objects selected without regard to their order is called a combination of the objects.*

There are two types of combinations: combinations without discounts and combinations with discounts.

Combination without discount

Definition 4.4.2 *We call combination without replacement any set of p elements taken without replacement from the n elements of E. The number of combinations of p objects among n objects is the number of combination without replacement is:*

$$C_n^p = \frac{n!}{p! \times (n-p)!}.$$

Remark 4.4.2 *The notation C_n^p is rarely used; instead we use $\binom{n}{p}$, pronounced n choose p"*

Example 4.4.1 *Form a delegation of 3 employees within a company of 50 employees.*

The number of delegations is:

$$C_{50}^3 = \frac{50!}{3! \times 47!} = \frac{50 \times 49 \times 48 \times 47!}{6 \times 47!} = 19600.$$

Combination with discount

Definition 4.4.3 *We call combination with replacement any set of p elements taken with replacement from the n elements of E. The number of combination with replacement is:*

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

Example 4.4.2 *There are five*(5) *colored balls in a pool. All balls are of different colors. In how many ways can we choose four*(4) *pool balls?(with discount)*

The order in which the balls can be selected does not matter in this case. The selection of balls can be repeated.
Use the following formula to get the number of way which the four pool balls can be chosen.

$$C_{5+4-1}^{4} = C_{8}^{4} = \frac{8!}{4! \times 4!} = \frac{8 \times 7 \times 6 \times 5 \times 4!}{(4 \times 3 \times 2 \times 1) \times 4!} = 70.$$

Hence, the pool balls can be selected in 70 different ways.

Remark 4.4.3 *1)When the order doesn't matter, it is a Combination.*
2)When the order does matter it is an arrangement.

Examples:

1) Choosing 3 desserts from a menu of 10.

- Combination: $C_{10}^{3} = 120$.

2) Listing your 3 favorite desserts, in order, from a menu of 10.

- Arrangement: $A_{10}^{3} = 720$.

Proposition 4.4.1 *Combination properties"*

1/

$$C_{n}^{0} = C_{n}^{n} = 1.$$

2/

$$C_{n}^{1} = C_{n}^{n-1} = n.$$

3/

$$C_{n}^{2} = C_{n}^{n-2} = \frac{n\,(n-1)}{2}.$$

4/

$$C_{n}^{p} = C_{n}^{n-p}.$$

5/

$$C_{n-1}^{p-1} + C_{n-1}^{p} = C_n^p.$$

Proof: To demonstrate the above properties, we use the cobinination formula given in definition

1/

$$C_n^0 = \frac{n!}{0! \times n!} = \frac{n!}{n! \times 0!} = C_n^n = 1.$$

2/

$$C_n^1 = \frac{n!}{1! \times (n-1)!} = \frac{n!}{(n-1)! \times 1!} = C_n^{n-1} = n.$$

3/

$$C_n^2 = \frac{n!}{2! \times (n-2)!} = \frac{n!}{(n-2)! \times 2!} = C_n^{n-2} = \frac{n\,(n-1)}{2}.$$

4/

$$C_n^p = \frac{n!}{p! \times (n-p)!} = \frac{n!}{(n-p)! \times p!} = C_n^{n-p}.$$

5/

$$\begin{aligned} C_{n-1}^{p-1} + C_{n-1}^{p} &= \frac{(n-1)!}{(p-1)! \times (n-p)!} + \frac{(n-1)!}{p! \times (n-1-p)!} \\ &= \frac{(n-1)!\,[p + (n-p)]}{p! \times (n-p)!} = \frac{n!}{p! \times (n-p)!} = C_n^p. \end{aligned}$$

4.5 Binomial coefficients and Pascal's Triangle

The expansion of the polynomial

$$(a+b)^n = \underbrace{(a+b) \times (a+b) \times ... \times (a+b)}_{n\ times}$$

requires calculating for each term a^k (or b^k) with $0 \leq k \leq n$, the number of distinct ways of choosing k times a (or b) among n possibilities. This number is the binomial coefficient given by:

$$C_n^k = \frac{n!}{k! \times (n-k)!}$$

and we have

$(a+b)^0 =$	1										
$(a+b)^1 =$	$1 \times a$	$+$	$1 \times b$								
$(a+b)^2 =$	a^2	$+$	$2 \times a \times b$	$+$	b^2						
$(a+b)^3 =$	a^3	$+$	$3 \times a^2 \times b$	$+$	$3 \times a \times b^2$	$+$	b^3				
$(a+b)^4 =$	a^4	$+$	$4 \times a^3 \times b$	$+$	$6 \times a^2 \times b^2$	$+$	$4 \times a \times b^3$	$+$	b^4		
.	.		.	.	.						
.	.		.	.	.						
	.		.	.	.						
$(a+b)^n =$	a^n	$+$	$C_n^1 \times a^{n-1} \times b$	$+$	$C_n^2 \times a^{n-2} \times b^2$	$+$	....	..	..	$+$	$C_n^n \times b^n$

Binomial coefficients form a triangle whose lines correspond to constant n, so the sum of two consecutive coefficients in one line is equal to the term of the next line below the second term.

4.5.1 Binomial Theorem

Theorem 4.5.1 *Binomial Theorem is as follows:*

$$(a+b)^n = \sum_{k=0}^{n} C_n^k \times a^{n-k} \times b^k = \sum_{k=0}^{n} \binom{n}{k} \times a^{n-k} \times b^k$$

Example 4.5.1 *Pascal's triangle in case* $n = 3$.

$$\begin{array}{cccc} 1 & & & \\ 1 & 1 & & \\ 1 & 2 & 1 & \\ 1 & 3 & 3 & 1 \end{array}$$

Example 4.5.2 *Expand* $(a+b)^3$.

$$\begin{aligned} (a+b)^3 &= \binom{3}{0} a^3 b^0 + \binom{3}{1} a^2 b^1 + \binom{3}{2} a^1 b^2 + \binom{3}{3} a^0 b^3 \\ &= a^3 + 3a^2 b + 3ab^2 + b^3 \end{aligned}$$

Chapter 5

Introduction to probability

5.1 Definitions

5.1.1 Set Definitions

- A set is a well-defined collection of objects.
- Each object in a set is called an **element** of the set.
- Two sets are equal if they have exactly the same elements in them.
- A set that contains no elements is called a null set or an empty set.
- If every element in Set A is also in Set B, then Set A is a subset of Set B.
- We call the partition of S the set of all the subsets of S, It is denoted by $\mathcal{F} = \mathcal{P}(S)$.

5.1.2 Set Notation

- A set is usually denoted by a capital letter, such as A, B, or C.
- An element of a set is usually denoted by a small letter, such as x, y, or z.
- The null set, or empty set, denoted by $\emptyset$, contains no elements at all. The empty set is a subset of any set.

5.1.3 Experiment

Definition 5.1.1 *An experiment(or statistical experiment) is a procedure which results in an outcome. In general, the outcome of an experiment depends on the conditions under which the experiment is performed.*

5.1.4 Random Experiment

Definition 5.1.2 *In a random experiment, the outcome is subject to chance. This means that if the experiment is repeated, the outcome may be different from time to time.*

Example 5.1.1 *Experiment 1:Consider an experiment involving a single coin toss. There are two possible outcomes, heads (H) and tails (T).*

Example 5.1.2 *Experiment 2:Toss a die and observe the number that appears on the upper face.*

Example 5.1.3 *Experiment 3:Record a person's blood type.*

Example 5.1.4 *Experiment4: Measuring the height (cms) of a girl on her first day at school.*

5.1.5 Sample Space

Sample Space

Definition 5.1.3 *The sample space* ***(or The probability space)*** *S of a random experiment is the set of all possible outcomes of the experiment.*

- A sample space is discrete if it consists of a finite or countable infinite set of outcomes.
- A sample space is continuous if it contains an interval of real numbers.

Remark:
Each **outcome** in a sample space is called an **element** or a **member** of the sample space. It is also called sample point(**A simple event)**.
Refer to the preceding examples:

Example 5.1.5 ***Experiment*** *1:, there are only two outcomes for tossing a coin, and the sample space is:*

$$S = \{heads, tails\} \text{ or } S = \{H, T\}$$

Consider another experiment :

Example 5.1.6 *If we toss a coin* ***three times****, then the sample space is:*

$$S = \{HHH, HHT, HTH, THH, HTT, TTH, THT, TTT\}$$

Example 5.1.7 ***Experiment*** *2:When the die is tossed once, there are six possible outcomes. The sample space of the experiment is* $S = \{1; 2; 3; 4; 5; 6\}$*, which contains* $n = 6$ *equally likely outcomes.*

Example 5.1.8 ***Experiment*** *3:The four possible outcomes are: "Blood type A, Blood type B, Blood type AB, Blood type O " (and Rh factor)these simple events:*

The sample space is:

$$S = \{A^+, A^-, B^+, B^-, AB^+, AB^-, O^+, O^-\}$$

Example 5.1.9 ***Experiment*** *4:Sample space* $S = \mathbb{R}$ *the set of all possible real numbers.*

5.1.6 Event

Definition 5.1.4 *An event is a subset of a sample space* S*.*

Remark 5.1.1 $A \subseteq S$

Refer to the preceding Examples:

Example 5.1.10 *(Experiment 1)There are only two outcomes for tossing a coin, and the sample space is*

$$S = \{heads, tails\} \text{ or } S = \{H, T\}$$

and the events are:

$$\{H, T\}, \{H\}, \{T\}, \emptyset$$

Example 5.1.11 *We can defin the events A, B and C for the die tossing experiment 2 by:*

A:" Observe an odd number"
B:" Observe a number less than 4"
C:" Observe a number greater than 7"

- Since event A occurs if the upper face is $1, 3$, or 5, it is a collection of three simple events and we write:

$$A = \{1, 3, 5\}$$

- The event B occurs if the upper face is $1, 2$, or 3 and is defined as a collection or set of these three simple events, so:

$$B = \{1, 2, 3\}$$

- C:" Observe a number greater than 7"

$$C = \emptyset$$

Example 5.1.12 *Consider rolling a fair die twice and observing the dots facing up on each roll. What is the sample space?*

There are $36 = 6 \times 6$ possible outcomes in the sample space S, where

$$S = \left\{ \begin{array}{c} (1;1),(1;2),(1;3),(1;4),(1;5),(1;6),(2;1),(2;2),(2;3),(2;4),(2;5),(2;6), \\ (3;1),(3;2),(3;3),(3;4),(3;5),(3;6)(4;1),(4;2),(4;3),(4;4),(4;5),(4;6), \\ (5;1),(5;2),(5;3),(5;4),(5;5),(5;6),(6;1),(6;2),(6;3),(6;4),(6;5),(6;6) \end{array} \right\}$$

- What is the event D " that the sum of faces is greater than 7", given that first outcome was a 3 ?

$$D = \{(3;5),(3;6)\}$$

5.1.7 Basic principle of counting

Theorem 5.1.1 *Suppose two experiments to be performed and*

- For Experiment 1, we have m possible outcomes .
- For each outcome of Experiment 1$\hookrightarrow$We have n outcomes for Experiment 2.

