CHAPTER 1

Fundamentals of Statistics

1.1 INTRODUCTION

In the modern world of information and communication technology, the importance of statistics is very well recognized by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on. As of today, there is no other human walk of life, where statistics cannot be applied.

Statistics is concerned with the scientific method of collecting, organizing, summarizing, presenting and analyzing statistical information (data) as well as drawing valid conclusion on the basis of such analysis. It could be simply defined as the "science of data". Thus, statistics uses facts or numerical data, assembled, classified and tabulated so as to present significant information about a given subject. Statistic is a science of understanding data and making decisions in the face of randomness.

The study of statistics is therefore essential for sound reasoning, precise judgment and objective decision in the face of up- to- date accurate and reliable data. Thus many researchers, educationalists, business men and government agencies at the national, state or local levels rely on data to answer operations and programs.

In this chapter, you begin by learning five basic words—population, sample, variable, parameter, and statistic (singular)—that identify the fundamental concepts of statistics. These five words, and the other concepts introduced in this chapter, help you explore and explain the statistical methods discussed in later chapters.

1.2 The First Three Words of Statistics

You've already learned that statistics is about analyzing things. Although *numbers* was the word used to represent things in the opening of this chapter, the first three words of statistics, *population*, *sample*, and *variable*, help you to better identify what you analyze with statistics.

Population

CONCEPT. All the members of a group about which you want to draw a conclusion.

EXAMPLES. All DZ citizens who are currently registered to vote, all patients treated at a particular hospital last year, the entire daily output of a cereal factory's production line.

Sample

CONCEPT. The part of the population selected for analysis.

EXAMPLES. The registered voters selected to participate in a recent survey concerning their intention to vote in the next election, the patients selected to fill out a patient satisfaction questionnaire, 100 boxes of cereal selected from a factory's production line.

Variable

CONCEPT. A characteristic of an item or an individual that will be analyzed using statistics.

EXAMPLES. Gender, the party affiliation of a registered voter, the household income of the citizens who live in a specific geographical area, the publishing category (hardcover, trade paperback, mass-market paperback, textbook) of a book, the number of televisions in a household.

INTERPRETATION. All the variables taken together form the data of an analysis. Although people often say that they are analyzing their data, they are, more precisely, analyzing their variables. (Consistent to everyday usage, the authors use these terms interchangeably throughout books.)

You should distinguish between a variable, such as gender, and its value for an individual, such as male. An observation is all the values for an individual item in the sample. For example, a survey might contain two variables, gender and age. The first observation might be male, 40. The second observation might be female, 45. The third observation might be female, 55. A variable is sometimes known as a column of data because of the convention of entering each observation as a unique row in a table of data. (Likewise, some people refer to an observation as a row of data.)

Variables can be divided into the following types:

	Categorical Variables	Numerical Variables
Concept	The values of these variables are selected from an established list of categories.	The values of these variables involve a counted or measured value.
Subtypes	*Nominal scale* is a naming scale, where variables are simply "named" or labeled, with no specific order. *Ordinal scale* has all its variables in a specific order, beyond just naming them.	*Discrete.* values are counts of things. *Continuous.* values are measures and any value can theoretically occur, limited only by the precision of the measuring process.
Examples	Gender, a variable that has the categories "male" and "female." Academic major, a variable that might have the categories "English," "Math," "Science," and "History," among others.	The number of people living in a household, a discrete numerical variable. The time it takes for someone to commute to work, a continuous variable.

1.3 The Fourth and Fifth Words

After you know what you are analyzing, or, using the words of the previous Section, after you have identified the variables from the population or sample under study, you can define the parameters and statistics that your analysis will determine.

Parameter

CONCEPT. A numerical measure that describes a variable (characteristic) of a population.

EXAMPLES. The percentage of all registered voters who intend to vote in the next election, the percentage of all patients who are very satisfied with the care they received, the mean weight of all the cereal boxes produced at a factory on a particular day.

