Part I

Descriptive Statistics

Chapter 1

Statistical Concepts

1.1 Introduction

The field of statistics is concerned with the scientific study of collecting, organizing, analyzing, and drawing conclusions from data. Statistical methods help us to transform data to knowledge. Statistical concepts enable us to solve problems in a diversity of contexts, add substance to decisions, and reduce guesswork. The discipline of statistics stemmed from the need to place knowledge management on a systematic evidence base. Earlier works on statistics dealt only with the collection, organization, and presentation of data in the form of tables and charts. In order to place statistical knowledge on a systematic evidence base, we require a study of the laws of probability. In mathematical statistics we create a probabilistic model and view the data as a set of random outcomes from that model. Advances probability theory enable us to draw valid conclusions and to make reasonable decisions on the basis of data.

Statistical methods are used in almost every discipline, including agriculture, astronomy, biology, business, communications, economics, education, electronics, geology, health sciences, and many other fields of science and engineering, and can aid us in several ways. Modern applications of statis tical techniques include statistical communication theory and signal processing, information theory, network security and denial of service problems, clinical trials, artificial and biological intelligence, quality control of manufactured items, software reliability, and survival analysis.

1.2 Definitions

▶ **Statistics** is a branch of science dealing with collecting, organizing, summarizing, analysing and making decisions from dat

▶ **Descriptive statistics** deals with methods for collecting, organizing, and describing data by using tables, graphs, and summary measures

▶ **A population** is the set of all elements (observations), items, or objects that bring them a common recipe and at least one that will be studied their properties for a particular goal. The components of the population are called individuals or elements.

Remark 1 Note that a population can be a collection of any things, like Ipad set, Books, animals or inanimate, therefore it does not necessary deal with people.

▶ **A sample** is a subset of the population selected for study

▶ **An element** (or member of a sample or population) is a specific subject or object about which the information is collected.

▶ **A variable** is a characteristic under study that takes different values for different elements.

▶ **The value of a variable** for an element is called an **observation** or **measurement** and also **modality**

Example 1.1 Scores of six students in a Statistics test can be $4, 6, 8, 3, 2$ and 9 marks

- **The variable** is **marks**
- **The modalities (values)** are $4, 6, ..., 9$

Quantitative variable gives us quantitative data	Qualitative variable gives us qualitative data
The age of people in years 19, 2, 45, 23, 88, ...	The gender of Organisms Male, Female .....
Number of children in family 5, 2, 4, 1, 14, ...	Results tossed a coin twice HH, HT, TH, TT (H=Head, T=Tail)
The heights of buildings in meters 15, 5.6, 12.7, 105, 27, ...	Eye color of people Black, Brown, Blue, Green, ...
The weights of cars in tons (ton=1000 Kg) 2.35, 1.65, 2.05, 2.10, 1.30, ...	Religious affiliation Muslim, Christian, Jew, ...
The speed of a car going on a main road in Km 110, 105, 85, 120, 90, ...	The pressure in a boiler High, Moderate, Low

1.3 Type of Data

We know that the variable is a characteristic under study that takes different values for different elements. In statistics, we have **two types of variables** according to their elements; first type is called **quantitative variable** and the second one is called **qualitative variable.**

When a subject can be measured numerically such as (the price of a shirt), then the subject in this case is quantitative variable. The following definition provides us with this concept.

Definition 1.1 Quantitative variable gives us **numbers** representing counts or measurements

When a subject cannot be measured numerically such as (eye color), then the subject in this case is qualitative variable. The following

definition provides us with this concept

.

Definition 1.2 Qualitative variable (or **categorical data**) gives us **names** or labels that are not numbers representing the observations.

The following examples illustrates the two type of variables

Example 1.2 The following table shows some examples of the two types of variables gives us quantitative data

Moreover, the variables measured in **quantitative data** divided into t**wo main types**, **discrete** and **continuous**. A variable that assumes countable values is refer to discrete variable, otherwise the variable is a continuous one. Accordingly, we provide the following definitions.

Definition 1.3 Discrete variables assume values that can be counted or that must take of set of certain values

Example 1.3

-The number of children in a family, where we have $1, 2, 3, ...$

-The number of students in a classroom, where we have $21, 25, 32, 18$ and so on.

-Number of accidents in a city, where we have $1, 2, 3, ...$or k accidents.

The other type of quantitative variable is the continuous variablewhich is assumed uncountable values definition.

Defintion 1.4 Continuous variables assume all values between any two specific values, i.e. they take all values in an interval. They often include fractions and decimals in values

Example1.4

- **Temperature**: For example the temperature in Riyadh city in last summer was between 15 and 56, i.e. the temperaturet ![15, 56].

- **Age**: For example the age of a horse is between 0 (Stillborn) and 62 years (Said the oldest horse was 62 years, but the middle age of a horse is 30 years), i.e. the age of a horse $x \in [0, 62]$

- **Height**: For example the height of a student in a Country is between 110 cm (person elf) and 226 cm (person giant), i.e. the height of a student $x \in [110, 226]$.

▶ **The qualitative variable** can be **nominal** or **ordinal** , the difference between both levels is explained in the following definitions.

Definition 1.5 The nominal level of measurement classifies data into mutually exclusive (disjoint) categories in which no order or ranking can be imposed on the data.

Example 1.5

– **Gender**: Male, Female.

– **Eye color**: Black, Brown, Blue, Green, ...

– **Religious** affiliation: Muslim, Christian, Jew, ...

– **Nationality**: Saudi, Syrian, Jordanian, Egyptian, Pakistani, ...

– **Scientific major** field: statistics, mathematics, computers, Geography

Definition 1.6 The ordinal level of measurement classifies data into categories that can be ordered, however precise differences between the ranks do not exist.

Example 1.6

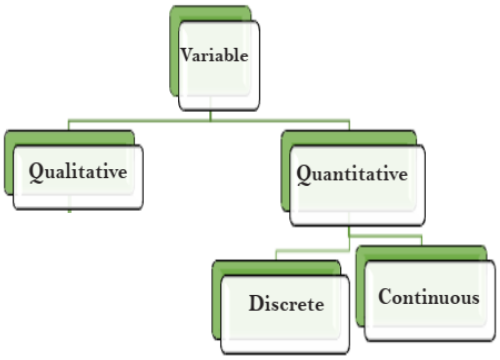
• **Grade** (h, A, B, C, D F) : Grading technique is the most common example on ordinal level. For example we find that the system

of appreciation in Saudi universities are (in descending order) A^+, A, B^+, B, C^+, C, D^+, D, F.

• **Rating scale** (bad, good, excellent and so on ...): To test the,quality of the canned product, we find that the state of the tested object either excellent or good or bad.

• **Ranking of football players:** A football player can be ranked in first grade, second grade, third grade, ...

• **Ranks of university faculty member**s: Academic ranks usually classified as professor, associate professor, assistant professor , and instructor.



1.4 Organizing Data

In this section we will learn how to organize the Qualitative and Quantitative data.

Frequency Distribution

Definitions

► **Absolute (Raw) frequency** n_i is the number of individuals who share the same modality
► **Relative frequency** f_i is the proportion of observations taking a certain modality such that

$$f_i = \frac{n_i}{N}, \tag{1}$$

where N is the size of a sample or a population whith $N = \sum_{i=1}^{k} n_i$ and $\sum_{i=1}^{k} f_i = 1$
► **Percentage of a Category,** we find it by multiplying the relative frequency of category by 100% such that

$$p_i = f_i \times 100$$

► **A Frequency Table** shows how the frequencies are distributed over various categories.
► **Frequency Distribution:** can be written as a table reporting frequencies of observations across the different modalities that the variable of interest can assume

x_i	x_1	x_2	...	x_K
n_i	n_1	n_2	...	n_K
f_i	f_1	f_2	...	f_K

Example 1.7

s	s	n	v	s	n	n	n	v	v
v	s	v	n	n	v	s	n	v	v
v	v	s	n	v	s	s	v	n	n
v	v	s	s	v	v	v	n	s	s
s	s	v	v	v	n	n	s	s	s

Type of Satisfaction	Relative Frequency (f)	Percentage
v	20/50=0.40	(0.40) (100%)=40%
s	17/50=0.34	(0.34) (100%)=34%
n	13/50=0.26	(0.26) (100%)=26%
	Sum= 1.00	**Sum**= 100%

Assume that a sample of 50 students from the preparatory year **(PY)** was selected, and those students were asked how they feel about the degree of their satisfaction of the program.The responses of those students are recorded below where **(v)** means very high satisfaction, **(s)** means somewhat satisfaction and **(n)** means no satisfaction.
Construct a frequency distribution table for these data

Complet the table by the relative frequency and percentage

► **Cumulative Frequency** is a form of frequency distribution in which each frequency beginning with the second (**from the top**) is added with the total of the previous ones This kind of cumulative frequency is known as **'less than type' cumulative frequency** when addition is done **from top**. Conversely if addition is done from below, then it will be **'greater than type'** cumulative frequency or **'more than type'**

Example1.8

Types of satifaction	n_i	**less than cumulative frequenc**	**greater than cumulati**
v	20	2O	50
s	17	37=17+20	30=50-20
n	13	50=13+20	13=30-17
$\sum$	50		

Chapter 2

Graphical Presentation of Frequency Distribution

Graphical Presentation of Frequency Distribution is the representation or presentation of data as Diagrams and Graphs.

► **Advantages of Graphical Representation of Data**

1. Data are presented pictorially.
2. Give better insight and understanding of the data.
3. Makes the presentation eye-catching.
4. The data become more logical (clear).
5. The comparison becomes easy.
6. Can derive the conclusion from data very quickly

.

► **Disadvantages of Graphical Representation of Data**

1. A graph cannot represent all details of the variables.
2. Very difficult to include and study the small differences in large measurements.
3. Graphs usually show approximate figures.
4. Graphs are only a supplement to the tabular presentationof data.

2.1 Graphical Representation of Qaulitative Data

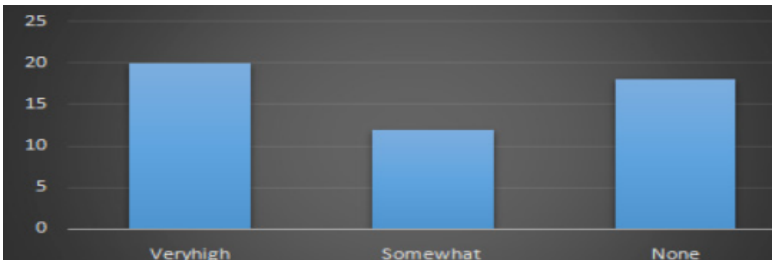
There are many types of graphs that are used to display qualitative data; in this part we will study and graph two of such graphs

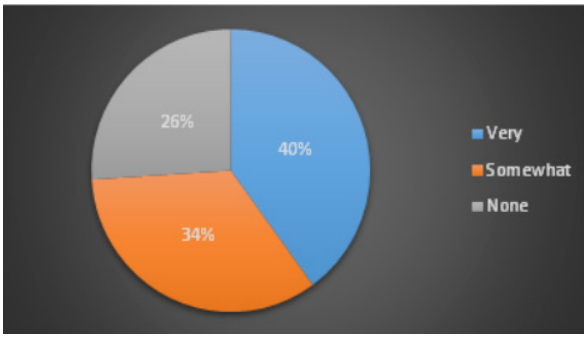
which they are commonly used to display the qualitative data, these graphs are the **Bar chart** and the **Pie chart**.