Then,
Total number of possible outcomes is $m \times n$.

Example 5.1.13 *Consider rolling a fair die twice and observing the dots facing up on each roll. What is the sample space?*

There are $36 = 6 \times 6$ possible outcomes in the sample space S.

Example 5.1.14 *If a* 25- *member club needs to elect a chair and a treasurer, how many different ways can these two to be elected?*

- For the chair position, there are 25 total possibilities. For each of those 25 possibilities, there are 24 possibilities to elect the treasurer.

Then, we obtain $n \times m = 25 \times 24 = 600$ different ways.

5.1.8 The generalized basic principle of counting

If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes; and if, for each of these n_1 possible outcomes, there are n_2 possible outcomes of the second experiment; and if, for each of the possible outcomes of the first two experiments, there are n_3 possible outcomes of the third experiment; and if ..., then there is a total of $(n_1 \times n_2 \times ..... \times n_r)$ possible outcomes of the r experiments.

Example 5.1.15 *1) How many different* 7*-place license plates are possible if the first* 3 *places are to be occupied by letters and the final* 4 *by numbers?*
2) How many different 7*-place license plates if the number can be used once?*

- By the generalized version of the basic principle:

1) The different 7-place license plates is :

$$26 \times 26 \times 26 \times \ 10 \times \ 10 \times \ 10 \times \ 10 = 175.760.000$$

2) If the number can be used once:

26 choices for each letter

10 choices for 1st number
9 choices for 2nd number
8 choices for 3rd number
7 choices for 4th number

Then,

$$26 \times 26 \times 26 \times \ 10 \times \ \ 9 \times 8 \times 7 = 88\ 583\ 040$$

5.2 Algebra of Events in Probability

In probability, any subset, say A of a sample space S, is called an event. We have different types of events in probability. Also, we can perform different operations on these called the algebra of events.

5.2.1 Operations on events

In a random experiment , let S be the sample space. Let $A \subseteq S$ and $B \subseteq S$ be the events in S. We say that

1)Intersection of Events

The intersection of two events A and B, denoted by $A \cap B$, is the event containing all elements that are common to A and B.

2)Union of Events

The union of the two events A and B, denoted by $A \cup B$,is the event containing all the elements that belong to A or B (or both).

3) Complement of An Event

The complement of A (denoted as $\overline{A}$)
$\overline{A}$ is an event that occurs only when A doesn't occur.

4)Disjoint

Events that have no outcomes in common are said to be **disjoint** (or mutually exclusive).
In the other word,
A and B are disjoint(or mutually exclusive) if and only if $A \cap B = \emptyset$.

5.2.2 Useful relationships

- $A \cap \overline{A} = \emptyset$
- $\overline{\overline{A}} = A$
- $A \cup \emptyset = A$
- $A \cap \emptyset = \emptyset$
- $A \cup \overline{A} = S$
- $\overline{S} = \emptyset$
- $\overline{\emptyset} = S$
- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$

Example 5.2.1 *Refer to the preceding example 5.1.11. Give $A \cap B, A \cup B$ and $\overline{A}$.*

We have $A = \{1, 3, 5\}$ and $B = \{1, 2, 3\}$. So,

$$\begin{aligned} A \cap B &= \{1, 3\} \\ A \cup B &= \{1, 2, 3, 5\} \\ \overline{A} &= \{2, 4, 6\} \end{aligned}$$

5.3 Probability spaces

5.3.1 Event Space

Definition 5.3.1 *The collection $\mathcal{F} = \mathcal{P}(S)$ of subsets of the sample space is called an event space if*

1. $S \in \mathcal{F}$.
2. $\forall A \in \mathcal{F} \Rightarrow \bar{A} \in \mathcal{F}$.
3. $\forall A,\ B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$.
4. $\forall (A_n)_{n\in I} \in \mathcal{F},\ I \subseteq \mathbb{N} \Rightarrow \underset{n\in I}{U} A_n \in \mathcal{F}$.

5.3.2 Probability measure

Definition 5.3.2 *A probability measure on $\mathcal{F}$ is a function:*

$$\begin{array}{rccc} P: & \mathcal{F} & \rightarrow & [0,1] \\ & A & \mapsto & P(A) \end{array}$$

satisfying

1) $\forall A \in \mathcal{F}/\ (A \neq \emptyset\ et\ A \neq S),\ \ 0 < P(A) < 1.$

2) $P(S) = 1$

3) $P(\emptyset) = 0$

4) $\forall A \in \mathcal{F},\ P\left(\bar{A}\right) = 1 - P(A)$.

5) $\forall A \in \mathcal{F}, \forall B \in \mathcal{F},\ P(A \cup B) = P(A) + P(B)$(If A and B are two disjoint events $A \cap B = \emptyset$).

6) $\forall (A_i)_{1\leq i\leq n} \in \mathcal{F}$ sequence of of disjoint events two by two in $\mathcal{F}$ (in that $A_i \cap A_j = \emptyset$ when ever $i \neq j$) then,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Remark 5.3.1 *If $P(A) = 1$, then event A will always happen if the experiment is performed $(A = S)$.*

Remark 5.3.2 *If $P(A) = 0$, event A will never happen $(A = \emptyset)$.*

5.3.3 Probability spaces

Definition 5.3.3 *A probability space is a triple* $(S,\ \mathcal{F}, P)$ *of objects such that*

(a) S is a non-empty set.
(b) $\mathcal{F}$ is an event space of subsets of S.
(c) P is a probability measure on $(S, \mathcal{F})$.

Remark 5.3.3 *The structure of probability space* $(S,\ \mathcal{F}, P)$ *depends greatly on whether* S *is a countable set (that is, a finite or countably infinite set) or an uncountable set. If* S *is a countable set, we normally take* $\mathcal{F}$ *to be the set of all subsets of* S *and* $(S,\ \mathcal{F}, P)$ *is a discrete probability space.*

5.3.4 The probability of an event

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element.

Definition 5.3.4 *The probability of an event* A *is equal to the sum of the probabilities of the simple events contained in* A.

- **In particular**, the probability of any event $A = \{e_1, e_2, ..., e_n\}$ is the sum of the probabilities of its elements:

$$P(A) = P(\{e_1, e_2, ..., e_n\}) = P(e_1) + P(e_2) + ... + P(e_n)$$

Note that: We are using here the simpler notation $P(e_i)$ to denote the probability of the event $\{e_i\}$, instead of the more precise $P(\{e_i\})$.

Example 5.3.1 *Consider experiment :Involving three coin tosses. The outcome will now be a* 3*-long string of heads or tails. The sample space is*

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

We assume that each possible outcome has the same probability of $\frac{1}{8}$.
Consider, as an example the event $A = \{\text{exactly 2 heads occur}\}$ then,

$$A = \{HHT, HTH, THH\}$$

Using additivity, the probability of A is the sum of the probabilities of its elements:

$$\begin{aligned} P(A) &= P(\{HHT, HTH.THH\}) \\ &= P(\{HHT\}) + P(\{HTH\}) + P(\{THH\}) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} \end{aligned}$$

Example 5.3.2 *The proportions of blood phenotypes A, B, AB, and O in the population of all Caucasians in the United States are reported as* 0.41, 0.1, 0.04 *and* 0.45, *respectively.*

If a single Caucasian is chosen randomly from the population, what is the probability that he or she will have either type A or type O blood?
Solution: The four simple events, A, B, AB, and O, do not have equally likely probabilities. Their probabilities are found using the relative frequency concept as:

$$P(A) = 0.41 \quad , \quad P(B) = 0.10, \quad P\ (AB) = 0.04, \ P(O) = 0.45$$

- The event of interest consists of two simple events, so

$$\boxed{\text{P(person is either type } A \text{ or type } O) = P(A) + P(O) = 0.41 + 0.45 = 0.86}$$

Application to an experiment with equally likely outcomes

Suppose that an experiment involves a large number N of simple events and you know that all the simple events are equally likely. Then each simple event has probability $\frac{1}{N}$, and the probability of an event A can be calculated as:

$$\begin{aligned} P(A) &= \frac{\text{number of outcomes in } A}{\text{total number of possible outcomes of the experiment}} \\ &= \frac{Card\ A}{Card\ S} = \frac{n_A}{N} \end{aligned}$$

where $n_A = Card\ A$ is the number of simple events that result in the event A and $N = Card\ S$.

Example 5.3.3 ***Experiment*** *1: $S = \{H, T\}$ and $\mathcal{F} = \mathcal{P}(S) = \{\emptyset, \ \{H\}, \ \{T\}, \ S\}$*

and the probability P defined for any event A by:

$$\begin{array}{rccc} P: & \mathcal{F} & \rightarrow & [0,1] \\ & A & \mapsto & P(A) \end{array}$$

1/ If $A = \emptyset : \ P(A) = 0$.

2/ If $A = \{H\} : P(A) = \frac{1}{2}$.

3/ If $A = \{T\} : P(A) = \frac{1}{2}$.

4/ If $A = S : P(A) = 1$.

Example 5.3.4 ***Experiment*** *2(Toss a die):* $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$.

(case : Equally likely outcomes: The probability of each outcomes $= \frac{1}{N} = \frac{1}{6}$)

$$\begin{aligned} P(\{1\}) &= P\{1\} = \frac{1}{6} \\ \text{and } P\{2\} &= P\{3\} = P\{4\} = P\{5\} = P\{6\} = \frac{1}{6}. \end{aligned}$$

1/ A' :"Observe an even number"

$$A' = \{2,\ 4,\ 6\}$$

and

$$P(A') = \frac{CardA'}{Card\ S} = \frac{3}{6} = \frac{1}{2}$$

2) A:" Observe an odd number". Then, $A = \{1,\ 3,\ 5\}$ and $P(A) = \frac{1}{2}$.

3) E:" Observe a number greater than 5". Then, $E = \{\ 6\}$ and $P(E) = \frac{1}{6}$.

4) B:" Observe a number less than 4". Then, $B = \{1, 2, 3\}$and $P(B) = \frac{1}{2}$

Example 5.3.5 *(Toss a die)(case :Not Equally likely outcomes)*

A dice is loaded in such a way that an even number is **twice** as likely to occur as odd number.