Statistic

CONCEPT. A numerical measure that describes a variable (characteristic) of a sample (part of a population).

EXAMPLES. The percentage of registered voters in a sample who intend to vote in the next election, the percentage of patients in a sample who are very satisfied with the care they received, the mean weight of a sample of cereal boxes produced at a factory on a particular day.

INTERPRETATION. Calculating statistics for a sample is the most common activity because collecting population data is impractical in most actual decision-making situations.

1.4 The Branches of Statistics

You can use parameters and statistics either to describe your variables or to reach conclusions about your data. These two uses define the two branches of statistics: **descriptive statistics** and **inferential statistics**.

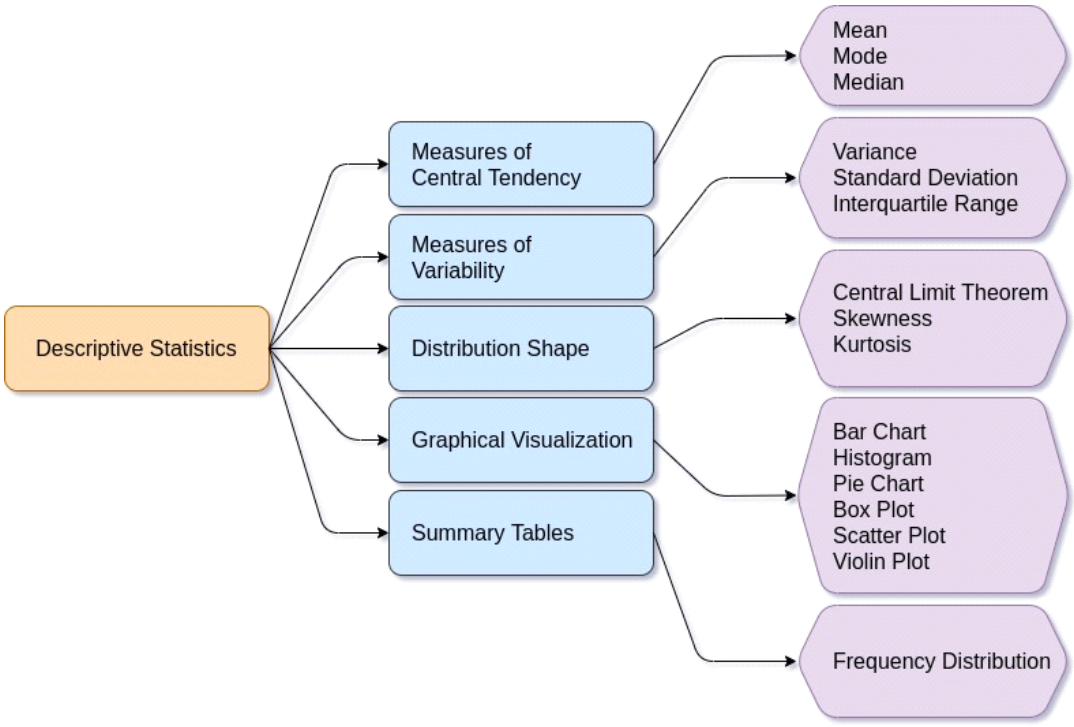
1.4.1 Descriptive Statistics

CONCEPT. The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES. The mean age of citizens who live in a certain geographical area, the

mean length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

INTERPRETATION. You are most likely to be familiar with this branch of statistics because many examples arise in everyday life. Descriptive statistics serves as the basis for analysis and discussion in fields as diverse as securities trading, the social sciences, government, the health sciences, and professional sports. Descriptive methods can seem deceptively easy to apply because they are often easily accessible in calculating and computing devices. However, this easiness does not mean that descriptive methods are without their pitfalls, as Chapters of Presenting Data in Charts and Tables and Descriptive Statistics explain.