1. Bar Chart:

To construct a bar graph (also called a bar chart), we use the following steps

⋆ **Construct a Bar Graph (Chart)**





1. Represent the categories on the horizontal axis (All categories are represented by intervals of the same width).

2. Mark the frequencies on the vertical axis.

3. Draw one bar for each category such that the bar graphs for relative frequency and percentage can be drawn simply by marking the relative frequencies or percentages, instead of the frequencies, on the vertical axis

Refer to the example 1.7, we construct **bar graph** for its data as follows:

2. Pie Graph (Chart)

1. Draw a circle.

2. Find the central angle for each category by the following equation:

$$\text{Measure of the central angle} = (\text{Relative frequency}) \times \mathbf{360}^\circ$$

3. Draw sectors corresponding to the angles that obtained in step 2.

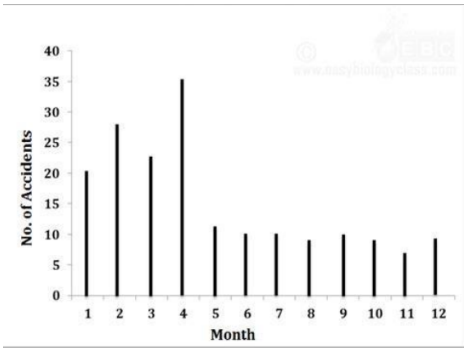
⋆ **Construct a pie chart for same example**

Applying step 2 for each category, we get

For category (**v**) the measure angle is $(0.40)(360^\circ) = 144^\circ$

For category (**s**) the measure angle is $(0.34)(360^\circ) = 122.4^\circ$

For category (**n**) the measure angle is$(0.26)(360^\circ) = 93.6^\circ$



2.2 Graphical Representation of Quantitative Data

A) Discret Data:

1 line diagram

The graphical representation of discret quntitative Data is called **line diagram** and to constract it we do the same steps in the constraction of the **bar graph** but we draw **stright line** for each value

Example 2.1 A study on the number of accidents in the year 2015 in a particular area is given below. Draw a line graph to represent the data.

Month	1	2	3	4	5	6	7	8	9	10	11	12
No, of accidents	21	27	23	34	12	11	9	8	9	8	7	8

Figure 2.3: line diagram

2. Frequency polygon

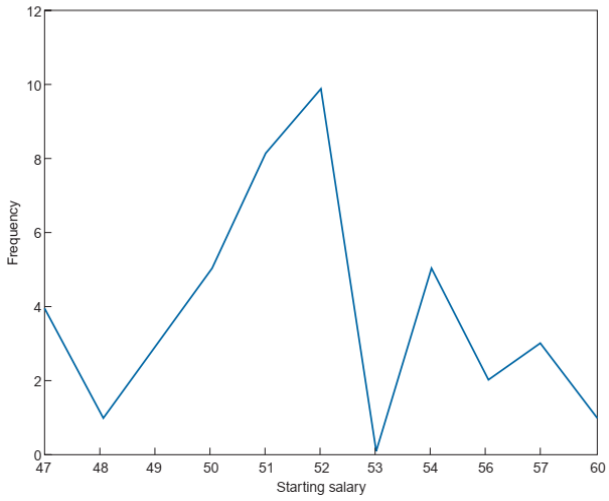
A **Frequency polygon** is a graphical tool used to display the distribution of discrete data, such that points are plotted for each modality
and connected by straight lines, forming a continuous line or "polygon".

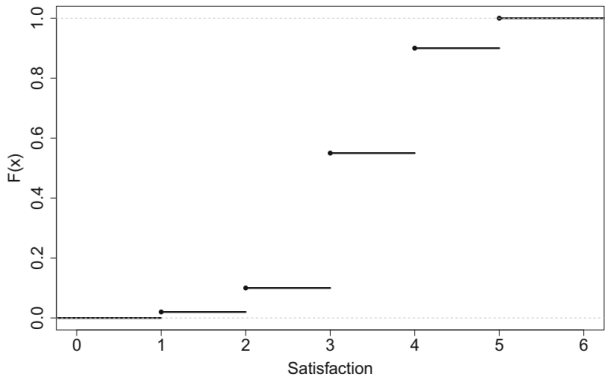
3. Empirical cumulative distribution function (ECDF)

Defintion 2.1 The empirical cumulative distribution function $F(x)$ is defined as the cumulativerelative frequencies of all values a_j , which are smaller than, or equal to x:

$$F(x) = \sum_{a_j \leq x} f(a_j)$$

This definition implies that
- $F(x)$ is a monotonically non-decreasing function,
- $0 \leq F(x) \leq 1$,
- $lim_{x \to -\infty}\ F(x) = 0$ (the lower limit of $F$is 0),
- $lim_{x \to +\infty}\ F(x) = 1$ (the upper limit of F is 1),





- $F(x)$ is right continuous.

► The empirical cumulative distribution function of discret variables is a **step function**.

B) Continous Data

When raw data are collected, they are organized numerically by distributing them into classes or categories in order to determine the number of individuals belonging to each class.

B.1 Procedure for forming frequency distribution

Given a set of observation , for a single variable.

1.Determine the range $(\mathbf{R}) = \mathbf{L} - \mathbf{S}$, where $\mathbf{L}$ = **largest observation** in the raw data; and $\mathbf{S}$ = **smallest observation** in the raw data.

2. Determine the appropriate number of classes or groups ($\mathbf{K}$). The choice of $\mathbf{K}$ is arbitrary but as a general rule, it should be a number (integer) depending on the size of the data given.

There are several suggested guide lines aimed at helping one decided on how many class intervals to employ.

Two of such methods are:

1.$\mathbf{K} = \mathbf{1} + \mathbf{3.322}\,(\log_{10}\mathbf{N})$

2. $\mathbf{K} = \sqrt{\mathbf{N}}$ where $\mathbf{N}$ = number of observations

3. Determine the **width** (**w**)**of the class interva**l. It is determined as $\mathbf{w} = \dfrac{\mathbf{R}}{\mathbf{K}}$

4. Determine the numbers of observations falling into each class interval i.e. find the class frequencies and the classes takes this form $[\mathbf{x}_i,\ \mathbf{x}_{i+1}[$.

Example 2.2 The following are the marks of 50 students : 48 70 60 47 51 55 59 63 68

Construct a frequency table for the above data

Answer: Range $(\mathbf{R}) = \mathbf{73 - 47 = 26}$

Number of classes $\mathbf{K} = \sqrt{\mathbf{N}} = \sqrt{50} = 7,07 \simeq 7$

Calsse size (width)$\mathbf{w} = \dfrac{\mathbf{R}}{\mathbf{K}} = \dfrac{26}{7} = \mathbf{3.7} \simeq \mathbf{4}$

We deduce the folloing distribution of grouped Data

Mark	$[47,51[$	$[51,55[$	$[55,59[$	$[59,63[$	$[63,67[$	$[67,71[$	$[71,74[$	$\sum$
Frecency	7	7	7	8	11	6	4	50

B.2 Graphs

Grouped data can be displayed in a **histogram**, **a polygon** or **ogive**.In this part we will learn how to construct such graphs.

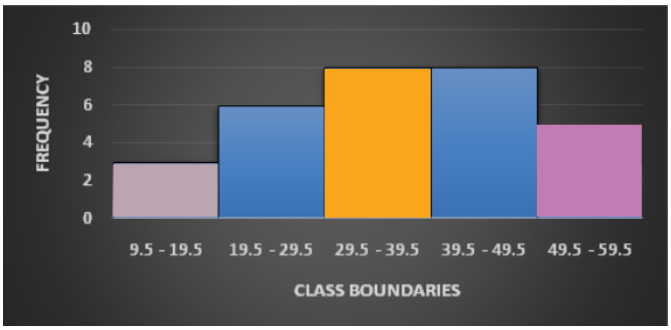
1. Histogram

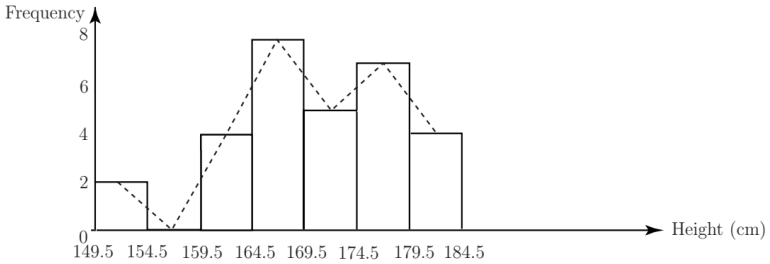
Definition 2.2 An histogram of grouped data in a frequency distribution table with **equal class widths** is a graph in which class boundaries are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on a vertical axis, were the bars are drawn **adjacent** to each other .

Frquency density When we construct a histogram, since the classes may **not have equal widths** or **standard interval** you must calculate the **frecency density** (d_i) wich marked on $y-$ axis such that $d_i = \dfrac{n_i}{a_i}$(or $= \dfrac{f_i}{a_i}$),were a_i is the with of the class $[x_i\ ,\ x_{i+1}[$.

2. Polygon

Definition 2.3 A polygon is another type of graphs that can be used to represent grouped quantitative data. To draw a frequency polygon we plot





the points **(class midpoint, frequency)** and connect these points by **line segments**
Remak Note that, we add one **midpoint** with **zero** frequency **before** the firstclass and **after** the last class midpoint to closed the graph., but when the classes has **equal with**

3. Ogive

► **The ogive** is a cumulative frequency curve.

There are two types of ogives:

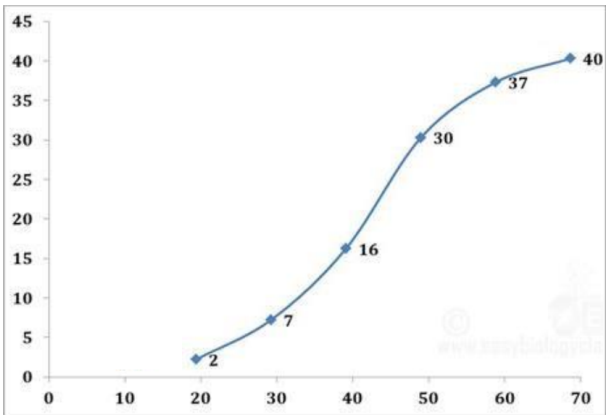
(1). **Less than ogive**

(2). **Greater than ogive** (more than ogive)

► **Less than ogive** is the graph of the **less than cumulative frequency distribution** ($n_i^c \nearrow$) which shows **the number of observations** less than the **upper class limit**

.