- We assign a probability of λ to each odd number and a probability of 2λ to each even number.

We have,

$$\begin{aligned} P(\{1\}) &= \lambda \\ P(\{2\}) &= 2\lambda \\ P(\{3\}) &= \lambda = P(\{5\}) \\ P(\{4\}) &= P(\{6\}) = 2\lambda \end{aligned}$$

Then,

$$9\lambda = 1 \Rightarrow \lambda = \frac{1}{9}$$

- Find $P(B)$:

B:" Observe a number less than 4"
So,

$$B = \{1, 2, 3\}$$

and

$$P(B) = P(\{1, 2, 3\}) = P(\{1\}) + P(\{2\}) + P(\{3\}) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}$$

5.4 General probability theorems

Consider a probability law, and let A, B, and C be events.
(a) If $A \subset B$, then $P(A) \leq P(B)$.
(b) $P(AUB) = P(A) + P(B) - P(A \cap B)$.(the Addition Rule)
(c) $P(AUB) \leq P(A) + P(B)$.
(d) $P(AUBUC) = P(A) + p(\overline{A} \cap B) + p(\overline{A} \cap \overline{B} \cap C)$.

Chapter 6

Conditional Probability and Independent Events

6.1 Conditional probability

Definition 6.1.1 *If $A, B \in \mathcal{F}$ and $P(B) > 0$, the (conditional) probability of A given B is denoted by $P(A \setminus B)$ and defined by:*

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

Example 6.1.1 *Eighty students were surveyed about playing an instrument. The results are shown in the two-way frequency table.*

Gender\Play an Instrument	*Yes*	*No*	*Total*
Male	20	15	*35*
Female	28	17	*45*
Total	*48*	*32*	*80*

One of these students is to be selected at random.
What is the probability that he is a male and given that he play an Instrument?

Let A " student Male" and B "student play an Instrument".

- Find $P(A \setminus B)$:

We have $P(A \cap B) = \frac{20}{80}$ and $P(B) = \frac{48}{80}$. Then,

$$P(A \setminus B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{20}{80}}{\frac{48}{80}} = \frac{20}{48}$$

Example 6.1.2 *Amina is a good student. The probability that she studies mathematics and passes her english test is* $\frac{18}{20}$ *. If the probability that Amina studies mathematics is*$\frac{15}{16}$ *. Find the probability that Amina passes her english test, given that she studied.*

Let A"studies mathematics" and B"passes english test"
We have: $P(A \cap B) = \frac{18}{20}$ and $P(A) = \frac{15}{16}$ then,

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{18}{20}}{\frac{15}{16}} = 0,96$$

6.1.1 Properties

Proposition 6.1.1 *Conditional probability satisfies the following properties:*

1/ $P(S \setminus B) = 1$.

2/ $P(A \cap B) = P(A \setminus B) \times P(B) = P(B \setminus A) \times P(A)$.(Multiplication Rule)

3/ $P\left(\bar{A} \setminus B\right) = 1 - P(A \setminus B)$.

4/ $P(A) = P(A \setminus E) \times P(E) + P\left(A \setminus \overline{E}\right) \times P(\overline{E})$(Total Probability Rule)

Proof:
1/

$$P(S \setminus B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

2/

$$P(A \setminus B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A \setminus B) \times P(B)$$

and

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B \setminus A) \times P(A)$$

Then,

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

3/The events A and $\overline{A}$ are disjoint and satisfy $S = A \cup \overline{A}$.

$$\begin{aligned} S &= A \cup \overline{A} \Longrightarrow S \cap B = \left(A \cup \overline{A}\right) \cap B \\ S \cap B &= (A \cap B) \cup \left(\overline{A} \cap B\right) \end{aligned}$$

Then,

$$\begin{aligned} P(S \cap B) &= P\left[(A \cap B) \cup \left(\overline{A} \cap B\right)\right] \\ &\Longrightarrow P(B) = P(A \cap B) + P\left(\overline{A} \cap B\right) \\ &\Longrightarrow \frac{P(B)}{P(B)} = \frac{P(A \cap B) + P\left(\overline{A} \cap B\right)}{P(B)} \\ &\Longrightarrow 1 = \frac{P(A \cap B)}{P(B)} + \frac{P\left(\overline{A} \cap B\right)}{P(B)} \\ &\Longrightarrow 1 - \frac{P(A \cap B)}{P(B)} = \frac{\left(\overline{A} \cap B\right)}{P(B)} \\ &\Longrightarrow P\left(\bar{A} \setminus B\right) = 1 - P(A \setminus B) \end{aligned}$$

$$P\left(\bar{A} \setminus B\right) = 1 - P(A \setminus B)$$

4/Demonstrate the Total Probability Rule:
The events E and $\overline{E}$ are disjoint and satisfy $S = E \cup \overline{E}$.
Therefore, we have

$$\begin{aligned} P(A) &= P\left[(A \cap E) \cup \left(A \cap \overline{E}\right)\right] \\ &= P(A \cap E) + P\left(A \cap \overline{E}\right) \\ &= P(A \setminus E) \times P(E) + P\left(A \setminus \overline{E}\right) \times P(\overline{E}) \end{aligned}$$

Note: Invoking the fact that $P(A \cap B) = P(B \setminus A) \times P(A)$, the Addition Rule can also be expressed as

$$P(AUB) = P(A) + P(B) - P(B \setminus A) \times P(A)$$

6.2 Independent Events

Definition 6.2.1 *Two events are said to be independent if knowing one occurs does not change the probability of the other occurring.*

This is equivalent to saying that,

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

where $P(A) > 0$
and

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

where $P(B) > 0$
These give the following formal definition.

Definition 6.2.2 *Events A and B of a probability space $(S,\ \mathcal{F}, P)$ are called **independent** if and only if*

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

If A and B are **not independent** then they are **dependent**.

Example 6.2.1 *A single fair die is rolled. Let $A = \{3\}$ and $B = \{1, 3, 5\}$. Are A and B independent?*

In this example we can compute all three probabilities

$$P(A) = \frac{1}{6}, P(B) = \frac{1}{2}, \text{and} \quad P(A \cap B) = P(\{3\}) = \frac{1}{6}$$

Since the product $P(A) \times P(B) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$ and

$$P(A) \times P(B) \neq P(A \cap B)$$

It's follows that, the events A and B are not independent.

6.3 Bayes' rule

Bayes' rule (also known as Bayes' theorem) is a useful tool for calculating conditional probabilities. Bayes' theorem can be stated as follows:

Theorem 6.3.1 *(Bayes' theorem) Let A and B be two events, each of positive probability. Then*

$$\begin{aligned} P(A \setminus B) &= \frac{P(A)}{P(B)} \times P(B \setminus A) \\ P(A \setminus B) &= \frac{P(B \setminus A) \times P(A)}{P(B)} \end{aligned}$$

- Applying the law of total probability with $S = A \cup \overline{A}$ to the probability $P(B)$ denominator

we can write

$$\begin{aligned} P(A \setminus B) &= \frac{P(B \setminus A) \times P(A)}{P(B)} \\ &= \frac{P(B \setminus A) \times P(A)}{P(B \setminus A) \times P(A) + P\left(B \setminus \overline{A}\right) \times P(\overline{A})} \end{aligned}$$

- Total Probability formula is an important formula for manipulating conditional probabilities.

The idea is that if you partition a sample space into events $A_1; A_2; ....; A_n$ such that the $A_i's$ are mutually exclusive and $\underset{i}{\cup} A_i = S$, then for any event B,

$$P(B) = \sum_{i=1}^{n} P(B \setminus A_i) \times P(A_i)$$

Example 6.3.1 *In a certain assembly plant, three machines, A_1, A_2, and A_3,make $30\%, 45\%$, and 25%, respectively, of the products. It is known from past experience that $2\%, 3\%$, and 2% of the products made by each machine, respectively, are defective. If a product was chosen randomly and found to be defective, what is the probability that it was made by machine A_3?*

Solution:
Let D " the product is defective".

- Find $P(A_3 \setminus D)$?

We have

- $P(A_1) = 0.3$, $P(A_2) = 0.45$ and $P(A_3) = 0.25$.
- $P(D \setminus A_1) = 0.02$, $P(D \setminus A_2) = 0.03$, and $P(D \setminus A_3) = 0.02$

Using Bayes' rule to write:

$$\begin{aligned} P(A_3 \setminus D) &= \frac{P(A_3) \times P(D \setminus A_3)}{P(D)} \\ &= \frac{0.25 \times 0.02}{P(D)} \end{aligned}$$

- Find $P(D)$:

Suppose that a finished product is randomly selected. What is the probability that it is defective? $P(D)$?

- Applying the total probability rule , we can write

$$\begin{aligned} P(D) &= P(D \setminus A_1) \times P(A_1) + P(D \setminus A_2) \times P(A_2) + P(D \setminus A_3) \times P(A_3) \\ &= 0.02 \times 0.3 + 0.03 \times 0.45 + 0.02 \times 0.25 = 0.0245 \end{aligned}$$

Thus,

$$\begin{aligned} P(A_3 \setminus D) &= \frac{P(D \setminus A_3) \times P(A_3)}{P(D)} \\ &= \frac{0.25 \times 0.02}{0.0245} = 0.204 \end{aligned}$$

6.4 Exercises

Exercise N°1:
A class consists of 6 boys and 4 girls.
1) An exam is given and no two students obtain the same score.
a) How many different rankings are possible?
b) What if boys and girls are ranked separately?
2)The teacher wants to select 1 girl and 1 boy to represent the class for a function. In how many ways can the teacher make this selection?

Exercise N°2:
If 3 books are picked at random from a box containing 5 novels, 3 books of mathematics , and a dictionary.

1) Find the total number of possibilities.
2)What is the number of possibilities that:

- the dictionary is selected.
- 2 novels and 1 books of mathematics are selected.

3) What is the probability that:

- the dictionary is selected.
- 2 novels and 1 books of mathematics are selected.

Exercise N°3:
A and B are two events. If $p(A) = 0.3, p(B) = 0.2$ and $p(A \cap B) = 0.1$
Find the following probabilities: $p\left(\overline{A}\right), p(A \cup B)$and $p\left(\overline{A \cup B}\right)$.

Exercise N°4:
A random experiment can result in one of the outcomes $\{e_1, e_2, e_3, e_4\}$with probabilities $0.2, 0.4, 0.1$and 0.3 respectively.
Let A denote the event$\{e_1, e_2\}$, B the event $\{e_2, e_3, e_4\}$ and $C = \{e_3\}$
Find $p(A), p(B), p(C), p\left(\overline{A}\right), p\left(\overline{B}\right)$and $p(A \cap C)$.