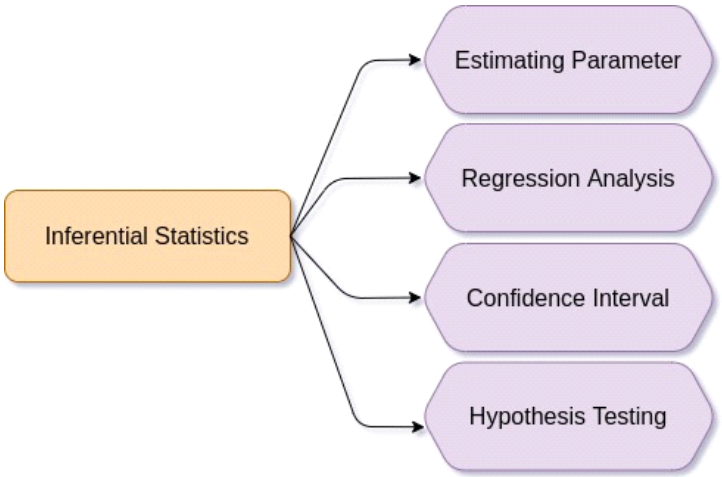


1.4.2 Inferential Statistics

CONCEPT. The branch of statistics that analyzes sample data to reach conclusions about a population.

EXAMPLE. A survey that sampled 1,264 women found that 45% of those polled considered friends or family as their most trusted shopping advisers and only 7% considered advertising as their most trusted shopping adviser. By using methods discussed in next Section, you can use these statistics to draw conclusions about the population of all women.

INTERPRETATION. When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. This deeper level of analysis means that inferential statistical methods can be easily misapplied or misconstrued, and that many inferential methods require a calculating or computing device.



1.4.3 USES OF STATISTICS

Statistics can be used among others for:

1) Planning and decision making by individuals, states, business organizations, research institution etc.

2) Forecasting and prediction for the future based on a good model provided that its basic assumptions are not violated.

3) Project implementation and control. This is especially useful in on-going projects such as network analysis, construction of roads and bridges and implementation of government programs and policies.

4) The assessment of the reliability and validity of measurements and general points significance tests including power and sample size determination.

1.5 STATISTICAL DATA

Data can be described as a mass of unprocessed information obtained from measurement of counting of a characteristics or phenomenon. They are raw facts that have to be processed in numerical form they are called **quantitative data**. For instance the collection of ages of students in a particular session is an example of this data. But when data are not presented in numerical form, they are called **qualitative data**. E.g.: status, sex, religion, etc.

Définition 1.5.1 *Statistical data are data obtained through objective measurement or enumeration of characteristics using the state of the art equipment that is precise and unbiased. Such data when subjected to statistical analysis produce results with high precision.*

1.5.1 SOURCES OF STATISTICAL DATA

1. Primary data: These are data generated by first hand or data obtained directly from respondents by personal interview, questionnaire, measurements or observation. Statistical data can be obtained from:

(i) Census – complete enumeration of all the unit of the population

(ii) Surveys – the study of representative part of a population

(iii) Experimentation – observation from experiment carried out in laboratories and research center.

(iv) Administrative process e.g. Record of births and deaths.

ADVANTAGES

- Comprises of actual data needed
- It is more reliable with clarity
- Comprises a more detail information

DISADVANTAGES

- Cost of data collection is high
- Time consuming
- There may larger range of non response

2. Secondary data: These are data obtained from publication, newspapers, and annual reports. They are usually summarized data used for purpose other than the intended one. These could be obtain from the following:

(i) Publication e.g. extract from publications

(ii) Research/Media organization

(iii) Educational institutions

ADVANTAGES

- The outcome is timely
- The information gathered more quickly
- It is less expensive to gather.