Example 2.3 Construct a **less than ogive** using the following daa



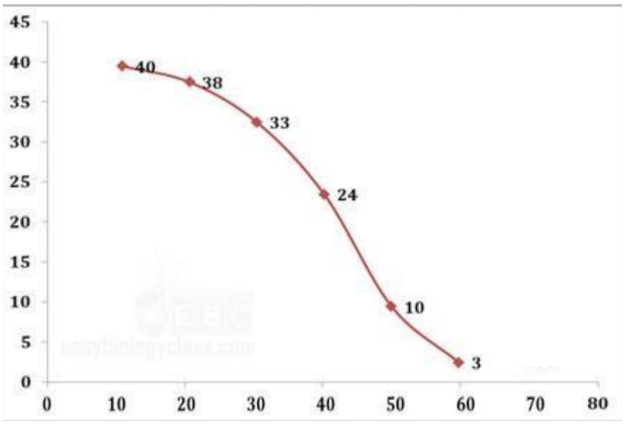
Class	Frecency	Upper limite of class	$n_i^c \nearrow$
$[10, 20[$	2	20	2
$[20, 30[$	5	30	7
$[30, 40[$	9	40	16
$[40, 50[$	14	50	30
$[50, 60[$	7	60	37
$[60, 70[$	3	70	40
$\sum$	40		

Figure 2.7: histogram with frecency polygone

▶ **Greater than ogive** is the graph of the **greater than cumulative frequency** ($n_i^c \searrow$)distribution which shows the number of observations greater than the **lower class limit**

.

Class	Frecency	Lower limite of class	$n_i^c \searrow$
$[10, 20[$	2	10	40
$[20, 30[$	5	20	38
$[30, 40[$	9	30	33
$[40, 50[$	14	40	24
$[50, 60[$	7	50	10
$[60, 70[$	3	60	3
$\sum$	40		


45
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
40
38
33
24
10
3

Chapter 3

Numerical Descriptive Mesures

If one is interested in conveying conciser information, **numerical summaries** are suitable since they describe variables with **numbers,** the features to focus on are:

• **centrality**: describing what is the "typical" value for the variables.

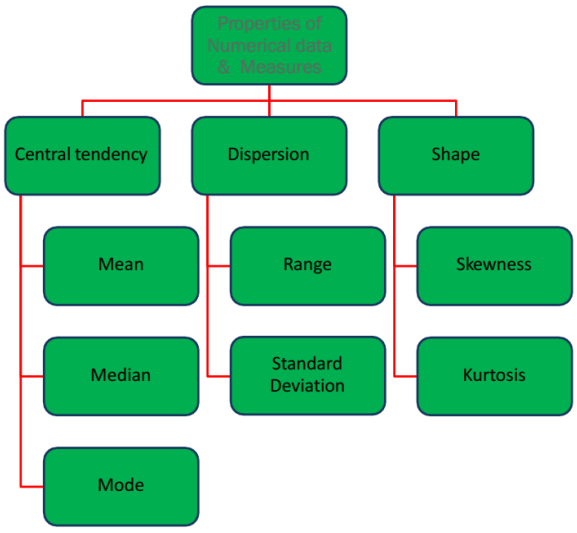
• **variability**: describing whether the observations take similar values or they differ from each other.

The following measures are used to describe a data set:

1 Measures of position ,

2 Measures of spread,

3 Measures of shape.



3.1 Measures of position

Measures of position are used to describethe relative location of an observation

Measurres of central tendency

Measurres of central tendency or **an average** refers to where the data is **centered**,and gives **a single representative value** for a set of usually **unequal values**. This value is the point around which all the values cluster. So, the measure of central tendency is also called a measure of **central location**

Various important measures of central tendency are: **Mode**, **Median**, **Mean**

The Mode

a) discrete data
Definition The mode (M_o) is the modality with the highest observed frequency

$$y_{Mode} = \{x_j : n_j \geq n_i \text{ or } f_j \geq f_i \ \forall j \neq i\}$$

This is the most general notion of centrality and applies to all type of data (numerical and categorical)

b) Grouped data

To calculate the mode for grouped data, you can use the following steps:

▶ Identify the modal class $[a, b[$: This is the class interval with the **highest frequency**.

▶ Use the mode formula

$$M_o = a + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} \times (b - a)$$

where:
n_m = frequency of the modal class
n_{m-1} = frequency of the class before the modal class
n_{m+1} = frequency of the class after the modal class

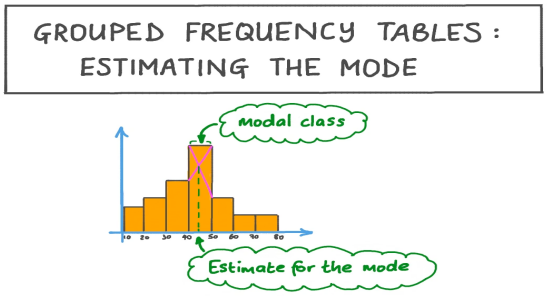
Example 3.1

Class Interval	frequency
$[10, 20[$	5
$[20, 30[$	15
$[30, 40[$	25
$[40, 50[$	10
$[50, 60[$	5

The modal class is $[30, 40[$ because it has a highest frecency 25

$$M_o = 30 + \frac{25 - 15}{(25 - 15) + (25 - 10)} \times (40 - 30) = 34$$

Remark 3.1



✓ You can take the **mid- value** of class modal $[a, b[$ as a values of the mode , it means $\left(M_o \simeq \frac{a+b}{2}\right)$

✓ there can be more than one mode!

Median

Definition The median $(Me)$is the value which **divides** the data into **two equal parts**. 50% of the observations will be less than median value and 50% of the values will be more than the median value.

⋙ **Calculation for Raw data**

1. If the number of observations is **odd** then

$$\mathbf{Me} = \mathbf{x}_{\frac{n+1}{2}} \mathbf{(e.i\ value\ of\ } \frac{n+1}{2}\mathbf{th\ )}$$

observation after the values are arranged in ascending order of magnitude.

Eexample 3.2

The median of $20, 30, 35, 64, 23, 46, 78, 34, 20$

Arranging the data in ascending order; $20, 20, 23, 30, 34, 35, 46, 64, 78$

$Me =$ value of $[\frac{9+1}{2} = 5th$ observation$] = 34$

1. If the number of observations is **even** then e **median** will be the **average** of two **central values**

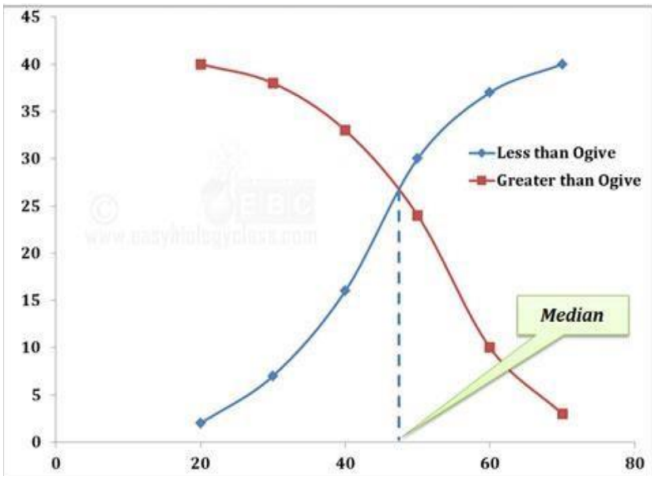
Example 3.3

If the data is the median of 20,30,35,64,23,46,78,34,20,56

Arranging the data in ascending order : 20,20,23,30,34,35,46,56,64,78

$$\begin{aligned} \mathbf{Me} &= \mathbf{value\ of\ (10+1)/2 = 5.5th\ observation} \\ &= \mathbf{(value\ of\ 5th\ observation + value\ of\ 6th\ observation)/2} \\ &= \mathbf{(34+35)/2 = 34.5} \end{aligned}$$

⋙ **Calculation for discrete data**



Me = value of x corresponding to the **cumulative frequency** just **greater than or equal to N/2**

i) Arrange the data is in ascending order

ii) Find the cumulative frecency ($c.f$).; Calculate $N/2$

iii) In $c.f.$ column see the value just $\geq N/2$

iv) **Me** = value of x corresponding to this $c.f.$

Remark 3.2

you can use relative cumulative frequency $f_i^c \nearrow$ such that $f_i^c \nearrow \geq \frac{1}{2} = 0.5$.

Example 3.4

x	2	4	6	$\underbrace{8}$	10	12	14	$\sum$
n_i	4	6	10	12	8	7	3	50
$n_i^c \nearrow$	4	10	20	$\underbrace{32}$	40	47	50	

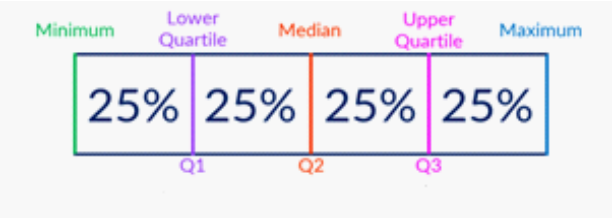
we have $\frac{N}{2} = 25$, so $Me = x_{25} = 8$

3.2 Finding Median Graphically

If both, less than and greater than, cumulative frequency curve is drawn on the same graph, we can easily find the median value. The point in which, both the curve **intersects**, corresponding to the **x-axis**, gives the **median value**.

QUARTILES

It is the value which **divides** the data into **four equal parts**. There are three quartiles.



• Q_1, the first quartile or the lower quartile divides the data in such a way that 25% of the observations will be less than Q_1 value and 75% of the values will be more than the Q_1 value.

• Q_3 , the Third quartile or upper quartile divides the data in such a way that 75% of the observations will be less than Q_3 value and 25% of the values will be more than the Q_3 value

• Q_2,The second quartile is the median. 50% of the observations will be less than median value and 50%of thevalues will be more than the median value

.

⋙ **Calculation for Raw data**

• Q_1 = value of $(n+1)/4$ th observation in ascending order data.
• Me = value of $(n+1)/2$ th observation in ascending order data.
• Q_3 = value of $3(n+1)/4$ th observation in ascending order data.