Exercise N°5:
Two dice are rolled. Consider the events:
$A =$ { sum of two dice equals 3 }
$B =$ { sum of two dice equals 7 } and $C =$ { at least one of the dice shows a 1 }.

(a) What is $P(A \setminus C)$?
(b) What is $P(B \setminus C)$?
(c) Are A and C independent? What about B and C?

6.5 Solutions

Exercise N°1:
1)
a) $10! = 3.628.800$ different rankings are possible.
b) $6! \times 4! = 17.280$ different rankings are possible if boys and girls are ranked separately.
2)The teacher wants to select 1 girl and 1 boy to represent the class for a function. Here the teacher is to perform two operations:
1°) Selecting a boy from among the 6 boys. So, these can be done in 6 ways
and
2°) Selecting a girl from among 4 girls. So, these can be performed in 4 ways.
By the fundamental principle of counting, the required number of ways is: $6\times 4 = 24$.
Exercise N°2:

1) The total number of possibilities is : $C_9^3 = 84$
2)
i) The number of possibilities that the dictionary is selected :

$$C_1^1 \times C_8^2 = 28$$

ii)The number of possibilities that 2 novels and 1 books of mathematics are selected is:

$$C_5^2 \times C_3^1 = 30$$

3) The probability that:

- the dictionary is selected $= \frac{28}{84}$
- 2 novels and 1 books of mathematics are selected $= \frac{30}{84}$

Exercise N°3:
We have :$p(A) = 0.3, p(B) = 0.2$ and $p(A \cap B) = 0.1$.
Then,

$$P\left(\bar{A}\right) = 1 - P(A) = 1 - 0.3 = 0.7$$

$$\begin{aligned} P(AUB) &= P(A) + P(B) - P(A \cap B) \\ &= 0.3 + 0.2 - 0.1 = 0.4 \end{aligned}$$

and

$$P\left(\overline{A \cup B}\right) = 1 - P(A \cup B) = 1 - 0.4 = 0.6$$

Exercise N°4:
We have:

$$\begin{aligned} p(\{e_1\}) &= p(e_1) = 0.2 \\ p(\{e_2\}) &= p(e_2) = 0.4 \\ p(\{e_3\}) &= p(e_3) = 0.1 \\ p(\{e_4\}) &= p(e_4) = 0.3 \end{aligned}$$

- Find $p(A)$:

We have $A = \{e_1, e_2\}$, then

$$p(A) = p(e_1) + p(e_2) = 0.2 + 0.4 = 0.6$$

- Find $p(B)$:

We have $B = \{e_2, e_3, e_4\}$, then

$$p(B) = P(e_2) + P(e_3) + P(e_4) = 0.4 + 0.1 + 0.3 = 0.8$$

- Find $p(C)$:

$C = \{e_3\}$, So:

$$p(C) = p(e_3) = 0.1$$

- Find $p\left(\overline{A}\right)$:

$$p\left(\overline{A}\right) = 1 - p(A) = 1 - 0.6 = 0.4$$

- Find $p\left(\overline{B}\right)$:

$$p\left(\overline{B}\right) = 1 - p(B) = 1 - 0.8 = 0.2$$

- Find $p(A \cap C)$:

$$p(A \cap C) = P(\emptyset) = 0$$

Exercise N°5:
Note that: the sample space S is

$$S = \{(i;j)/\ i;j = 1;2;3;4;5;6\}$$

So,

$$S = \left\{ \begin{array}{c} (1;1),(1;2),(1;3),(1;4),(1;5),(1;6),(2;1),(2;2),(2;3),(2;4),(2;5),(2;6), \\ (3;1),(3;2),(3;3),(3;4),(3;5),(3;6)(4;1),(4;2),(4;3),(4;4),(4;5),(4;6), \\ (5;1),(5;2),(5;3),(5;4),(5;5),(5;6),(6;1),(6;2),(6;3),(6;4),(6;5),(6;6) \end{array} \right\}$$

with each outcome equally likely.
Then

$$A = \{(1;2);(2;1)\}$$

$$B = \{(1;6);(2;5);(3;4);(4;3);(5;2);(6;1)\}$$

$$C = \{(1;1);(1;2);(1;3);(1;4);(1;5);(1;6);(2;1);(3;1);(4;1);(5;1);(6;1)\}$$

$$A \cap C = A$$

$$B \cap C = \{(1;6);(6;1)\}$$

Thus,

$$\begin{aligned} P(A) &= \frac{2}{36} \\ P(B) &= \frac{6}{36} \\ P(C) &= \frac{11}{36} \end{aligned}$$

Therefore,

$$P(A \setminus C) = \frac{P(A \cap C)}{P(C)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$$

and

$$P(B \setminus C) = \frac{P(B \cap C)}{P(C)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$$

i) Note that: $P(A) = \frac{2}{36} \neq P(A \setminus C)$. So, A and C are not independent.
ii) Similarly, $P(B) = \frac{6}{36} \neq P(B \setminus C)$. Then, B and C are not independent.

Chapter 7

Random variables

7.1 Random Variables

Let S be the sample space associated with an experiment . A random variable X is a function that assigns a real number $X(s)$ to each elementary event $s \in S$. Since the outcome of the experiment, i.e., the specific s, is not predetermined, the value of $X(s)$ is not fixed in advance. This implies that the value of the random variable is determined by the specific outcome of the experiment.
A more formal definition of a random variable is as follows.

7.1.1 Random Variable

Definition 7.1.1 *A random variable is a function from the sample space S to the set $\mathbb{R}$ of all real numbers*

$$\begin{array}{rccc} X & : & S & \rightarrow & \mathbb{R} \\ & & s & \rightarrow & X(s) \end{array}$$

7.1.2 Types of random variables

There are two types of random variables: discrete and continuous

A Discrete Random Variable

- If a random variable has a finite or countably infinite set of values it is called discrete random variable.
- The possible valus of X may be listed as: $x_{1,}x_2, ...$

A Continuous Random Variable

When the random variable can take any value on the real line it is called a continuous random variable.

7.2 Discrete Random Variables

7.2.1 Probability mass function

Definition 7.2.1 *For a discrete random variable X with possible values $x_{1,}x_2, ..., x_n$, a probability mass function(pmf) (or distribution function) is function f such that:*

1. $f(x_i) = p(x_i) = P\{X = x_i\}$
2. $f(x_i) \geq 0$
3. $\sum_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} p(x_i) = \sum_{i} p\{x = x_i\} = 1$

- The probability distribution for a discrete random variable is a formula, table, or graph that gives the possible values of x, and the probability $p(x)$ associated with each value of x.

$X = x_i$	x_1	x_2	.....				.....	x_n
$f(x_i) = p(x_i)$	$p(x_1)$	$p(x_2)$	.....				.....	$p(x_n)$

Refer to the preceding examples.

Example 7.2.1 *Consider example 5.1.6: involving three coin tosses.*

The outcome will now be a 3-long string of heads or tails. The sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Let the random variable X be **the number of heads** in three coin tosses. The table below lists the eight outcomes of s and the corresponding values of X.

Outcome(s)	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(s)$	3	2	2	2	1	1	1	0

X assigns each outcome in s a number from the set $\{0, 1, 2, 3\}$. Then

$$X(S) = \{0,\ 1,\ 2,\ 3\}$$

$$\begin{aligned}
P(X = 0) &= P(\{TTT\}) = \frac{1}{8} \\
P(X = 1) &= P\ (\{HTT, THT, TTH\}) = \frac{3}{8} \\
P(X = 2) &= P(\{HHT, HTH\ , THH\}) = \frac{3}{8} \\
P(X = 3) &= P(\{HHH\}) = \frac{1}{8}
\end{aligned}$$

These results are summarized in the following table:(**discrete probability distribution**)

x	0	1	2	3	$\sum_{x=0}^{3} P(X = x)$
$P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Example 7.2.2 *Letting X denote the random variable that is defined as the sum of two fair dice,*

$$S = \left\{ \begin{array}{c} (1;1),(1;2),(1;3),(1;4),(1;5),(1;6),(2;1),(2;2),(2;3),(2;4),(2;5),(2;6), \\ (3;1),(3;2),(3;3),(3;4),(3;5),(3;6)(4;1),(4;2),(4;3),(4;4),(4;5),(4;6), \\ (5;1),(5;2),(5;3),(5;4),(5;5),(5;6),(6;1),(6;2),(6;3),(6;4),(6;5),(6;6) \end{array} \right\}$$

with each outcome equally likely.

Then

$$X(S) = \{\ 2,\ 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

and

$$P\{X = 2\} = P\{(1,1)\} = \frac{1}{36}$$
$$P\{X = 3\} = P(\{(1,2),(2,1)\}) = \frac{2}{36}$$
$$P\{X = 4\} = P(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36}$$
$$P\{X = 5\} = P(\{(1,4),(2,3),(3,2),(4,1)\}) = \frac{4}{36}$$
$$P\{X = 6\} = P(\{(1,5),(2,4),(3,3),(4,2),(5,1)\}) = \frac{5}{36}$$
$$P\{X = 7\} = P(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = \frac{6}{36}$$
$$P\{X = 8\} = P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36}$$
$$P\{X = 9\} = P(\{(3,6),(4,5),(5,4),(6,3)\}) = \frac{4}{36}$$
$$P\{X = 10\} = P(\{(4,6),(5,5),(6,4)\}) = \frac{3}{36}$$
$$P\{X = 11\} = P(\{(5,6),(6,5)\}) = \frac{2}{36}$$
$$P\{X = 12\} = P(\{(6,6)\}) = \frac{1}{36}$$

These results are summarized in the following table:
(**probability distribution or probability mass function)**

x	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example 7.2.3 *Probability distribution for example 7.2.1.*

$X = x$	0	1	2	3
$f(x) = P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

A graph of $f(x)$ is presented in Figure 1.

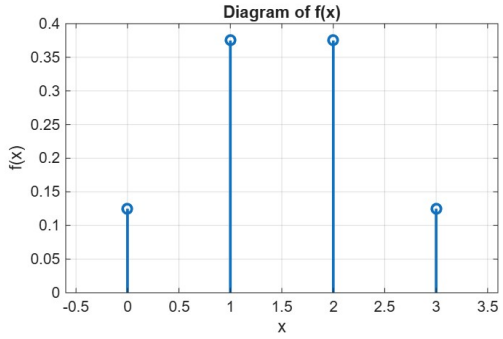


Figure 1

7.2.2 Cumulative Distribution Function

If X is a discrete random variable on $(S,\ \mathcal{F}, P)$, the cumulative distribution function(CDF) of X is the measures the probability that random variable X assumes a value less than or equal to x, that is $P(X \leq x)$.