DISADVANTAGES

- Most time information are suppressed when working with secondary data
- The information may not be reliable

1.5.2 METHODS OF COLLECTION OF DATA

There are various methods we can use to collect data. The method used depends on the problem and type of data to be collected. Some of these methods include:

1. Direct observation
2. Interviewing
3. Questionnaire
4. Abstraction from published statistics.

DIRECT OBSERVATION

Observational methods are used mostly in scientific enquiry where data are observed directly from controlled experiment. It is used more in the natural sciences through laboratory works than in social sciences. But this is very useful studying small communities and institutions.

INTERVIEWING

In this method, the person collecting the data is called the interviewer goes to ask the person (interviewed) direct questions. The interviewer has to go to the interviewed personally to collect the information required verbally. This makes it different from the next method called questionnaire method.

QUESTIONNAIRE

A set of questions or statement is assembled to get information on a variable (or a set of variable). The entire package of questions or statement is called a questionnaire. Human beings usually are required to respond to the questions or statements on the questionnaire. Copies of the questionnaire can be administered personally by its user or sent to people by post. Both interviewing and questionnaire methods are used in the social sciences where human population is mostly involved.

ABSTRACTIONS FROM THE PUBLISHED STATISTICS

These are pieces of data (information) found in published materials such as figures related to population or accident figures. This method of collecting data could be useful as preliminary to other methods.
Other methods includes: Telephone method, Document/Report method, Mail or Postal questionnaire, On-line interview method, etc.

1.6 PRESENTATION OF 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. Most cases, it is necessary to present data in tables, charts and diagrams in order to have a clear understanding of the data, and to illustrate the relationship existing between the variables being examined.

1.6.1 FREQUENCY TABLE

This is a tabular arrangement of data into various classes together with their corresponding frequencies.

Procedure for forming frequency distribution

Given a set of observation $x_1, x_2, ..., x_n$, for a single variable.

1. Determine the range $(R) = L - S$ where L = largest observation in the raw data; and S = smallest observation in the raw data.

2. Determine the appropriate number of classes or groups (K). The choice of K is arbitrary but as a general rule, it should be a number (integer) between 5 and 20 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:

(a) $K = 1 + 3.322 \times \log_{10}(n)$

(b) $K = \sqrt{n}$ where n = number of observations.

3. Determine the width (w) of the class interval. It is determined as $w = \frac{R}{K}$

4. Determine the numbers of observations falling into each class interval i.e. find the class frequencies.

NOTE: With advent of computers, all these steps can be accomplishes easily.

SOME BASIC DEFINITIONS

Variable: This is a characteristic of a population which can take different values. Basically, we have two types, namely: continuous variable and discrete variable. A *continuous variable* is a variable which may take all values within a given range. Its values are obtained by measurements e.g. height, volume, time, exam score etc. A *discrete variable* is one whose value change by steps. Its value may be obtained by counting. It normally takes integer values e.g. number of cars, number of chairs.

Class interval: This is a sub-division of the total range of values which a (continuous) variable may take. It is a symbol defining a class E.g. 0-9, 10-19 etc. there are three types of class interval, namely: Exclusive, inclusive and open-end classes method.

Exclusive method:

When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification. E.g. Let some expenditures of some families be as follows: 0 – 1000, 1000 – 2000, etc. It is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the above example, there are so families whose expenditure is between 0 and 999.99. A family whose expenditure is 1000 would be included in the class interval 1000-2000.

Inclusive method:

In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval. This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may take only integral values.

It cannot be used with fractional values like age, height, weight etc. In case of continuous variables, the exclusive method should be used. The inclusive method should be used in case of discrete variable.

Open end classes:

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

Class limit: it represents the end points of a class interval. {Lower class limit & Upper class limit}. A class interval which has neither upper class limit nor lower class limit indicated is called an open class interval e.g. "less than 25", "25 and above"

Cumulative frequency: This is the sum of a frequency of the particular class to the frequencies of the class before it.

Example 1. The following are the marks of 50 students :

48 70 60 47 51 55 59 63 68 63 47 53 72 53 67 62 64 70 57 56 48 51 58 63 65 62 49 64 53 59 63 50 61 67 72 56 64 66 49 52 62 71 58 53 63 69 59 64 73 56.