Exmple 3.5

Find Q_1 and Q_3, for the following data: $20, 30, 35, 64, 23, 46, 78, 34, 20$

Solution

Arranging the data in ascending order: $20, 20, 23, 30, 34, 35, 46, 64, 78$

• $n = 2k + 1 = 9$
► Q_1 = value of $(9+1)/4 = 2.5th$ observation
= value of 2nd observation +0.5(3rd value−2nd value)
$= 20 + 0.5(23 - 20) = 20 + 0.5 \times 3 = 20 + 1.5 = 21.5$
► Q_3 = value of $3(9+1)/4 = 7.5th$ observation
$= 7th$ observation $+0.5(8th$ value $-7th$ value)
$= 46 + 0.5(64 - 46) = 46 + (0.5 \times 18) = 46 + 9 = 55$
• $n = 2k$

Example 3.6

Find Q_1 and Q_3 for this data $20, 30, 35, 64, 23, 46, 78, 34, 20, 56$

Answer:

Arranging the data in ascending order: $20, 20, 23, 30, 34, 35, 46, 56, 64, 78$
► Q_1 = value of $(10+1)/4 = 2.75th$ observation
= 2nd value +0.75 (3rd value $-2nd$ value)
$= 20 + 0.75(23 - 20)$
$= 20 + (0.75 \times 3) = 20 + 2.25 = 22.25$
► Q_3 = value of $3(n + 1\backslash 4)th$ observation

$= (3 \times 2.75 = 8.25th)$ observation
$= 8th$ value $+0.25$ ($9th$ value $-8th$ value)
$= 56 + 0.25(64 - 56)$
$= 56 + 0.25(8) = 56 + 2 = 58$

⋙ **Calculation for discrete data**

o Me = value of x corresponding to the cumulative frequency just $\geq N/2$
o Q_1 = value of x corresponding to the cumulative frequency just $\geq N/4$
o Q_3 = value of x corresponding to the cumulative frequency just $\geq 3N/4$

Example 3.7

Find the median and the quartiles

x	2	4	$\overbrace{6}$	$\underbrace{8}$	$\widetilde{10}$	12	14	$\sum$
n_i	4	6	10	12	8	7	3	50
$n_i^c \nearrow$	4	10	$\overbrace{20}$	$\underbrace{32}$	$\widetilde{40}$	47	50	

• $N/2 = 50/2 = 25$; Therefore $Me = 8$
• $N/4 = 50/4 = 12.5$

Q_1 = value of x corresponding to the cumulative frequency just greater than or equal to $N/4 = 20$. $Q_1 = 6$

• $3N/4 = 37.5$

Q_3 = value of x corresponding to the cumulative frequency just greater than or equal to $3N/4$; Q_3 = value of x corresponding to the cumulative frequency just greater than 37.5 i.e.,40 ;$Q_3 = 10$

⋙ **Calculation for continous da datata**

If $n_i^c \nearrow (Q_i) \geq \dfrac{iN}{4} \Rightarrow Q_i \in b_i, a_i$ and Q_i for all $i = 1, 2, 3$ are given by the following formula

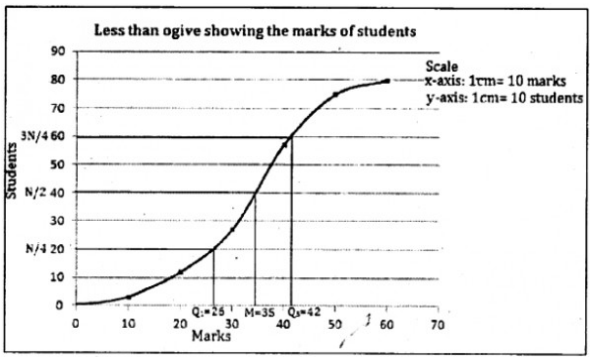
$$Q_i = a_i + \frac{\dfrac{iN}{4} - n^c \uparrow}{n_{Q_i}} \times (b_i - a_i) \text{ for } i = 1, 2, 3$$

$i = 1, 2, 3$
N =Total frecency
$n^c \nearrow$= cumulative frequency before Q_i class
n_{Q_i} = frecuency of $Q_i$class

Example 3.8

find the three quartiles

Marks	n_i	$n_i^c \nearrow$
$[10, 25[$	6	6
$[25, 40[$	20	26
$[40, 55[$	44	70
$[55, 70[$	26	96
$[70, 85[$	3	99
$[85, 100[$	1	100
$\sum$	100	



► $i = 1, Q_1 = a_1 + \dfrac{\dfrac{N}{4} - n^c_{before}}{n_{Q_1}} \times (b_1 - a_1)$

$$\frac{N}{4} = \frac{100}{4} = 25 \Rightarrow Q_1 \in [25, 40[$$

$$\Rightarrow Q_1 = 25 + \frac{25 - 6}{20} \times (15) = 39.25 \in [25, 40[$$

► $i = 2, Q_2 = a_2 + \dfrac{\dfrac{N}{2} - n^c_{before}}{n_{Q_2}} \times (b_2 - a_2)$

$$\frac{N}{2} = \frac{100}{2} = 50 \Rightarrow Q_2 \in [40, 55[$$

$$\Rightarrow Q_2 = Me = 40 + \frac{50 - 26}{44} \times (15) = 48.18 \in [40, 55[$$

► $i = 3, Q_3 = a_3 + \dfrac{3\dfrac{N}{4} - n^c_{before}}{n_{Q_3}} \times (b_3 - a_3)$

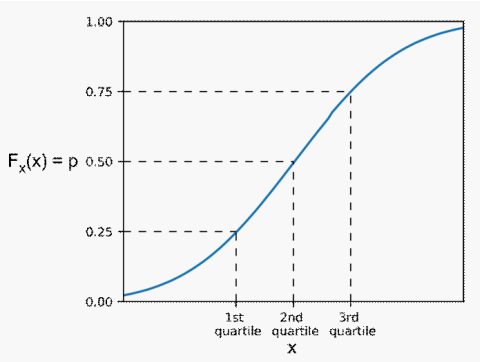
$$\frac{3N}{4} = \frac{3100}{4} = 75 \Rightarrow Q_3 \in [55, 70[$$

$$\Rightarrow Q_3 = 55 + \frac{75 - 77}{26} \times (15) = 57.88 \in [55, 70[$$

Finding quartiles Graphically

Draw horizontal lines from these values on the y-axis to intersect the Ogive, then drop vertical lines from these intersection points to the $x-$axis.

or you can use the **relative frequency**



Arithmetic Mean

Definition Arithmetic mean is the total (sum) of all values divided by the number ofobservations.

⋙ **Calculation of Arithmetic mean for Raw data**

When the observed values are given individually such as $x_1, x_2, x_3 \ldots . x_k$ the arithmetic mean is given by

$$\overline{X} = \frac{x_1 + x_2 + x_3 \ldots + x_k}{N} = \frac{\sum_{i=1}^{k} x_i}{N}$$

Example 3.9

Calculate the arithmetic mean for the following: $1600, 1590, 1560, 1610, 1640, 10$.

Answer

$$\overline{X} = \frac{1600 + 1590 + 1560 + 1610 + 1640 + 10}{6} = \frac{8010}{6} = 1335$$

⋙ **Calculation of Arithmetic mean for discrete data**

Let $\sum_{i=1}^{k}$ be the n values of the variable X with corresponding frequencies $n_1, n_2, n_3 \ldots . n_k$

$$\overline{X} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 \ldots + x_k n_k}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

Example 3.10

Calculate the arithmetic mean

x_i	2	4	6	8	10	12	14	$\sum$
n_i	4	6	10	12	8	7	3	50
$x_i \times n_i$	8	24	60	96	80	84	42	394

Arithmetic mean is $\overline{X} = \dfrac{394}{50} = 78.8$

⋙ **Calculation of Arithmetic mean for continous data**

Let $m_i, m_2, m_3 \ldots .m_n$ be **the mid values** of the class interval of the variable X with corresponding frequency $n_1, n_2, n_3 \ldots .n_k$.

then the arithmetic mean

$$\overline{X} = \frac{m_1n_1 + m_2n_2 + m_3n_3 \ldots + m_kn_k}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} m_in_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} m_in_i}{N}$$

where $\sum n_i = N$ and $m_i = \dfrac{x_i + x_{i+1}}{2}$

Remark 3.3

you can use the relative frequency to calculate the atithmetic mean and we get $\overline{X} = \sum_{i=1}^{k} x_i f_i$ if X is a discret variable, if it is continous you replace x_i with the mid value of class (m_i)

Example 3.11

classes	m_i	n_i	$m_i \times n_i$
$[20, 40[$	$\dfrac{20+40}{2} = 30$	4	120
$[40, 60[$	50	6	300
$[60, 80[$	70	10	700
$[80, 100[$	90	12	1080
$[100, 120[$	110	8	88
$\sum$		40	3080

Arithmetic mean, $\overline{X} = \dfrac{3080}{40} = 77.$

Remark 3.4 The numerical center tendency mod, median and mean coincid $\left(M_o = Me = \overline{X}\right)$ only when the distibution is symmetric.

3.3 Measures of Spread

The Spread refers to how the data **deviates** from the **position** measure, and it gives an indication of the amount of **variation** in the process.

There are different statistics by which we can describe the spread of a data set:

▶ Range: Is the difference between the largest observed value in the data set and the smallest one, so, while considering range great deal of information

is ignored.Often denoted by 'R' such that

$$R = x_{\max} - x_{\min}$$

⋙ **Caracterstics of the range**

• The simplest measure of variability..

• It is good enough in many practical cases.

• It does not make full use of the available data.

• It can be misleading when the data is skewed or in the presence of outliers.

• Just one outlier will increase the range dramatically.

▶ Interquartile range is defined as the difference between the 75th and25th quartiles as

$$IQR = Q_3 - Q_1$$

It covers the centre of the distribution and contains 50% of the observations.

▶ Standard deviation: The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean. **A low value** of the standard deviation indicates that the values **are highly concentrated** around the mean. **A high value** of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may even be far away from the mean

The standard deviation of a sample is denoted by 's'and is computed as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{n} n_i},$$

where $x_1, x_2, x_3 \ldots . x_n$ given as a set of n observation

Remark 3.5 Replace x_i whith m_i in continous data.

Caracterstics of the Standard deviation

• A low standard deviation indicates that the data points are clustered around the mean.

• A large standard deviation indicates that they are widely scattered around the mean.

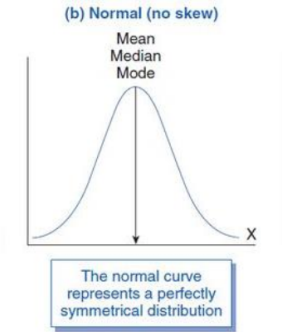
• The standard deviation of a populationis denoted by "σ"

•.Perceived as difficult to understand because it is not easy to picture what it is.

• It is however a more robust measure of variability.

▶ Variance- square of standard deviation, so it is calclated as

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 n_i = \frac{1}{n}\sum_{i=1}^{n} x_i^2 n_i - \left(\overline{X}\right)^2 \\ &= \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 f_i = \sum_{i=1}^{n} x_i^2 f_i - \left(\overline{X}\right)^2. \end{aligned}$$



▸ Cofficient of variation

This is a dimension less quantity that measures the relative variation between two servers observed in different units. The coefficients of variation are obtained by dividing the standard deviation by the mean and multiply it by 100. Symbolically

$$\mathbf{C.V} = \frac{S}{\overline{X}} \times 100\%$$

The distribution with **smaller C.V** is said to **be better**.