Definition 7.2.2 *The cumulative distribution function (CDF) denoted by F is the function $F : \mathbb{R} \to [0, 1]$ defined by:*

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

And for any a :

$$F(a) = \sum_{\text{all } x \leq a} p(x)$$

Remark 7.2.1 *We sometimes write $F_X(x)$ for $F(x)$.*

Example 7.2.4 *Refer to the preceding example*

Probability distribution is:

$X = x$	0	1	2	3
$f(x) = P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Then, the cumulative distribution function F of X is given by:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{1}{8} & \text{if} \quad 0 \leq x < 1 \\ \frac{4}{8} & \text{if} \quad 1 \leq x < 2 \\ \frac{7}{8} & \text{if} \quad 2 \leq x < 3 \\ 1 & \text{if} \quad x \geq 3 \end{cases}$$

The graph of $F(x)$ is shown in Figure 2.

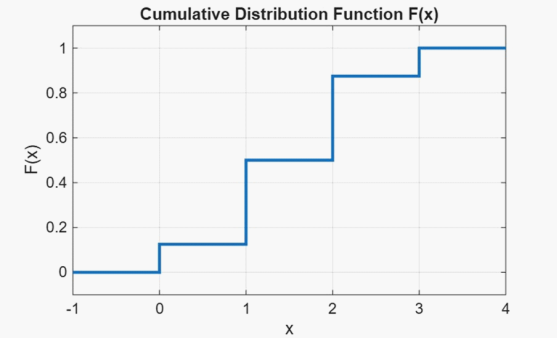


*Figure*2

7.2.3 Expected Value

The expected value or mean is mesure of center or middle of the probability distribution.

Definition 7.2.3 *The expected value (the expectation or mean value or mean) of the discrete random variable X, which we denote by $E(X)$ or μ, is given by:*

$$\mu = E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

$$\mu = \sum x f(x)$$

Remark 7.2.2 *The expected value of the random variable X is the value that you would expect to observe on average if the experiment is repeated over and over again.*

Example 7.2.5

$X = x$	0	1	2	3
$f(x) = P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

So the mean is:

$$\mu = E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = \frac{3}{2} = 1.5$$

Proposition 7.2.1 *Lets X and Y be two discrete random variables defined on the same probability space.*

So:

1/ $E(X+Y) = E(X) + E(Y)$.

2/ $E(aX+b) = aE(X) + b, \forall (a,b) \in \mathbb{R}^2$.

Example 7.2.6 *A discrete random variable with $E[X] = 0.5$, evaluate $E[2X+3]$*

$$E(aX+b) = aE(X) + b$$
$$E[2X+3] = 2E(X) + 3 = 2 \times 0.5 + 3 = 4$$

7.2.4 Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF $f(x)$, and let $g(X)$ be a realvalued function of X. Then, the expected value of the random variable $g(X)$ is given by:

$$E[g(X)] = \sum_x g(x) f(x)$$

7.2.5 Variance

The variance is a mesure of the dispersion or variability in the distribution.

Definition 7.2.4 *The variance of the random variable X, which we denote by $Var(X)$(or $V(X)$) is given by:*

$$Var(X) = \sum_{i=1}^{n} p(x_i)\,[x_i - E(X)]^2 \text{ or } Var(X) = E[X - E(X)]^2$$

Proposition 7.2.2 *Lets X be a discrete random variable and a, b two real numbers.*

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

2/ $Var(aX + b) = a^2 Var(X)$.

Example 7.2.7 *A discrete random variable with $Var(X) = 1.5$ evaluate $Var(2X + 3)$.*

$$Var(aX + b) = a^2 Var(X)$$

$$Var(2X + 3) = 2^2 Var(X) = 4 \times 1.5 = 6$$

7.2.6 Standard deviation

Standard deviation is defined as the square root of the variance. That is

$$\sigma(X) = \sqrt{Var(X)}$$

Example 7.2.8 *Consider the random variable X of Example 7.2.1, which has the* ***pmf***

$X = x$	0	1	2	3
$f(x) = P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The mean: $E[X] = \frac{3}{2}$

- Calculations: The variance

METHOD1:
$\mu = E(X) = \frac{3}{2}$

x	$p(x)$	$xp(x)$	$(x-\mu)^2$	$(x-\mu)^2 p(x)$
0	$\frac{1}{8}$	0	$(0-\frac{3}{2})^2 = \frac{9}{4}$	$\frac{9}{32}$
1	$\frac{3}{8}$	$\frac{3}{8}$	$(1-\frac{3}{2})^2 = \frac{1}{4}$	$\frac{3}{32}$
2	$\frac{3}{8}$	$\frac{6}{8}$	$(2-\frac{3}{2})^2 = \frac{1}{4}$	$\frac{3}{32}$
3	$\frac{1}{8}$	$\frac{3}{8}$	$(3-\frac{3}{2})^2 = \frac{9}{4}$	$\frac{9}{32}$
Total	/	$\mu = E(X) = \frac{3}{2}$	/	$V(x) = \frac{3}{4}$

So the variance is:

$$\begin{aligned} Var(X) &= \sum (x-\mu)^2 p(x) \\ Var(X) &= \frac{3}{4} \end{aligned}$$

and the Standard deviation is:

$$\sigma = \sqrt{\frac{3}{4}}$$

METHOD2:

x	$p(x)$	$xp(x)$	x^2	$x^2 p(x)$
0	$\frac{1}{8}$	0	0	0
1	$\frac{3}{8}$	$\frac{3}{8}$	1	$\frac{3}{8}$
2	$\frac{3}{8}$	$\frac{6}{8}$	4	$\frac{12}{8}$
3	$\frac{1}{8}$	$\frac{3}{8}$	9	$\frac{9}{8}$
Total	/	$E(X) = \frac{3}{2}$	/	$E(X^2) = 3$

So the variance is:

$$\begin{aligned} Var(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4} \end{aligned}$$

and the Standard deviation is:

$$\sigma = \sqrt{\frac{3}{4}}$$

7.2.7 Moments

The n^{th} moment is defined as: $E[X^n]$the expected value of the random variable X^n. the n^{th} moment is given by

$$m_n = E[X^n] = \sum x^n f(x)$$

and there is no need to calculate the **pmf** of X^n.

Example 7.2.9 *Compute second moment when X represents the outcome when we roll a fair die.*

Since $P\{X = i\} = \frac{1}{6}$, $i = 1, 2, 3, 4, 5, 6$, we obtain

$$m_2 = E[X^2] = 1^2\frac{1}{6} + 2^2\frac{1}{6} + 3^2\frac{1}{6} + 4^2\frac{1}{6} + 5^2\frac{1}{6} + 6^2\frac{1}{6} = \frac{91}{6}$$

Remark 7.2.3 *With this terminology the first moment of the random variable X is just the mean.*

7.2.8 Central moment

Definition 7.2.5 *The r^{th} central moment (or the r^{th} moment about the mean), denoted by $\mu_r(X)$ and defined as:*

$$\mu_r(X) = E[(X - E[X])^r]$$

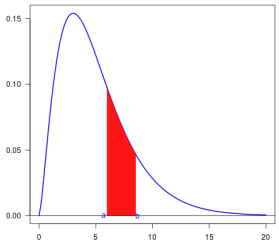
7.3 Continuous Random Variable

7.3.1 Probability density function

Definition 7.3.1 *For a continuous random variable X . a probability density function(**pdf**) is function $f : \mathbb{R} \to \mathbb{R}$ such that*

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

3. $P(a \leq X \leq b) = \int_a^b \mathrm{f(x)dx}$
(ares under $f(x)$ from a to b for any a and $b\,(a < b)$).



Example 7.3.1 *Suppose that X is a continuous random variable whose probability density function is given by:*

$$f(x) = \begin{cases} C.x & if \;\; 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

a) What is the value of C?
b) Find $P\{X > \frac{1}{2}\}$.
Solution:
a) Find the value of C:
Since f is a probability density function, we must have that: $\int_{-\infty}^{\infty} \mathrm{f(x)dx}=1 \int_{-\infty}^{\infty} f(x)dx =$
$1 \Rightarrow \int_0^1 f(x)dx = 1 \Rightarrow \int_0^1 C.xdx = 1 \Rightarrow \frac{C}{2} = 1 \Rightarrow C = 2$
Thus,

$$f(x) = \begin{cases} 2.x & \text{if} \;\; 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

b) Find $P\{X > \frac{1}{2}\}$:

$$P\{X > \frac{1}{2}\} = \int_{\frac{1}{2}}^{\infty} f(x)dx = \int_{\frac{1}{2}}^{1} 2xdx = \left[x^2\right]_{\frac{1}{2}}^{1} = 1 - \frac{1}{4} = \frac{3}{4}$$

7.3.2 Cumulative Distribution Function

If X is a discrete random variable on $(S,\ \mathcal{F}, P)$, the cumulative distribution function (CDF) of X is the measures the probability that random variable X assumes a value less than or equal to x, that is $P(X \leq x)$.

Definition 7.3.2 *The cumulative distribution function of a continuous random variable X is:*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

for $-\infty < x < \infty$

7.3.3 Relation of CDF and pdf

Differentiating both sides yields:

$$\frac{dF(x)}{dx} = f\left(x\right)$$

For a continuous random variable the pdf is the derivative of the CDF.

$$\boxed{\begin{array}{c} P\left(a \leq X \leq b\right) = \int_{a}^{b} f(x)dx = F(b) - F(a) \\ \hline \text{and } \frac{dF(x)}{dx} = f\left(x\right) \text{ if the derivative exists.} \end{array}}$$

Remark 7.3.1 *For any random variable X,*

$$P(a < X < b) = F(b) - F(a)$$

Example 7.3.2

$$f\left(x\right) = \begin{cases} 2.x & if\ \ 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

The cumulative distribution function is given by:

$$F\left(x\right) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \int_{0}^{x} 2.udu = x^2 & \text{if} \quad 0 \leq x \leq 1 \\ 1 & \text{if} \quad x \geq 1 \end{cases}$$

7.3.4 Expected value and Variance

Suppose that X is a continuous random variable with probability density function $f(x)$.