(a) Construct a frequency table for the above data.

(b) Answer the following questions using the table obtained:

(i) how many students scored between 51 and 62?

(ii) how many students scored above 50?

(iii) what is the probability that a student selected at random from the class will score less than 63?

Solution 1 *(a) Range (R) =73-47=26*

# of classes (k) $= \sqrt{n} = \sqrt{50} = 7.07 \approx 7$

Class size (w) $= \frac{R}{k} = \frac{26}{7} = 3.7 \approx 4$

Frequency table

Mark	***frequency***
47-50	*7*
51-54	*7*
55-58	*7*
59-62	*8*
63-66	*11*
67-70	*6*
71-74	*4*
$\sum$	*50*

(b) (i) 22 (ii) 43 (iii) 0.58

Example 2. The following data represent the ages (in years) of people living in a housing estate :

18 31 30 6 16 17 18 43 2 8 32 33 9 18 33 19 21 13 13 14 14 6 52 45 61 23 26 15 14 15 14 27 36 19 37 11 12 11 20 12 39 20 40 69 63 29 64 27 15 28.

Present the above data in a frequency table showing the following columns; class interval, class mark (mid-point), frequency and cumulative frequency in that order.

Solution 2 *Range (R) =69-2=67*

# of classes (k) $= \sqrt{n} = \sqrt{50} = 7.07 \approx 7$

Class size (w) $= \frac{R}{k} = \frac{67}{7} = 9.5 \approx 10$

class interval	*class mark*	*frequency*	*cumulative frequency*
2-11	*6.5*	*7*	*7*
12-21	*16.5*	*21*	*28*
22-31	*26.5*	*8*	*36*
32-41	*36.5*	*7*	*43*
42-51	*46.5*	*2*	*45*
52-61	*56.5*	*2*	*47*
62-71	*66.5*	*3*	*50*

1.6.2 GRAPHICAL PRESENTATION OF DATA

It is not enough to represent data in a tabular form. The most attractive way of representing data is through charts or graphs.

PIE CHART

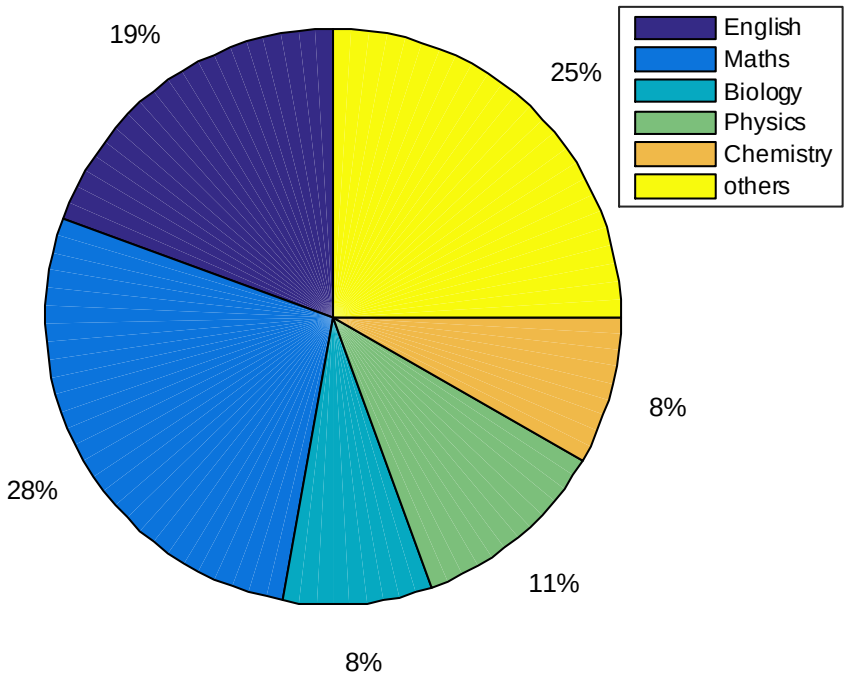
A pie chart is a circular graph in which numerical data are represented by sectors of a circle. The angles of the sectors are proportional to the frequencies of the items they represent

Example. In a school, the lesson periods for each week are given below.