3.4 Measures of Shape

The shape helps identifying which descriptive statistic is more appropriate to use in a given situation.

Two common statistics that measure the shape of the data:

- Skewness.
- Kurtosis.

Skewness:

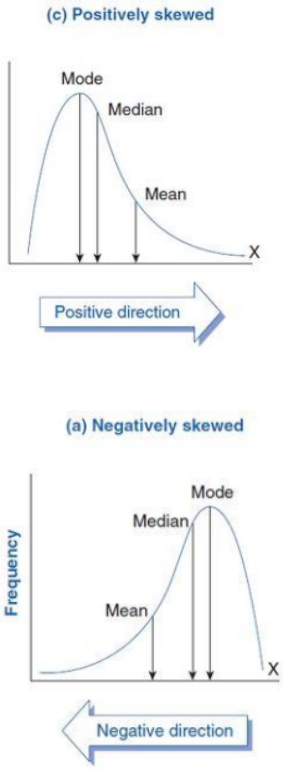
The term '**skewness**' refers to lack of symmetry or departure from symmetry, e.g.,when a distribution is not symmetrical (or is asymmetrical) it is called a **skewed** distribution.

5The measures of skewness indicate the difference between the manner in which the observations are distributed in a particular distribution compared with a **symmetrical** (or **normal**) distribution.

▸ In a symmetrical distribution, the values of mean, median and mode are alike

▸ If the value of **mean is greater than the mode**, skewness is said to be **positiv**e.

▸A distribution is **positively skewed** when the long tail is on the positive side of the peak.



▶ If the value of **mode is greater than mean**, skewness is said to be **negativ**

▶ A distribution is **negatively skewed** when the long tail is on the negative side of the peak

⊛ **Generally**,

If $\overline{X} > Mo$, the skewness is **positive**.

If $\overline{X} < Mo$, the skewness is **negative.**

If $\overline{X} = Mo$, the skewness is **zero**.

Or w can calculate the coefficient of skewness p_1

⊛ Skewness is measured in the following ways:

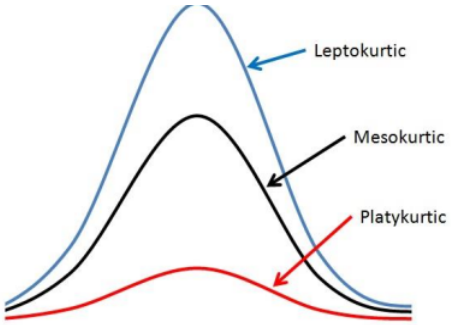
Karl pearsson's coefficient of skewness

$$p_1 = \frac{\text{Mean-Mode}}{\text{Sandard deviation}} = \frac{\overline{X} - Mo}{s}$$

$p_1 = 0$,the distribution is **symmetrical** or **normal**

$p_1 < 0$, the distribution is **negatively skewed**

$p_1 > 0$,he distribution is **positively skewed**



3.5 Kurtosis

Kurtosis refers to the degree of peakedness of a frequency curve.

It tells how tall and sharp the central peak is, relative to a standard bell curve of a distribution.

Kurtosis can be described in the following ways:

• **Platykurtic:** When the **kurtosis** $< \mathbf{0}$, the frequencies throughout the curve are closer to be equal (i.e., the curve is more flat and wide)

• **Leptokurtic**: When the **kurtosis** $> \mathbf{0}$, there are high frequencies in only a small part of the curve (i.e, the curve is more peaked)

• **Mesokurtic:** When the **kurtosis** $= \mathbf{0}$

▶ Kurtosis is measured in the following ways:

$$\gamma_2 = \beta_2 - 3$$

where $\beta_2 = \dfrac{u_4}{u_2^2}$ and $u_4 = \dfrac{1}{N}\sum \left(x_i - \overline{X}\right)^4 n_i, u_2 = \dfrac{1}{N}\sum \left(x_i - \overline{X}\right)^2 n_i.$

Part I

Theory of probabilty

Chapter 1

Combinatorial Analysis

Intoduction

Combinatorial Analysis is an area of Matematics that studies the different **ways combine** or **organise** elements of a set, so this

chapter deals with finding effective methods for counting the number of ways that things can occur.

In fact, many problems in probability theory can be solved simply by counting the number of different ways that a certain event can occur.

Fundamental principal of counting

1. **Multiplication principle**

- If **first operation** can be done by **m way**s and
- Second operation can b done by **n ways**
- Then total number of ways in wich both operation can be done simutaneously equal to $\mathbf{m} \times \mathbf{n}$

2. Addition principle

If certain operation can be performed in **m ways** and **another** operation can be performed in **n ways** then **total number of ways** in wich **either** of the two operation can be performed is $\mathbf{m} + \mathbf{n}$

Example 1.1

Suppose you want to get a policy to get tax relief.

Suppose **3** policy scheme avaiabe with **L.I.C** and and **5** policy scheme with **Birla life insurance**, in how many ways this can be done

Answer: Using addion principle

(**1**) that here first operation is to get policy from **L.I.C** which can be done **3** ways

(**2**) Second operation means to get plicy from **Birla life insurance** which can be done **5** ways

So you can do either of the two opertion

then the total number of ways or choices is $\mathbf{3} + \mathbf{5} = \mathbf{8}$

Example 1.2

▶ How many **3** digit number can be formed by using digits $\mathbf{8}, \mathbf{9}, \mathbf{7}, \mathbf{2}$ without repeating digit?

▶ How many are greater than **800**?

Answer:

The three digit number has three places to be filled

Hendred place Tenth place Unit place

▶ Hendred th place can be filled by **4** ways

▶ After this, tenth place can be filled by **3** ways

▶ After unit place can be filled by **2** ways

Total **3** digits nombers we can form are: $\mathbf{4 \times 3 \times 2 = 24}$

✠ To find number greater than **800**

We uvserved that numbers like **877** or 927 **starting** with either **8** or **9** are greater than 800 in this case

▷ Hence

▷ Hendred th place can be filled by **2** ways

▷ After this, tenth place can be filled by **3** ways

▷ After unit place can be filled by **2** ways

▷ Total **3** digits nombers greater han **800** are: $\mathbf{2 \times 3 \times 2 = 12}$

1.1 Permutation

Permutation of given objects is **arrangement** of that objects in a **specific order**

Example 1.3

Suppose three objects $\mathbf{A}, \mathbf{B}, \mathbf{C}$

So there are **6** different **arrangemant** or **permutation:** $\mathbf{ABC}, \mathbf{ACB}, \mathbf{BAC}, \mathbf{BCA}, \mathbf{CAB}, \mathbf{CBA}$

Note: In permutation order of objects is imortant such that $\mathbf{ABC} \neq \mathbf{ACB}$

Types of Permutation

Permutation can be classified in three different categories:

- **Permutation** of **n** different objects (when **repetition is not allowed**)
- Repetition, where repetition is allowed
- Permutation when the objects **are not distinc**t (Permutation of **multi sets**)

Let us understand all the cases of permutation in details.

Permutation of **n** different objects

⋙ If **n** is a positive integer and **r** is a whole number, such that $\mathbf{r < n}$, then $\mathbf{P}_n^r$) represents the number of all possible arrangements or permutations of **n distinct** objects taken **r** at a time. In the case of permutation without **repetition**, the number of available choices will be reduced each time. It can also be represented as:

$$\mathbf{P}_n^r = \frac{n!}{(n-r)!} = n \times (n-1) \times ...(n-r+1),$$

where $n! = n \times (n-1) \times ...(n-r+1) \times ... \times 4 \times 3 \times 2 \times 1$

Example 1.4: How many 3 letter words with or without meaning can be formed out of the letters of the word **SWING** when repetition of letters is not allowed?

Answer: Here $\mathbf{n = 5}$, as the word **SWING** has 5 letters. Since we have to frame **3** letter words with or without meaning and without , therefore total permutations possible are:

$$\mathbf{P}_n^r = \frac{5!}{(5-3)!} = 5 \times 4 \times 3 = 60$$

⋙ When the number of different object is "**n**,"so the number of ways to arrange all **n** object is

$$\mathbf{P_n = n!}, \qquad (1! = 1; 0! = 1)$$

Permutation when repetition is allowed

We can easily calculate the permutation with repetition. The **permutation with repetition** of objects can be written using the exponent form.

When the number of object is "**n**," and we have "**r**" to be the selection of object, then;

Choosing an object can be in **n** different ways (each time).

Thus, the permutation of objects when repetition is allowed will be equal to,

$$\mathbf{\tilde{P}_n^r = n \times n \times n \times ..... (r\ times) = n^r}$$

Example 1.5 How many 3 letter words with or without meaning can be formed out of the letters of the word **SMOKE** when **repetition** of words is allowed?

The number of objects, in this case, is **5**, as the word **SMOKE** has **5** alphabets.

and $\mathbf{r = 3}$, as **3**-letter word has to be chosen.

Thus, the permutation will be:

Permutation (when repetition is allowed) $\mathbf{\tilde{P}_n^r = n^r = 5^3}$

Permutation of multi-sets

Permutation of **n** different objects when $\mathbf{P_1}$ objects among '**n**' objects are **similar**, $\mathbf{P_2}$ objects of the **second kind** are **similar**, $\mathbf{P_3}$ objects of the **third kind** are **similar** and so on, $\mathbf{P_k}$ objects of the **k th** kind are **similar** and the remaining of all are of a **differen**t **kind**, Thus it forms a **multiset**, where the permutation is given as:

$$\mathbf{P}_n = \frac{n!}{P_1! \times P_2! \times P_3! \times ... \times P_k!}$$

where $\sum_{i=1}^{k} \mathbf{P}_i = \mathbf{n}$

Example:**1.6** How many permutations of letter (**BIOSTATICS**)?

$\mathbf{B = 1, I = 3, O = 1, S = 3, T = 3, A = 1, C = 1}$

$\mathbf{n = 1 + 3 + 1 + 3 + 3 + 1 + 1 = 13}$

$\mathbf{P}_{13} = \dfrac{13!}{1! \times 3! \times 1! \times 3! \times 3! \times 1! \times 1!} = 208828800$__

1.2 Combination

• Suppose we want to select different groups of r objects from a set o
n distinct objects

• The order of objects in a group is not important here

• How many distinct subsets of size r can be formed from n distinct objects?

Example 1.7

• There are five gift cards, each having a different design. You randomly pick two from these five cards.

• How many possible distinct subsets of two cards are there?

• Denote the five designs as $\mathbf{A, B, C, D, E}$

• The possible outcomes are
$\{AB, AC, AD, AE, BC, BD, BE, CD, CE, DE, BA, CA, DA, EA, CB, DB, EB, DC, EC, ED\}$
and the total number of ordered arrangements is 10

• Here, AB and BA are two different ordered arrangements, but they form the same subset. We are interested in finding out the total
number of possible **distinct subsets**. The **order does not matter**.