- The expected value (or mean value or mean) of X , denote as $E(X)$ or μ, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

- The variance of X, denote as $Var(X)$ or σ^2, is given by,

$$\begin{aligned} \sigma^2 &= V(X) = Var(X) = \int_{-\infty}^{\infty} f(x)\,[x-\mu]^2\, dx \\ &= E(X^2) - \mu^2 \end{aligned}$$

Where

$$\mu = E(X)$$

and

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$

- The standard deviation of X is

$$\sigma = \sqrt{V(X)}$$

- The n^{th} moment of X is

$$m_n = E[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$$

Example 7.3.3 *Consider:*

$$f(x) = \begin{cases} 2.x & if \;\; 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

- The expected value $E(X)$ is given by:

$$\mu = E(X) = \int_{-\infty}^{\infty} x.f\,(x)\,dx = \int_{0}^{1} 2.x^2 dx = \left[\frac{2x^3}{3}\right]_0^1 = \frac{2}{3}$$

- The variance of X is

$$V\,(X) = E(X^2) - (E(X))^2$$

We have,

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f\,(x)\,dx = \int_{0}^{1} 2.x^3 dx = \left[\frac{x^4}{2}\right]_0^1 = \frac{1}{2}$$

Then,

$$\begin{aligned} V\,(X) &= E(X^2) - (E(X))^2 \\ &= \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} \end{aligned}$$

- The standard deviation of X is:

$$\sigma = \sqrt{V(X)} = \sqrt{\frac{1}{18}}$$

Remark 7.3.2 *The method for calculating the expected value for a continuous random variable is similar to what you have done. Nevertheless, the basic results concerning expectations are the same for continuous and discrete random variables.*

7.4 Exercise

Exercise N°1:
Let X be the number of heads in the experiment of tossing two fair coins.

1. Find the probability distribution(PD) (or probability mass function(pmf)).
2. Find the cumulative distribution function (CDF).

Exercise N°2:
A random variable X can assume five values: $1, 2, 3, 4, 5, 6$.
A portion of the probability distribution is shown here:

x	1	2	3	4	5	6
$f(x) = p(x)$	0.1	0.2	?	0.3	0.1	0.2

1. Find $p(3)$.
2. Find the cumulative distribution function $F(x)$ and Construct a graph of $F(x)$.
3. Calculate the expected value (mean), variance, and standard deviation.
4. What is the probability that x is greater than 2?
5. What is the probability that x is 3 or less?

Exercise N°3:
Suppose that X is a continuous random variable whose probability density function is given by:

$$f(x) = \begin{cases} \frac{x^2}{3} & \text{if} \quad -1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the cumulative distribution function $F(x)$ of density function $f(x)$.
2. Use $F(x)$ to evaluate $p(0 < x \leq 1)$.
3. Find $\mu = E(X)$,$V(X)$, and σ.

Exercise N°4:
Find the Mean and Variance of the Bernoulli. Consider the experiment of tossing a biased coin, which comes up a head with probability p and a tail with probability $1-p$.

Exercise N°5:
Let X a discrete random variable with mean 5 and variance 65.
Consider another random variable Y such that $Y = 3X + 2$.
Evaluate the mean and variance of Y.

7.5 Solutions

Exercise N°1:
1.Let X be the number of heads in the experiment of tossing two fair coins.
Then the probability function is

$$P(X=0)=\frac{1}{4}; P(X=1)=\frac{1}{2}; P(X=2)=\frac{1}{4}$$

Thus, The probability distribution is

$X=x$	0	1	2
$f(x)=P(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

2. From the definition, the CDF is given by

$$F(x)=\begin{cases} 0 & \text{if} \quad x<0 \\ \frac{1}{4} & \text{if} \quad 0\leq x<1 \\ \frac{3}{4} & \text{if} \quad 1\leq x<2 \\ 1 & \text{if} \quad x\geq 2 \end{cases}$$

Exercise N°2:

x	1	2	3	4	5	6
$f(x)=p(x)$	0.1	0.2	?	0.3	0.1	0.2

1. Find $p(3)$:

The probabilities must sum to 1.
Therefore,

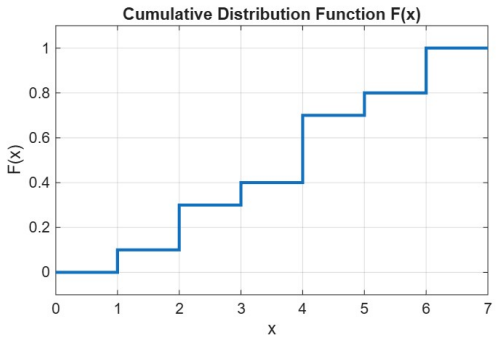
$$p(1)+p(2)+p(3)+p(4)+p(5)+p(6)=1$$

$$\Rightarrow \quad p(3)=1-[p(1)+p(2)+p(4)+p(5)+p(6)]$$
$$\Rightarrow \quad p(3)=0.1$$

2. The cumulative distribution function F of X is given by:

$$F(x)=\begin{cases} 0 & \text{if} \quad x<1 \\ 0.1 & \text{if} \quad 1\leq x<2 \\ 0.3 & \text{if} \quad 2\leq x<3 \\ 0.4 & \text{if} \quad 3\leq x<4 \\ 0.7 & \text{if} \quad 4\leq x<5 \\ 0.8 & \text{if} \quad 5\leq x<6 \\ 1 & \text{if} \quad x\geq 6 \end{cases}$$

A graph of $F(x)$.



4. The expected value (the mean)

- The mean is:

$$\mu=E(X)=\sum xp(x)=1\times 0.1+2\times 0.2+3\times 0.1+4\times 0.3+5\times 0.1+6\times 0.2=3.7$$

We have:

$$E(X^2)=\sum x^2p(x)=1^2\times 0.1+2^2\times 0.2+3^2\times 0.1+4^2\times 0.3+5^2\times 0.1+6^2\times 0.2=16.3$$

- The variance is:

$$\begin{aligned} Var(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= 16.3 - (3.7)^2 = 2.61 \end{aligned}$$

- The standard deviation is

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{2.61} = 1.615$$

5. The probability that x is greater than 2.

$$p(x > 2) = p(3) + p(4) + p(5) + p(6) = 0.7$$

6. The probability that x is 3 or less:

$$p(x \leq 3) = p(3) + p(2) + p(1) = 0.4$$

Exercise N°3:
1.Find the cumulative distribution function $F(x)$ of density function $f(x)$:
We have,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du = \int_{-1}^{x} \frac{u^2}{3}du = \frac{x^2+1}{9}$$

So,

$$F(x) = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{x^2+1}{9} & \text{if} \quad -1 \leq x < 2 \\ 1 & \text{if} \quad x \geq 2 \end{cases}$$

2. Find $p(0 < x \leq 1)$:

$$p(0 < x \leq 1) = F(1) - F(0) = \frac{1}{9}$$

3. Find $E(X), V(X)$ and σ:

- The expected value $E(X)$ is:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{-1}^{2} \frac{x^3}{3} dx = \left[\frac{x^4}{12}\right]_{-1}^{2} = \frac{15}{12}$$

- The variance of X is

$$V(X) = E(X^2) - (E(X))^2$$

We have,

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_{-1}^{2} \frac{x^4}{3} dx = \left[\frac{x^5}{15}\right]_{-1}^{2} = \frac{33}{15}$$

Then,

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= \frac{33}{15} - \left(\frac{15}{12}\right)^2 = 0.6375 \end{aligned}$$

- The standard deviation of X is

$$\sigma = \sqrt{V(X)} = \sqrt{0.6375} = 0.798$$

Exercise N°4:
Mean and Variance of the Bernoulli random:
The Bernoulli random variable X with pmf:

x	0	1
$f(x)$	$1-p$	p

The variance are given by the following calculations:

$$\begin{aligned} E[X] &= 1{\cdot}p + 0{\cdot}(1-p) = p \\ E[X^2] &= 1^2{\cdot}p + 0^2{\cdot}(1-p) = p \\ Var(X) &= E[X^2] - (E[X])^2 = p - p2 = p(1-p) \end{aligned}$$

Chapter 8

Usual discrete probability laws

A discrete random variable X can only take a finite or countable set of distinct values. Some notable and significant discrete distributions include the Bernoulli, Binomial, Geometric, and Poisson distributions. In this chapter, we will explore these distributions and their properties.

8.1 The Bernoulli Distribution

Many probability problems involve only two possible outcomes or can be simplified to have two outcomes. For instance:

1. When a coin is flipped, it can either land on heads or tails.
2. A true or false question can be answered with either "true" or "false."

These types of situations are referred to as **Bernoulli experiments**.

Definition 8.1.1 *We say that the discrete random variable X has the Bernoulli distribution with parameter p if a response variable that takes only two possible values, and we label one of these outcomes as* 1 *and the other as* 0.

So that X takes the values 0 and 1 only.

There exists $p \in [0,1]$ such that $P(X=0)=q$, $P(X=1)=p$ and the mass function of X is given by:

$$f(x)=P(X=x)=\begin{cases} q & \text{if} \quad x=0 \\ p & \text{if} \quad x=1 \\ 0 & \text{if} \quad x\neq 0,1 \end{cases}$$

or

X	0	1
$f(x)=p(x)$	q	p

Note that: The random variable X is said to have the Bernoulli distribution, we write this as

$$X \sim Bernoulli\,(p)$$

The expected value, variance and standard deviation of X are given by the following calculations:

1/

$$E(X)=p$$

2/

$$\begin{aligned} Var(X) &= E(X^2)-(E(X))^2 \\ &= p-p^2 \\ &= p(1-p) \\ &= pq \end{aligned}$$

3/

$$\sigma(X)=\sqrt{pq}$$

Example 8.1.1 *Consider the experiment of tossing a biased coin. Let the random variable X be* ***the number of head****.*

Then,

$$X(S)=\{0,\ 1\}$$

and the mass function of X is given by:

X	0	1
$f(x) = P(X = x)$	$\frac{1}{2}$	$\frac{1}{2}$

$X \sim$Bernoulli$\left(\frac{1}{2}\right)$ and we get:

1) $E(X) = \frac{1}{2}$.

2) $Var(X) = \frac{1}{4}$.

3) $\sigma(X) = \frac{1}{2}$.