English 7, Maths 10, Biology 3, Physics 4, Chemistry 3, others 9. Draw a pie chart to illustrate this information.

Solution 3 *Total no. of periods in a week = 7+10+3+4+3+9 =36*

Subject	***N° of Periods***	***Angle of sector***
English	*7*	$\frac{7}{36} \times 360° = 70°$
Maths	*10*	$\frac{10}{36} \times 360° = 100°$
Biology	*3*	$\frac{3}{36} \times 360° = 30°$
Physics	*4*	$\frac{4}{36} \times 360° = 40°$
Chemistry	*3*	$\frac{3}{36} \times 360° = 30°$
others	*9*	$\frac{9}{36} \times 360° = 90°$
$\sum$	*36*	$360°$

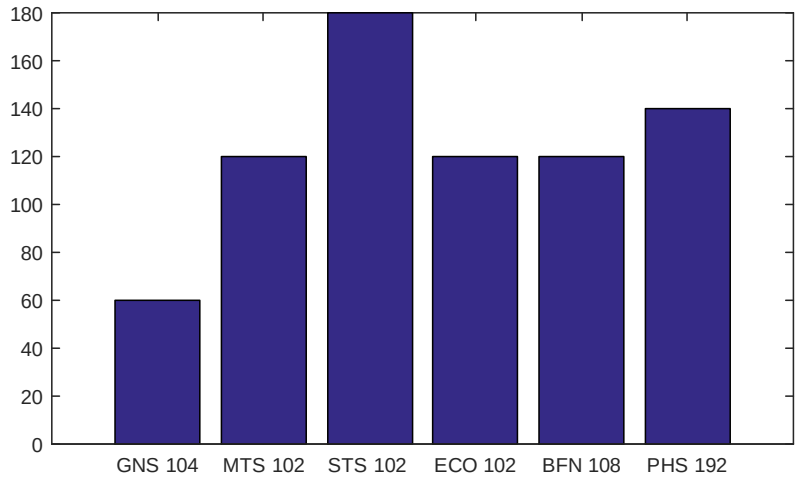


BAR CHART

A bar chart is a statistical graph in which bars (rectangular bars) are drawn such that their lengths or heights are proportional to the quantities or item they represent. Each bar is separated by equal gaps.

Example. The allotment of time in minutes per week for some of the university courses in second semester is :

Courses	**Minutes**
GNS 104	60
MTS 102	120
STS 102	180
ECO 102	120
BFN 108	120
PHS 192	140



HISTOGRAM

Histograms are similar to bar charts; they are a way to display counts of data. A bar graph charts actual counts against categories; The height of the bar indicates the number of items in that category. A histogram displays the same categorical variables in "bins".

A bin shows how many data points are within a range (an interval). Normally, you choose the range that best fits your data. There are no set rules about how many bins you can have, but the rule of thumb is 5-20 bins. Any more than 20 bins and your graph will be hard to read. Fewer than 5 bins and your graph will have little (if any) meaning.

What does the height of a bar in a histogram represent?

Unlike a bar chart, the *area* of a bar in a histogram represents the frequency, not the height. The frequency is calculated by multiplying the width of the bin by the height. The height of a bar in a histogram indicates frequency (counts) only if the bin widths are evenly spaced. For example, if you are plotting magnitudes of earthquakes and your bins are 3-5, 5-7 and 7-9, each bin is spaced two numbers apart and so the height of the bar would equal the frequency. However, histograms don't always have even bins. When a histogram has uneven bins, the height does not equal the frequency.

The table below gives the marks of 80 students on an exam. The data has already been grouped for us into 10 classes. The exam scores are given in whole marks.