• Total number of distinct subsets of two cards is **10**

General result

The number of subsets or combinations of size r that can be formed from **n** distinct objects ($\mathbf{r \leq n}$) is given by

$$\mathbf{C}_n^r = \frac{\mathbf{n!}}{\mathbf{(n-r)\,!r!}}, \qquad \text{(Binomial coefficient)}$$

Note:

• Number of combination is fewer than permutation because $\mathbf{C}_n^r = \dfrac{\mathbf{P}_n^r}{\mathbf{r!}} \Rightarrow \mathbf{P}_n^r = \mathbf{C}_n^r \times \mathbf{r!}$

• $\mathbf{C}_n^0 = \mathbf{C}_n^n = 1$, $\mathbf{C}_n^1 = n$

• $\mathbf{C}_n^r = \mathbf{C}_{n-1}^{r-1} + \mathbf{C}_{n-1}^r$, $1 \leq r \leq n$

• $\mathbf{C}_n^r = \mathbf{C}_n^{n-r}$

• $(a+b)^n = \sum_{k=1}^{n} \mathbf{C}_n^k a^k b^k$

Difference Between Permutation and Combination

The major difference between the permutation and combination are given below:

Permutation **Combination**

1. Permutation means the selection of objects, where the order of selection matters

1.The combination means the selection of objects, in which the order of selection does not matter.

2.In other words, it is the arrangement of **r** objects taken out of **n** objects.

2. In other words, it is the selection of r objects taken out of **n** objects **irrespective** of the object arrangement.

3. The formula for permutation is

$$\mathbf{P}_n^r = \frac{\mathbf{n!}}{\mathbf{(n-r)\,!}}$$

3. The formula for combination is

$$\mathbf{C}_n^r = \frac{\mathbf{n!}}{\mathbf{(n-r)\,!r!}}$$

1.3 Application examples

Example 1. In how many ways 6 students can be arranged in a line, such that

(**i**) Two particular students of them are always together

(**ii**) Two particular students of them are never together

Answer:

(**i**) The given condition states that 2 students need to be together, hence we can consider them 1.

Thus, we get **5!** ways, i.e.**120**.

Also, the two children in a line can be arranged in **2!** Ways.

Hence, the total number of arrangements will be, $\mathbf{5!} \times \mathbf{2!} = \mathbf{120} \times \mathbf{2} = \mathbf{240}$ ways

(**ii**) The total number of arrangements of **6** students will be **6!**, i.e. **720** ways.

Out of the total arrangement, we know that two particular students when together can be arranged in 240 ways.

Therefore, total arrangement of students in which two particular students are never together will be $\mathbf{720 - 240}$ ways, i.e. **480** ways.

Example 2. Consider a set having **5** elements $\mathbf{a, b, c, d, e}$. In how many ways **3** elements can be selected (**without repetition**) out of the total number of elements.

Answer: Given $\mathbf{X} = \{\mathbf{a, b, c, d, e}\}$

- 3 are to be selected.
- Therefore, C_{10}^3

Example 3: It is required to seat **5** men and **4** women in a row so that the women occupy the **even** places. How many such arrangements are possible?

Solution:

• We are given that there are **5** men and **4** women.i.e. there are **9 positions**.

• The **even** positions are: **2nd**, **4th**, **6th** and the**8th** places

• These **4** places can be occupied by **4** women in $P_4 = 4! = 4\times3\times2\times1 = 24$ ways

The remaining **5** positions can be occupied by **5** men in $\mathbf{P}_5 = \mathbf{5}! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ ways

Therefore, by **the Fundamental Counting Principle**,

Total number of ways of seating arrangements $= 24 \times 120 = 2880$

1.4 Concept of Probability

What is the probability?

Probability is a measure (or number) used to measure the chance of the occurrence of some event, this number is between 0 and 1.

Experiments and random events.

Definition 2.1. In probability theory, **random experiment** means a **repeatable process** that yields a **result** or an **observation**.

Example 2.1

▶ Tossing a coin, rolling a die, extracting a ball from a box are **random experiments**.

▶ When tossing a coin, we get one of the following elementary results: (**heads**), (**tails**):

▶ When throwing a d**ie**, if we denote by (1) the appearance of the face with **one dot**, with(2) the appearance of the face with **two dots**, etc., then we get the following elementary results: $(1), (2), (3), (4), (5), (6)$.

Definition 2.2. **A random event** is an event that either happens or fails to happen as a result of an experiment.

▶ When tossing a coin, the event **(heads)** may happen or may fail to happen, so this is a random event.

▶ **A random event** depends on the combined action of several factors which may not have been taken into consideration when setting up the experiment.

Sample space of an experiment

Definition 2.3. The **sample space** (S) of an experiment is a set of all **elementary results** (**possible outcomes**) of the experiment. The elements of a sample space are called **sample points**.

Certain event. Impossible event.

There are **two special events** for every experiment: the **certain event** and the **impossible even**t.

Definition 2.4. **The certain event** (**denoted by S**) is an event which happens with **certitude** at each repetition of an experiment.

Example 2.2.

• When tossing a coin, the event (one of the two faces appears) is a certain event of the experiment.

• When rolling a die, the event (one of the six faces appears) is a certain event of the experiment.

Definition 2.5. The impossible event (denoted by $\not\in$) is an event which never happens in a random experiment.

▶ When extracting a ball from a box which contains only white balls, the event (a red ball is extracted) is an impossible event.

Contrary events.

In the case of rolling a die, let's denote by **A** the event consisting of the appearance of one of the faces **2** or **5**, and **B** the event consisting of the appearance of one of the faces **1, 3, 4** or **6**.

We observe that if the event **A** does not take place, then the event B takes place, and the other ay round.

Definition 2.6. The contrary of an event **A** is an event **B** satisfying the property that, at any repetition of the experiment, if the event **A** occurs then **B** does not occur, and if the event **B** occurs then **A** does not occur. The events **A** and **B** are also called **mutually exclusive events**.

▶ If **B** is the contrary of **A** then **A** is the contrary of **B**.

▶ We denote the contrary of an event **A** by $\overline{A}$ or **CA**

Definition 2.7 The events **A** and **B** are **compatible** if they can occur **simultaneously**

▶ When throwing a die, the event **A=(an even number appears)** and the event **B=(one of the numbers 2 or 6 appears**), are **compatible**. If the **outcome** of the experiment is the appearance of the face with the **number 2**, then **both** events **A** and **B** take place.

Definition 2.8. The events **A** and **C** are **incompatible** if they **cannot occur** simultaneously.

▶ When rolling a die, the events **A = (an even number appears)** and **C = (an odd number appears)** are **incompatible**. They cannot take place at the **same time**. One may notice that the events **A** and **C** are **contrary events**.

On the other hand, if we consider the event **D = (the number 5 appears**), we can see that A and **D** are **incompatible**, but they are **not contrary events**: the non-occurrence of **A** does not imply the occurrence of **D**.

Definition 2.9. The events $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_n$ are **compatible** if they can **occur simultaneously**.

▶ When throwing a die, the events:

$\mathbf{A_1}$ = (one of the faces 2 or 4 appears)

$\mathbf{A_2}$= (one of the faces 2 or 6 appears)

$\mathbf{A_3}$= (one of the faces 2, 4 or 6 appears)

are compatible: if the outcome of the experiment is the appearance of the face with the number 2, all three events take place.

Event implied by another event

Definition 2.10.. We say that the event **A implies** the event **B** (or the event **B** is **implied** by the event **A**) if the occurrence of the event **A** means that the event **B occurs as well**.

▶ When throwing a die, the event **A** =(one of the faces 1 or 3 appears) implies the event **B** = (one of the faces 1, 2, 3 or 5 appears).

▶ Any event implies the certain event.

Operations with events.

▶ The set of results representing the event **A** is **included** in the set of results representing the event **B**: $\mathbf{A} \subset \mathbf{B}$.

▶ The sets representing two **incompatible** events are **disjoint**.

Definition 2.11. The **union** $\mathbf{A} \cup \mathbf{B}$ of two events A and **B** is the event which takes place when at **least one** of the events **A or B** occur.

Definition 2.12. The intersection $\mathbf{A} \cap \mathbf{B}$ of two events **A** and **B** is the event which occurs when **both events A** and **B** take place at the **same time**.

Example 2.3 . For the experiment of rolling one die, let's consider the following events: $A = \{1; 2; 5\}$g; $B = \{3; 4; 5\}$

The event **A** occurs if one of the following results is obtained: {1}, {2}or {5}; the event **B** occurs if one of the results {3}, {4} or {5} is obtained.

To insure that at least one of the events **A or B** take place, we must obtain one of the results {1}, {2},{3}, {4}, {5}. Therefore:

$\mathbf{A} \cup \mathbf{B} = \{12, 3, 4, 5\}$

On the other hand, both events take place at the **same time** only in the case when the face with,number **5** is obtained, so we get:

$\mathbf{A} \cap \mathbf{B} = \{5\}$

1.5 Probability of an event

Definition 2.13. If the events of the sample space associated to an experiment are equally possible, we say that they are equally probable and the probability of each event is equal to the inverse of the number of events from the sample space.

Example 2.4. Let's consider the experiment of tossing a coin and the sample space **S** which includes the two possible elementary results of this experiment:

H =(the outcome is heads)

T =(the outcome is tails)

$\mathbf{S} = \{\mathbf{H}, \mathbf{T}.\}$

As these two events **H** and **T** are equally possible, it is natural to estimate (to measure) the chance of occurrence of each of them by $\frac{1}{2}$= the inverse of the number of elementary events from **S**.

Example 2.5. We consider the experiment of rolling a die and the associated sample space $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$. As these six events are equally possible, it is natural to evaluate the chance of occurrence of each of them by $\frac{1}{6}$ =the inverse of the number of events from **S**.

Example 2.6. Consider the experiment of tossing two coins and the sample space $\mathbf{S} = \{(\mathbf{H}, \mathbf{H}), (\mathbf{H}, \mathbf{T}), (\mathbf{T}; \mathbf{H}), (\mathbf{T}, \mathbf{T})\}$. As the four events are equally possible, the chance of occurrence of each of them is evaluated by $\frac{1}{4}$ =the inverse of the number of events from **S**

Example 2.7.When tossing two coins and considering the sample space $\mathbf{S} = \{(\text{same symbol}), (\text{difierent symbols})\}$

as the events are equally possible, we evaluate the chance of occurrence of each event by $\frac{1}{2}$= the inverse of the number of events from **S**.

Definition 2.14

Probability of an event: If the experiment has (**n**) equally likely outcomes, then the probability of the event (**A**) is:

$$P(A) = \frac{n(A)}{n(S)} = \frac{\text{number of outcoms of } \mathbf{A}}{\text{number of outcoms of } \mathbf{S}}$$

1.6 Finite sample space. Elementary event.

Definition 1.15. A flnite sample space associated to an experiment is a flnite set $\mathbf{S} = \mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}$ of abstract elements.