8.2 The Binomial distribution

Suppose that we have a sequence of n Bernoulli trials such that we get a success (S) or failure (F) with probabilities $P\{S\} = p$ and $P\{F\} = q$ respectively$(q = 1 - p)$. Let X be the number of successs in the n trials. Then X is called a binomial random variable with parameters n, p and q, we write as

$$X \sim B(n,\ p,\ q)$$

Definition 8.2.1 *We say that X has the binomial distribution with parameters n, p and q if X takes values in $\{0, 1, ..., n\}$ $(n \in \mathbb{N}^* - \{1\})$ and*

$$P(X = k) = C_n^k p^k q^{n-k}$$

for $k = 0, 1, 2, ..., n$.with

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

By the binomial theorem a mass function satisfying:

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = (p + (1-p))^n = 1$$

Proposition 8.2.1 *Let X a binomial random variable with parameters n, p and q (or $X \sim B(n,\ p,\ q)$) .The binomial distribution has the following properties:*

1/ The expected value of X is:

$$E(X) = np$$

2/ The variance of X is:

$$Var(X) = npq = np\,(1-p)$$

3/ The standard deviation of X is:

$$\sigma\,(X) = \sqrt{npq} = \sqrt{np\,(1-p)}$$

Example 8.2.1 *Flip a coin* 3 *times. Let* X *= number of heads obtained, what are the expected value and the variance of* X*?*

$$X \sim B\left(3, \frac{1}{2}, \frac{1}{2}\right)$$

- Binomial trial $\longrightarrow$ Flip a coin
- Number of trials $(n = 3)$
- Success $\longrightarrow$Head occur $\left(p = \frac{1}{2}\right)$
- Let X = number of heads obtained
- Values $x = 0, 1, 2, 3$

Then,

1)

$$E(X) = np = \frac{3}{2}$$

2)

$$Var(X) = npq = \frac{3}{4}$$

3)

$$\sigma\,(X) = \sqrt{npq} = \frac{\sqrt{3}}{2}$$

Example 8.2.2 *Find the Binomial distribution for selected values of* $(n = 3)$ *(the probability mass function):*

The Binomial distribution is given by:

X	0	1	2	3
$f(x) = P(X = x)$				

We have:

$$P(X = k) = C_n^k p^k q^{n-k}$$

Then,

$$P(X = 0) = C_3^0 (0.5)^0 (0.5)^{3-0} = \frac{1}{8} = 0,125$$

$$P(X = 1) = C_3^1 (0.5)^1 (0.5)^{3-1} = \frac{3}{8} = 0.375$$

$$P(X = 2) = C_3^2 (0.5)^2 (0.5)^{3-2} = \frac{3}{8} = 0.375$$

and

$$P(X = 3) = C_3^3 (0.5)^3 (0.5)^{3-3} = \frac{1}{8} = 0,125$$

So,

X	0	1	2	3
$f(x) = P(X = x)$	$0,125$	0.375	0.375	$0,125$

Example 8.2.3 *Find* $P(x = 2)$ *for a binomial random variable with* $n = 10$ *and* $p = 0.1$.

$$P(X = 2) = C_{10}^2 (0.1)^2 (0.9)^{10-2} = 0.1937$$

Example 8.2.4 *A survey found that one out of five Americans say he or she has visited a doctor in any given month. If* 10 *people are selected at random, find the probability that exactly* 3 *will have visited a doctor last month.*

- Here $n = 10$: $p = \frac{1}{5}$ and $q = \frac{4}{5}$

$$P(X = 3) = C_{10}^3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^{10-3} = 0.201$$

8.3 The Poisson Distribution

The Poisson distribution is a discrete random variable often used in the study of rare phenomena under certain conditions. We cite for example, X: "the number of people aged over 100 in a population".

Definition 8.3.1 *Let $\lambda \in \mathbb{R}_+^*$,we say that a random variable X has the Poisson distribution, and write $X \sim Poisson(\lambda)$ $(or\ X \sim \boldsymbol{P}(\lambda))$, if $X(S) = \mathbb{N}$ and*

$$\forall k \in \mathbb{N} : P(X = k) = P_k = e^{-\lambda}\frac{\lambda^k}{k!},$$

We note that since (from calculus)$\sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ it is indeed true (as it must be) that:

$$\begin{aligned}\sum_{k\in\mathbb{N}} P_k &= \sum_{k=0}^{+\infty} e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!} = 1\end{aligned}$$

Proposition 8.3.1 *Let X be a discrete random variable following the Poisson law $P(\ \lambda)$, $\lambda \in \mathbb{R}^{*,+}$.The Poisson distribution has the following properties:*

1°) The expected value of X is:

$$E(X) = \lambda$$

2°) The variance of X is:

$$Var(x) = \lambda$$

3°) The standard deviation of X is:

$$\sigma(X) = \sqrt{\boldsymbol{\lambda}}$$

Example 8.3.1 *We consider the random variable X " number of typing errors per page of the Mathematics course".*

Here $X \sim P(\lambda)$, such that : $\lambda = 0.1$.
We have:

$$\forall k \in \mathbb{N} : P(X = k) = P_k = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-0.1}\frac{(0.1)^k}{k!}$$

Then, the probability having an error per page is:

$$\begin{aligned} P(X = 1) = P_1 &= e^{-0.1}\frac{(0.1)^1}{1!} \\ &= 0,09 \end{aligned}$$

Chapter 9

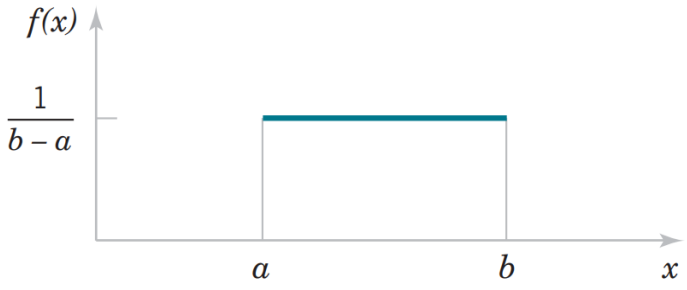
Usual continuous probability laws

Certain continuous distributions are so important that we list them here. Uniform, normal, exponential,...

9.1 The Uniform Random Variable

Definition 9.1.1 *We say that X is a uniform random variable on the interval (a, b) if its probability density function is given by:*

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$



Since the (cdf) if defined as: $F(x) = \int_{-\infty}^{x} f(u)du$.

$$F(x) = \begin{cases} 0 & \text{if} \quad x < a \\ \dfrac{x-a}{b-a} & \text{if} \quad a \leq x < b \\ 1 & \text{if} \quad x \geq b \end{cases}$$

9.1.1 Properties

The uniformly distributed random variable X over (a, b) has the following properties:

1/ The expected value of X is:

$$E[X] = \frac{a+b}{2}$$

and

2/ The variance of X is:

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

Example 9.1.1 *What is the probability density function?*

$$f(x) = \begin{cases} \dfrac{1}{4} = 0.25 & \text{if} \quad 0 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

i) The probability density function is an uniformly distributed random variable X over $(0, 4)$ and the (cdf) if defined as:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \dfrac{x}{4} & \text{if} \quad 0 \leq x < 4 \\ 1 & \text{if} \quad x \geq 4 \end{cases}$$

ii) Find $P(x \geq 3)$:

- **Method 1:**

$$P(x \geq 3) = \int_3^{+\infty} f(x)dx = \int_3^4 \frac{1}{4}dx = \left[\frac{x}{4}\right]_3^4 = \frac{1}{4}$$

- **Method 2:**

$$P(x \geq 3) = 1 - P(x < 3) = 1 - F(3) = \frac{1}{4}$$

9.2 The Normal Random Variables

Definition 9.2.1 *We say that X is a normal random variable, or simply that X is normally distributed,with parameters μ and σ^2 if the density function of X is given by*

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

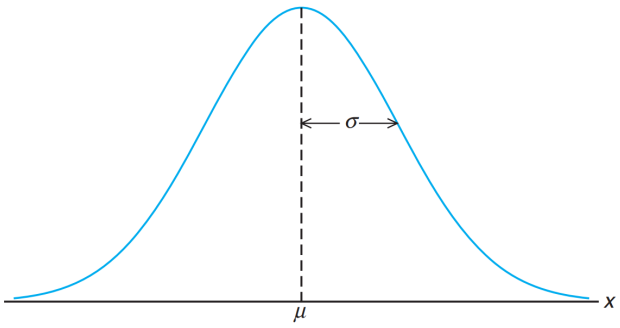
For a normally distributed random variable with parameters μ and σ^2 ($X \sim N(\mu; \sigma^2)$),

$$E[X] = \mu$$

and

$$Var(X) = \sigma^2$$

where $-\infty < \mu < \infty$ and $\sigma > 0$.



Normal Curve

9.2.1 Cumulative distribution function

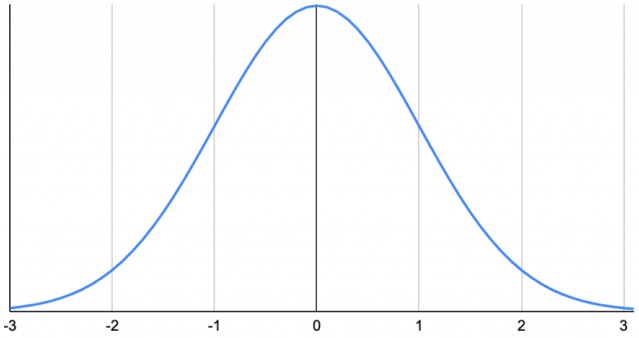
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

for -$\infty < x < \infty$

9.2.2 The Standard Normal Distribution

Definition 9.2.2 *The standard normal distribution is a normal distribution with a mean of* 0 *and a standard deviation of* 1.

The standard normal distribution is shown in Figure below.