Range of marks	**Frequency**
1–10	2
11–20	2
21–30	4
31–40	6
41–50	7
51–60	8
61–70	15
71–80	22
81–90	10
91–100	4
Total	80

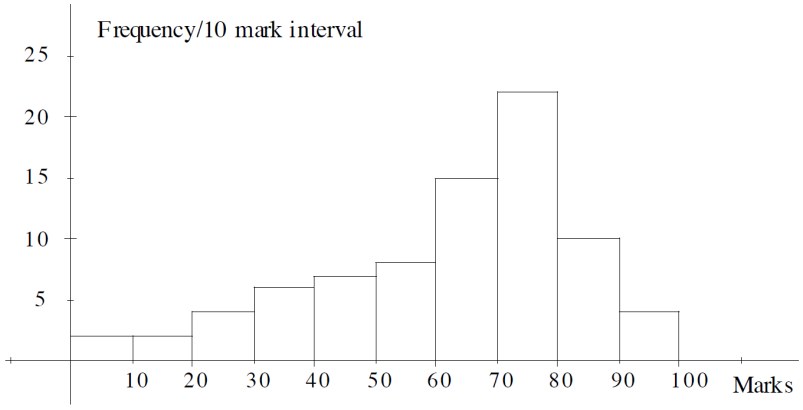
Each of the intervals from 1–10 marks, 11–20 marks and so on is called a class interval. In this example, each class interval is an interval of 10 marks, namely the marks 1 to 10 including both 1 and 10, or 11 to 20 including both 11 and 20, etc. The table tells us that, for example, the class interval 21–30 has a frequency of 4. This means that 4 students scored marks between 21 and 30 inclusive but we don't know their exact marks. A histogram of these data has been drawn below.

Here we have used the right hand endpoint of the class intervals to indicate our horizontal scale. All the class intervals have the same width, 10 marks.

The height of each column represents the frequency per 10 mark interval.

The area of each column represents the number of members in each class interval, or frequency. For the interval 21– 30, Area of the rectangle = no. of 10 mark intervals $\times$ frequency/10 mark interval $= 1 \times 4 = 4$.

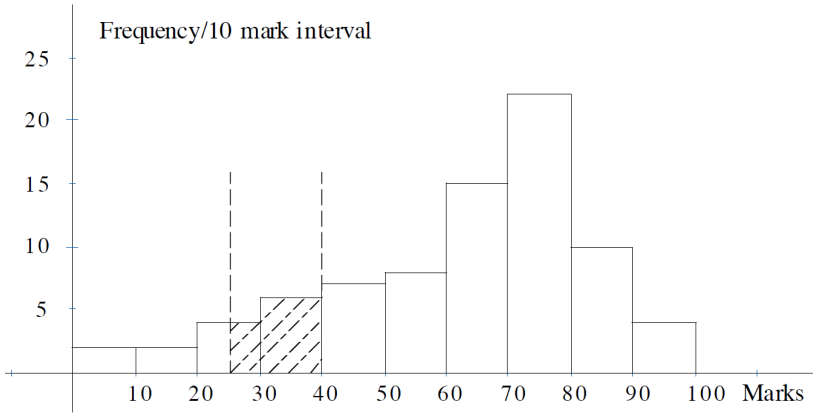
Since each column has the same width, i.e. one, its height is equal to its area. The



total area enclosed represents the total number in the sample.

If we are given a histogram, we can use it to get information about the sample. For example, we can use the histogram in the above figure to estimate the number of people with marks between 26 and 40. We want to find the area of the histogram between the two dotted lines in the next figure. The area is shaded to help you.