The parts of the set S are called **events**. An **event** is called **elementary** if it consists of **a single point** of the space **S**. The **empty part** of **S** $\emptyset$, is called **impossible event**, and **S** is called **certain event**.

Example 2.8. The teachers of a school are asked the following questions:

1. Is it necessary to modernize the school?

2.Is it necessary to build a sport facility for the school?

The answers given by one teacher can be one of the followings: e_1 = (YES; YES), e_2 =?(YES; NO), $\mathbf{e_3}$ = (NO; YES),$\mathbf{e_4}$= (NO; NO). The set $S = \{\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}, \mathbf{e_4}\}$ is a possible sample space for this experiment for a given teacher. The subsets of this space are:

$\mathbf{P(S)} = \{\emptyset, \{\mathbf{e_1}\}, \{\mathbf{e_1}\}, \{\mathbf{e_1}\}, \{\mathbf{e_1}\}, \{\mathbf{e_2}, \mathbf{e_3}\}, \{\mathbf{e_2}, \mathbf{e_4}\}, \{\mathbf{e_3}, \mathbf{e_4}\}, \{\mathbf{e_1}, \mathbf{e_3}\}, \{\mathbf{e_1}, \mathbf{e_2}\}, \{\mathbf{e_1}, \mathbf{e_4}\}, \{\mathbf{e_2}, \mathbf{e_3}, \mathbf{e_4}\}, \{\mathbf{e_1}, \mathbf{e_2},$

Each of these subsets is an event. The subsets $\mathbf{E}_1$= $\{e_1\}$, $\mathbf{E_2}$ = $\{e_2\}$, $\mathbf{E_3}$=$\{\mathbf{e_3}\}$, $\mathbf{E_4}$=$\{\mathbf{e_4}\}$.

contain only one point and they are **elementary events.** Any event (besides the impossible event) is a union of elementary events.

1.7 Axiomatic deflnition of probability

Defnition 2.16. We call probability on the sample space $S = \{\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}, \mathbf{e_4}\}$ a function $\mathbf{P}$ which associates to every event $\mathbf{A} \in \mathbf{P(S)}$ a number $\mathbf{P\ (A)}$, called **probability of A**, such that the following conditions (axioms) are fulfilled:

i) $\mathbf{P(A)} \geq 0$, $\forall \mathbf{A} \in \mathbf{P(S)}$,

ii) $\mathbf{P(S)} = \mathbf{1}$,

iii) $A \cap B = \emptyset \Rightarrow \mathbf{P(A \cup B)} = \mathbf{P(A)} + \mathbf{P(B)}$; $\forall \mathbf{A; B} \in \mathbf{P(S)}$.

The function $\mathbf{P} : \mathbf{P(S)} \to \mathbf{R_+^1}$ is called **probability measure**.

The sample space $\mathbf{S}$ together with the probability measure $\mathbf{P}$ (the pair $(\mathbf{S}; \mathbf{P})$) is called **probability space**

.

Proposition 2.1. Let $\mathbf{A} \in \mathbf{P(S)}$.

1.If $\mathbf{A} = \emptyset$, then $\mathbf{P(A)} = \mathbf{0}$.

2. If $\mathbf{A} = \{\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_k}\}$ then $\mathbf{P(A)} = \sum_{i=1}^{k} \mathbf{P}\left(\{\mathbf{e_i}\}\right)$

Consequence 2.1. If all $\mathbf{n}$ elementary events $\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}$ of the sample space $\mathbf{S}$ have the **same probability** (**are equally probable**), i.e. $\mathbf{P}(\{\mathbf{e_i}\}) = \mathbf{P}(\{\mathbf{e_j}\}), \forall$ $\mathbf{i}, \mathbf{j} = \overline{\mathbf{1, n}}$ then $\mathbf{P}(\{\mathbf{e_i}\}) = \dfrac{\mathbf{1}}{n}$ $\forall$ $\mathbf{i} = \overline{\mathbf{1, n}}$

Proposition .2.2 For any $\mathbf{A} \in \mathbf{P(S)}$, we have:

$$\mathbf{P}(\overline{\mathbf{A}}) = \mathbf{1} - \mathbf{P(A)}$$

Proof. As $\mathbf{A} \cap \overline{\mathbf{A}} = \emptyset$; and $\mathbf{A} \cup \overline{\mathbf{A}} = \mathbf{S}$, we have $\mathbf{P(A)} + \mathbf{P}(\overline{\mathbf{A}}) = \mathbf{P(S)} = \mathbf{1}$, so $\mathbf{P}(\overline{\mathbf{A}}) = \mathbf{1} - \mathbf{P(A)}$

Proposition 2.3. If $\mathbf{A}$, $\mathbf{B} \in \mathbf{P(S)}$ and $\mathbf{A} \subset \mathbf{B}$ then $\mathbf{P(A)} \leq \mathbf{P(B)}$

Proposition 2.4. If $\mathbf{A_1}, \mathbf{A_2}, ..., \mathbf{A_n} \in \mathbf{P(S)}$ and $\mathbf{A_i} \cap \mathbf{A_j} = \emptyset, \forall \mathbf{i} \neq \mathbf{j}$, then

$$P\left(\overset{n}{\underset{i=1}{\cup}} \mathbf{A_i}\right) = \sum_{i=1}^{n} P\left(\mathbf{A_i}\right)$$

Proposition 2.5

For any $\mathbf{A}, \mathbf{B} \in \mathbf{P(S)}$ the following equality holds:

$$\mathbf{P(A \cup B)} = \mathbf{P(A)} + \mathbf{P(B)} - \mathbf{P(A \cap B)}$$

1.8 Independent and dependent events.

Definition 2.16. The events $\mathbf{A}$ and $\mathbf{B}$ from $\mathbf{P(S)}$are called **independent** if

$$\mathbf{P(A \cap B)} = \mathbf{P(A)} \times \mathbf{P(B)}$$

Definition 2.17 We say that the events $\mathbf{A_1}, \mathbf{A_2}, ..., \mathbf{A_n}$ are **totally independent**, or **independent**, if for any $\mathbf{1} \leq \mathbf{i}_1 < \mathbf{i}_2 < ::: < \mathbf{i}_k \leq \mathbf{n}$, we have:

$$\mathbf{P}(\mathbf{A}_{i1} \cap \mathbf{A}_{i2} \cap ... \cap \mathbf{A}_{ik}) = \mathbf{P(A_{i1})} \times \mathbf{P(A_{i2})} \times ... \times \mathbf{P(A_{ik})}$$

Remark

▶ The independence of the events $\mathbf{A_1}, \mathbf{A_2}, ..., \mathbf{A_n}$ means that

$$C_n^2 + C_n^3 + ... + C_n^n = 2^n - \mathbf{n - 1}$$

▶ The independence of three events $\mathbf{A_1}, \mathbf{A_2}, \mathbf{A_3}$ means that we must have:

$$\begin{aligned} \mathbf{P}(\mathbf{A}_1 \cap \mathbf{A}_2) &= \mathbf{P}(\mathbf{A}_1) \times \mathbf{P}(\mathbf{A}_2) \\ \mathbf{P}(\mathbf{A}_1 \cap \mathbf{A}_3) &= \mathbf{P}(\mathbf{A}_1) \times \mathbf{P}(\mathbf{A}_3) \\ \mathbf{P}(\mathbf{A}_1 \cap \mathbf{A}_2 \cap \mathbf{A}_3) &= \mathbf{P}(\mathbf{A}_1) \times \mathbf{P}(\mathbf{A}_2) \times \mathbf{P}(\mathbf{A}_3) \end{aligned}$$

▶ If $\mathbf{A}$ and B are independent events, then the events $\mathbf{A}$ and $\mathbf{CB}$, $\mathbf{CA}$ and $\mathbf{B}$,$\mathbf{CA}$ and $\mathbf{CB}$ are also **independent**.

Definition 2.18. We say that the events $\mathbf{B_1}, \mathbf{B_2}, ..., \mathbf{A_k} \in \mathbf{P(S)}$ form a partition of the sample space $\mathbf{S}$ if the following conditions are fulfilled:

i) $\mathbf{B_i} \cap \mathbf{B_j} = \emptyset$, **for** $\mathbf{i \neq j}$,

ii) $\overset{\mathbf{k}}{\underset{\mathbf{i=k}}{\cup}} \mathbf{B_i} = \mathbf{S}$,

iii) $\mathbf{P(B_i)} > \mathbf{0}$, $\forall \mathbf{i} = \mathbf{1, 2, ..., k}$.

The events of a partition of the sample space are called **hypotheses**.

Definition 2.18. Let $\mathbf{A_1}, \mathbf{A}, ..., \mathbf{A_n}$, $\mathbf{B_1}, \mathbf{B_2}, ..., \mathbf{A_k}$ two partitions of the sample space $\mathbf{S}$. We say that these partitions are independent if

$$\mathbf{P}(\mathbf{A}_i \cap \mathbf{B}_j) = \mathbf{P}(\mathbf{A}_i) \cap \mathbf{P}(\mathbf{B}_j)$$

for any $\mathbf{i}$, $\mathbf{j}$, $\mathbf{i} = \mathbf{1, 2, ..., n}$, $\mathbf{j} = \mathbf{1, 2, ..., k}$.

Example 2.9. When tossing two coins, consider the following events:

$\mathbf{A_1}$ ="obtain heads on the first coin",

$\mathbf{A_2}$ ="obtain tails on the second coin",

$\mathbf{A_3}$="obtain heads and tails".

The events $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ are not $\mathbf{3}-$independent.

Answser: A sample space of this experiment is

$\mathbf{S} = \{\mathbf{e_1} = (\mathbf{H}, \mathbf{H}), \mathbf{e_2} = (\mathbf{H}, \mathbf{T}), \mathbf{e_3} = (\mathbf{T}; \mathbf{H}), \mathbf{e_4} = (\mathbf{T}, \mathbf{T})\}$:

The events $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ are

$\mathbf{A_1} = \{\mathbf{e_1}, \mathbf{e_2}\}, \mathbf{A2} = \{\mathbf{e_2}, \mathbf{e_4}\}, \mathbf{A3} = \{\mathbf{e_2}, \mathbf{e_3}\}$:

We have:

• $\mathbf{A_1} \cap \mathbf{A_2} = \{\mathbf{e_2}\} \Rightarrow \mathbf{P}(\mathbf{A_1} \cap \mathbf{A_2}) = \mathbf{P}(\mathbf{A_1}) \times \mathbf{P}(\mathbf{A_2}) = \dfrac{\mathbf{1}}{4}$

• $\mathbf{A_1} \cap \mathbf{A_3} = \{\mathbf{e_2}\} \Rightarrow \mathbf{P}(\mathbf{A_1} \cap \mathbf{A_3}) = \mathbf{P}(\mathbf{A_1}) \times \mathbf{P}(\mathbf{A_3}) = \dfrac{\mathbf{1}}{4}$

$\mathbf{A_1} \cap \mathbf{A_2} \cap \mathbf{A_3} = \{\mathbf{e_2}\} \Rightarrow \mathbf{P}(\mathbf{A_1} \cap \mathbf{A_2} \cap \mathbf{A_3}) = \dfrac{\mathbf{1}}{\mathbf{4}} \neq \dfrac{\mathbf{1}}{\mathbf{8}} = \mathbf{P}(\mathbf{A_1}) \times \mathbf{P}(\mathbf{A}_2) \times \mathbf{P}(\mathbf{A_3}) = \dfrac{\mathbf{1}}{4}$

1.9 Conditional probability

We illustrate the meaning of "**conditional probability**" using the following example:

Example 2.10 Consider rolling two dice. Let **a** be the number which appears on the first die and **b** the number appearing on the second die. What is the probability that $\mathbf{b = 3}$, knowing that $\mathbf{a + b > 8}$?