Density function of standard normal random variable Z :

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

for -$\infty < Z < \infty$

- All normally distributed variables can be transformed into the standard normally distributed variable by using the formula for the standard score:

$$Z = \frac{X - \mu}{\sigma}$$

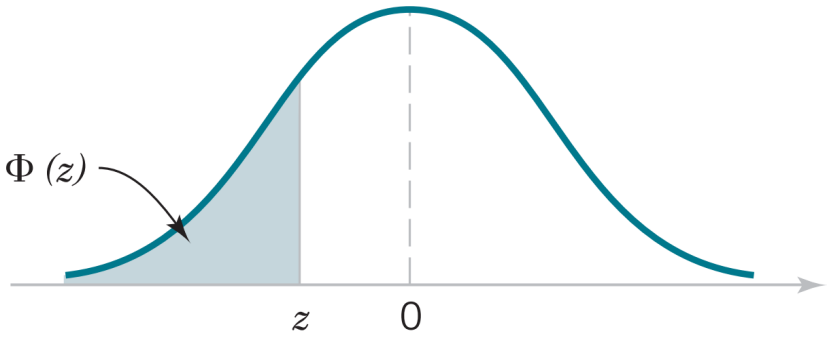
- The "standard normal" random variable is typically denoted Z and has mean 0 and variance 1:

If $X \sim N(\mu; \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0; 1)$$

The cumulative distribution function of standard normal random variable Z as:

$$\phi(z) = F(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du$$

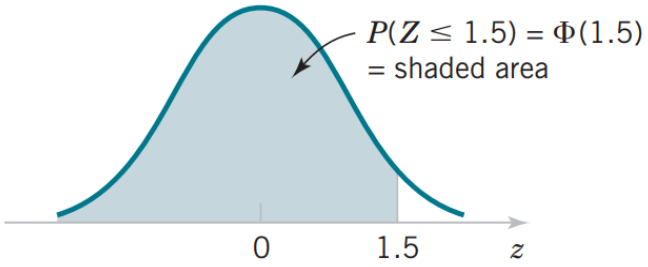


Some Properties

- Assume that Z is a standard normal random variable. The Table provides probabilities of the form
$$\phi(\alpha) = P(Z \leq \alpha)$$
- Note from symmetry of the probability density function about $z = 0$ that:
$$\phi(-\alpha) = 1 - \phi(\alpha)$$
- $P(Z < -\alpha) = 1 - P(Z \leq \alpha)$
- $P(\alpha < Z < \beta) = P(Z < \beta) - P(Z < \alpha)$
- $P(Z > \alpha) = 1 - P(Z \leq \alpha)$
- $P(Z > -\beta) = P(Z < \beta)$

Example 9.2.1 *The use of Table 1 to find* $P(Z \leq 1.5)$

$$P(Z \leq 1.5) = \Phi(1.5) = 0.93319$$



z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50800	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

*Table*1

Example 9.2.2 *Find the positive value of z if the normal curve area between* 0 *and z is* 0.3934.

We have

$$P(0 \leq Z \leq z) = 0.3934$$

and from table

$$P(0 \leq Z \leq 1.25) = 0.3934$$

So,

$$z = 1.25$$

Definition 9.2.3 *A continuous random variable whose probability density function is given, for some $\lambda > 0$, by:*

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

is said to be an exponential random variable (or, more simply, is said to be exponentially distributed) with parameter λ.

- The cumulative distribution $F(x)$ of an exponential random variable is given by:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)du = \int_{0}^{x} f(u)du = \int_{0}^{x} \lambda e^{-\lambda u}du = \left[-e^{-\lambda u}\right]_0^x \\ F(x) &= 1 - e^{-\lambda x} \end{aligned}$$

for $x \geq 0$.

- For an exponential random variable X with parameter λ,

$$E[X] = \frac{1}{\lambda}$$

and

$$Var(X) = \frac{1}{\lambda^2}$$

9.3 Exercises

Exercise N°1:
1) A discrete random variable with $E[X] = 0.5$, evaluate $E[2X + 3]$
2) A discrete random variable with $Var(X) = 1.5$ evaluate $Var(2X + 3)$.

Exercise N°2:
Let X is a random variable with mean 6 and variance 100.

1. Consider another random variable Y such that $Y = 3X + 6$.
2. Evaluate the mean and variance of Y?

Exercise N°3:
Consider the experiment of tossing a biased coin, which comes up a head with probability p and a tail with probability $1 - p$.

1. Find the probability mass function(pmf). Identify the distribution by name.
2. Find the Mean and the Variance.

Exercise N°4:
In a large consignment of electric bulb 10% are defective. A random sample of 20 is taken for inspection.
i) Find the probability that:
1°)All are good bulbs.
2°)There are exactly 3 defective bulbs.
3°)There are almost 3 defective bulbs.
ii) Find the expected value and the variance.

Exercise N°5:
We set that X is a uniform random variable on the interval (a, b) if its probability density function is given by:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } \ a < x < b \\ 0 & \text{otherwise} \end{cases}$$

1. Find the cumulative distribution function $F(x)$.
2. Calculate the expected value (mean) and variance.

Exercise N°6:
Suppose that the lifetime of a phone (e.g. the time until the phone does not function even after repairs), denoted by X, manufactured by the company A Pale, is exponentially distributed with mean 550 days.
1) Find the parameter λ of exponentially distributed.
2)Find the probability that a randomly selected phone will still function after two years, i.e. $X > 730$?

9.4 Solutions

Exercise N°1:
1) We have

$$E(aX+b) = aE(X) + b$$

then,

$$E\left[2X+3\right] = 2E(X) + 3 = 2 \times 0.5 + 3 = 4$$

2)

$$Var\left(aX+b\right) = a^2 Var(X)$$

$$Var\left(2X+3\right) = 2^2 Var(X) = 4 \times 1.5 = 6$$

Exercise N°2:
The mean and the variance of Y;
We have

$$E(X) = 6 \text{ and } V(X) = 100$$

then,

$$E\left(Y\right) = E\left[3X+6\right] = 3E(X) + 6 = 3 \times 6 + 6 = 24$$

and

$$V\left(Y\right) = V\left(3X+6\right) = 3^2 V(X) = 9 \times 100 = 900$$

Exercise N°3:
1) The Bernoulli random variable X with (pmf)

x	0	1
$f\left(x\right)$	$1-p$	p

2) The mean and variance are given by the following calculations:

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

and

$$E[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$$

Then,

$$Var(X) = E[X^2] - \left(E[X]\right)^2 = p - p2 = p(1-p)$$

Exercise N°4:
Here $n = 20$.

$$p = \frac{10}{100} = 0.1 \text{ and } \ q = 0.9.$$

i)By binomial distribution, the probability of getting X defective bulbs.

$$P\left(X = k\right) = C_n^k p^k q^{n-k}$$

1°)Find the probability that all are good bulbs:

- The probability of getting all good bulbs **equal to** probability of getting zero defective bulbs.

So,

$$P(x = 0) = C_{20}^0 \left(0.1\right)^0 \left(0.9\right)^{20-0} = \left(0.9\right)^{20}$$

2°)Find the probability that: there are exactly 3 defective bulbs:

$$P(x = 3) = 0.1901$$

3°)Find the probability that:there are almost 3 defective bulbs.

$$\begin{aligned} P(x &\leq 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) \\ &= 0.8671 \end{aligned}$$

ii) Find the expected value and the variance:

$$E(X) = np = 20 \times 0.1 = 2$$

and

$$V(X) = npq = 20 \times 0.1 \times 0.9 = 1.8$$

Exercise N°5:
1) Set X a uniform random variable on the interval (a, b), its probability density function $f\left(x\right)$ is given by:

$$f\left(x\right) = \begin{cases} \dfrac{1}{b-a} & \text{if } \ a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Since

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{-a}^{x} \frac{1}{b-a}du = \frac{x-a}{b-a}$$

Then, the cumulative distribution function is given by:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < a \\ \dfrac{x-a}{b-a} & \text{if} \quad a \leq x < b \\ 1 & \text{if} \quad x \geq b \end{cases}$$

2)

- Compute $E(X)$:

$$\begin{aligned} \mu &= E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-a}^{b} \frac{1}{b-a}xdx \\ E[X] &= \frac{1}{b-a}\int_{-a}^{b} xdx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\ E[X] &= \frac{b-a}{2} \end{aligned}$$

- Compute $Var(X)$:

We have:

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_{-a}^{b} \frac{1}{b-a}x^2dx = \frac{1}{b-a}\int_{-a}^{b} x^2dx \\ &= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{(b-a)^3}{3(b-a)} = \frac{(b-a)^2}{3} \end{aligned}$$

Then,

$$\begin{aligned} Var(X) &= E(X^2) - \mu^2 \\ &= \frac{(b-a)^2}{3} - \left(\frac{b-a}{2}\right)^2 \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Exercise N°6:
1) Find the parameter λ :
Here the mean $\frac{1}{\lambda} = 550$. Hence $\lambda = \frac{1}{550}$ is the rate parameter.
2) Find $P(X > 730)$:
We have :

$$\begin{aligned} P(X &> 730) = 1 - P(X \leq 730) \\ &= 1 - F(730) \end{aligned}$$

Where $F(x)$ is a cumulative distribution of an exponential random variable.
Then,

$$P(X > 730) = 1 - (1 - e^{-\frac{730}{550}}) = e^{-\frac{730}{550}} = 0.2652$$

Bibliography

[1] Pierre Dagnélie. Statistique théorique et appliquée. De BoeckUniversité, 1998.

[2] Rick Durrett. Elementary probability for applications. Cambridge university press, 2009.

[3] Richard Arnold Johnson etGouri K. Bhattacharyya. Statistics : principles and methods. Wiley, 1996.

[4] Aurelio Mattei. Inférence et décision statistiques : théorie et application à la gestion des affaires. P. Lang, 2000.

[5] Sheldon M. Ross. Initiation aux probabilités. Presses polytechniques et universitaires romandes, 2007.

[6] Gilbert Saporta. Probabilités, analyse des données et statistique. Technip, 1990

[7] SERIE S CHAUM, Théorie et applications de la statistique, 1991.

Abstract

This course offers a basic introduction to statistics and probability for first-year engineering students. It is divided into two parts: the first covers descriptive statistics across three chapters (definitions, data summarization with tables/graphs, and two-variable datasets), and the second addresses probability in six chapters (combinatorial analysis, event probability, conditional probability, random variables, and discrete/continuous probability laws).

Keywords: Statistical probabilities, random variable, univariate statistics, bivariate statistics.

Résumé

Ce polycopié propose une introduction aux concepts fondamentaux de la statistique et des probabilités pour les étudiants de première année en ingénierie. Il est divisé en deux parties: la première traite des statistiques descriptives en trois chapitres (définitions, résumés de données avec tableaux/graphes, ensembles de deux variables), et la deuxième partie sur la probabilité en six chapitres (analyse combinatoire, probabilité d'événements, probabilité conditionnelle, variables aléatoires, lois discrètes et continues).

Mots-clés: Probabilités, statistiques, variable aléatoire, statistique à une variable, statistique à deux variables.

ملخص

هذه المطبوعة مقدمة للمفاهيم الأساسية في الإحصاء والاحتمالات لطلبة السنة الأولى هندسة. وينقسم إلى قسمين: يتناول القسم الأول الإحصاء الوصفي في ثلاثة فصول (التعريفات، تلخيص البيانات باستخدام الجداول/الرسوم البيانية، بيانات ذات متغيّرين)، أما القسم الثاني فيتعلق بالاحتمالات ويتضمن ستة فصول (التحليل التوافقي، احتمال الأحداث، الاحتمال الشرطي، المتغيرات العشوائية، التوزيعات المنفصلة والمستمرة).

الكلمات المفتاحية: الاحتمالات الإحصائية، المتغير العشوائي، الإحصاء أحادي المتغير، الإحصاء ثنائي المتغير