Answser:

The sample space associated to this experiment is the set **S** of the following pairs:

(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)

All these events are equally probable and $\mathbf{P}((\mathbf{i};\mathbf{j})) = \dfrac{\mathbf{1}}{36}$, for any $\mathbf{i} = \overline{\mathbf{1,\ 6}}$, $\mathbf{j} = \overline{\mathbf{1,6}}$. Looking at all the elementary events of the sample space **S**, only in the case of the events $(6;3), (5;4), (4;5), (3;6), (6;4), (5;5), (4;6), (6;5), (5;6), (6;6)$ the condition $\mathbf{a + b > 8}$ is fulfilled. We consider the set $\acute{\mathbf{S}}$ formed by these events:

$\acute{\mathbf{S}} = \{(6;3); (5;4); (4;5); (3;6); (6;4); (5;5); (4;6); (6;5); (5;6); (6;6)\}$

The set $\acute{\mathbf{S}}$ is another sample space associated to this experiment, built up by taking into account that $\mathbf{a + b > 8}$. The elementary events of the sample space $\acute{\mathbf{S}}$ are equally probable and their probability is $\dfrac{1}{10}$.

We find only one element of the sample space $\acute{\mathbf{S}}$ for which $\mathbf{b = 3} : (\mathbf{6}, \mathbf{3})$. Therefore, in the sample space $\acute{\mathbf{S}}$, the probability of the event $\mathbf{b = 3}$ is $\dfrac{1}{10}$. This result will be called **the probability of "b = 3" conditioned by "a + b > 8".**

Definition 2.19. The probability of the event **A conditioned by the occurrence of the event B** is denoted by $\mathbf{P}(\mathbf{A} \mid \mathbf{B})$ or $\mathbf{P_B}(\mathbf{A})$ and is deflned by

$$\mathbf{P}(\mathbf{A} \mid \mathbf{B}) = \frac{\mathbf{P}(\mathbf{A} \cap \mathbf{B})}{\mathbf{P}(\mathbf{B})}, \qquad \mathbf{P}(\mathbf{B}) \neq 0$$

Instead of "probability of the event **A** conditioned by the occurrence of the event B" we simply say "**probability of A, given B**".

The reduced sample space is **B** (**the conditioning event**).

Proposition 2.6. For a flxed event $\mathbf{B} \in \mathbf{P}(\mathbf{S})$ such that $\mathbf{P}(\mathbf{B}) \neq \mathbf{0}$, for any two events $\mathbf{A_1}$; $\mathbf{A_2}$ from $\mathbf{P}(\mathbf{S})$, we have:

1) $0 \leq \mathbf{P}(\mathbf{A}_1 \mid \mathbf{B}) \leq 1$,

2) $\mathbf{P}(\mathbf{S} \mid \mathbf{B}) = \mathbf{1}$;

3) $\mathbf{A_1}, \mathbf{A_2}$ **– incompatible** $\Rightarrow \mathbf{P}((\mathbf{A}_1 \cup \mathbf{A_2}) \mid \mathbf{B}) = \mathbf{P}(\mathbf{A_1} \mid \mathbf{B}) + \mathbf{P}(\mathbf{A_2} \mid \mathbf{B})$.

Theorem 2.1 If **A** and **B** are independent events with non-zero probabilities, then:

$$\mathbf{P}(\mathbf{A} \mid \mathbf{B}) = \mathbf{P}(\mathbf{A}) \text{ and } \mathbf{P}(\mathbf{B} \mid \mathbf{A}) = \mathbf{P}(\mathbf{B})$$

Theorem 2.2 (**Total probability formula**). If the events $\mathbf{A_1}, \mathbf{A}, ..., \mathbf{A_n}$ form a partition of the sample space$\mathbf{S}$ and $\mathbf{X} \in \mathbf{P(S)}$, then:

$$\begin{aligned}\mathbf{P(X)} &= \sum_{i=1}^{n} \mathbf{P}\left(\mathbf{X} \mid \mathbf{A_i}\right) \times \mathbf{P}(\mathbf{A_i}) \\ &= \mathbf{P}\left(\mathbf{X} \mid \mathbf{A}_1\right) \times \mathbf{P}(\mathbf{A}_1) + \mathbf{P}\left(\mathbf{X} \mid \mathbf{A}_2\right) \times \mathbf{P}(\mathbf{A}_2) ... + \mathbf{P}\left(\mathbf{X} \mid \mathbf{A}_n\right) \times \mathbf{P}(\mathbf{A}_n)\end{aligned}$$

Theorem 2.3 (Bayes' formula). If the events $\mathbf{A_1}, \mathbf{A}, ..., \mathbf{A_n}$ form a partition of the sample space $\mathbf{S}$ and are the cause of the occurrence of an event $\mathbf{X}$, then:

$$\mathbf{P}(\mathbf{A}_k \mid \mathbf{X}) = \frac{\mathbf{P}(\mathbf{X} \mid \mathbf{A_k}) \times \mathbf{P}(\mathbf{A_k})}{\sum_{i=1}^{n} \mathbf{P}\left(\mathbf{X} \mid \mathbf{A_i}\right) \times \mathbf{P}(\mathbf{A_i})}$$

Definition 2.19. The probabilities $\mathbf{P}(\mathbf{A_i})$, $\mathbf{P}(\mathbf{X} \mid \mathbf{A_i})$, $\mathbf{i} = \overline{\mathbf{1,n}}$ are called **prior probabilities** and $\mathbf{P}(\mathbf{A_i} \mid \mathbf{X})$ are called **posterior probabilities**. The event $\mathbf{X}$ is called **e**vidence.

Before we receive the evidence, we have a set of **prior probabilities** $\mathbf{P}(\mathbf{A_i})$, $\mathbf{i} = \overline{\mathbf{1,n}}$ for **the hypotheses**. If we know the correct **hypothesis**, we know the probability for the evidence. That is, we know $\mathbf{P}(\mathbf{X} \mid \mathbf{A_i})$, $\mathbf{i} = \overline{\mathbf{1,n}}$. If we want to find the probabilities for the **hypothesis**, given the evidence, that is, we want to find $\mathbf{P}(\mathbf{A_i} \mid \mathbf{X})$, we can use the **Bayes' formula.**

Example 2.11. Consider two boxes. The first one contains **2 white balls** and 3 black balls, and the second contains **7 white ball**s and **5 black balls**. The event $\mathbf{A_1}$ means choosing **the first bo**x, and the event $\mathbf{A_2}$ means choosing **the second box**. It is known that the probability of the event $\mathbf{A_1}$ is $\mathbf{P(A_1) = 0.4}$, and the probability of the event $\mathbf{A_2}$ is $\mathbf{P(A_2) = 0.6}$. We randomly choose **one box** and a **black ball**. What is the probability that this **black ball** is chosen from the **first box** and **second box**?

Answer: Let $\mathbf{X}$ be the event "**a black ball has been extracted**". By **Bayes' formula**, we have:

$$\mathbf{P}(\mathbf{A}_1 \mid \mathbf{X}) = \frac{\mathbf{P}(\mathbf{X} \mid \mathbf{A}_1) \times \mathbf{P}(\mathbf{A}_1)}{\mathbf{P}\left(\mathbf{X} \mid \mathbf{A_1}\right) \times \mathbf{P}(\mathbf{A}_1) + \mathbf{P}(\mathbf{X} \mid \mathbf{A}_2) \times \mathbf{P}(\mathbf{A}_2)} = \frac{0.4 \times \frac{3}{5}}{0.4 \times \frac{3}{5} + 0.6 \times \frac{5}{12}} \simeq 0.49$$

$$\mathbf{P}(\mathbf{A}_2 \mid \mathbf{X}) = \frac{\mathbf{P}(\mathbf{X} \mid \mathbf{A}_2) \times \mathbf{P}(\mathbf{A}_2)}{\mathbf{P}\left(\mathbf{X} \mid \mathbf{A_1}\right) \times \mathbf{P}(\mathbf{A}_1) + \mathbf{P}(\mathbf{X} \mid \mathbf{A}_2) \times \mathbf{P}(\mathbf{A}_2)} = \frac{0.6 \times \frac{5}{12}}{0.4 \times \frac{3}{5} + 0.6 \times \frac{5}{12}} \simeq 0.52$$

Univ. 8 Mai 1945 2023-2024

Tutorial Series 1

Exercise 1: In a survey, people were asked how many times they visited a store before making a major purchase. The results are shown in the Table below.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

1. What is the population studied and the total size of the sample?
2. Find the variable and its type.
3. Find relative frequencies and cumulative relative frequencies for the survey.
4. What is the number of people who visited the store once?
5. Find the number of people who visited the store 3 times at least.
6. Find the number of people who visited the store more than 3 times.
7. Draw the polygon and bar graph.

Exercise 2

Given the following series of data on Gender and Height for 8 patients, for each variable fill in a frequency table (for Height, use the classes 140-160,160-170,170-20)

1. What are the variables and their types?
2. Complete the table with the central values, the class widths..
3. Compute the mean of the Height for the eight patients. Use first the series of individual data,
4. Compute the mean starting from the frequency table.
5. Do you expect to find a difference, and why?
6. Create an appropriate graph to represent the frequency distribution.

Id	Height	Gender
1	165	M
2	157	F
3	168	F
4	178	M
5	171	F
6	182	M
7	182	M
8	153	F

Exercise 3: The manager of a store selling laptops recorded the number of laptops sold per day for fifty days. The data series is represented in the table below:

7	13	8	10	9	12	10	8	9	10	6	14	7
15	9	11	12	11	12	11	12	5	14	8	10	14
12	8	5	7	13	16	12	11	9	11	11	12	12
9	14	5	14	9	14	11	13	10	11	9		

1. What are the population studied the variable and its type and the modalities.
2. Find relative frequencies, cumulative relative frequencies, and the range.
3. Find the number of days the store sold 15 items.
4. Find the number of days the store sold more than 12 items.
5. Find the number of days the store sold at most 12 items.