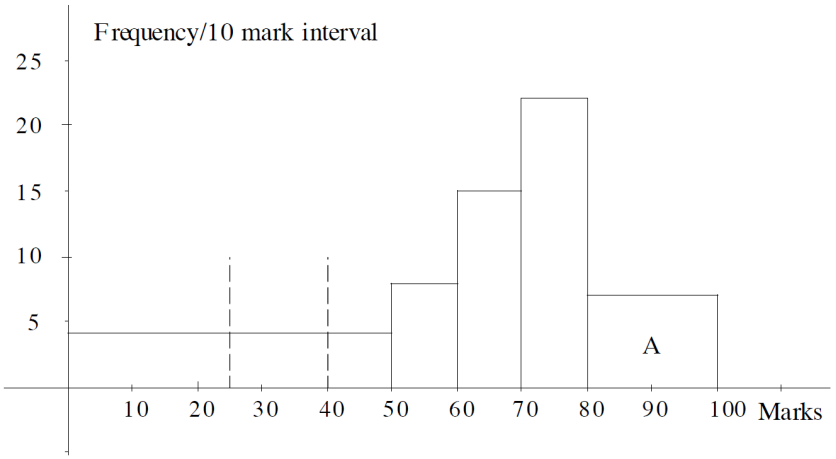
This area is $(\frac{1}{2} \times 4) + (1 \times 6) = 8$. That is, we take half the area of the rectangle 21–30 and add the area of the rectangle 31–40. So we estimate 8 people have marks between 26 and 40.



Now suppose the information is grouped differently.

Range of marks	**Frequency**	**Frequency per 10 marks**
1–50	21	4.2
51–60	8	8
61–70	15	15
71–80	22	22
81–100	14	7
Total	80	

Here all marks of 50 and below are grouped in one class interval, and marks above 80 are also grouped together. In drawing this histogram it is extremely important that the area of each column, rather than its height, represents the frequency. The correct units for the vertical axis is again frequency/10 mark interval. The histogram for these data is drawn below.



For example, the number of people who obtained more than 80 marks is the area of rectangle A.

Area of rectangle A = no. of 10 mark intervals × frequency/10 mark interval = 2×7 = 14.

This histogram can also be used to estimate the number of people with marks between 26 and 40. Again we find the area enclosed by the dotted lines drawn at 25 and 40. This time the estimate is 1.5 × 4.2 = 6.3, so our estimate would be 6.

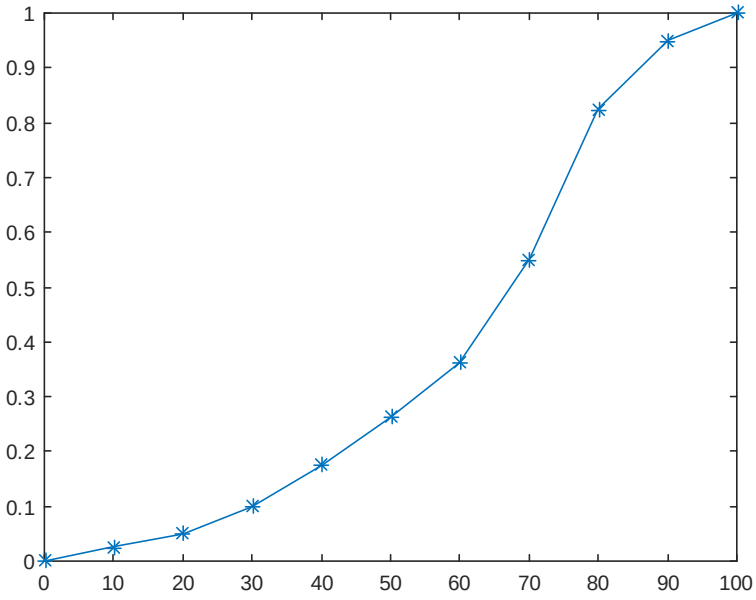
Ogive Graph / Cumulative Frequency Polygon

An ogive (oh-jive), sometimes called a cumulative frequency polygon, is a type of frequency polygon that shows cumulative frequencies. In other words, the cumulative percents are added on the graph from left to right.

An ogive graph plots cumulative (relative) frequency on the y-axis and class boundaries along the x-axis. It's very similar to a histogram, only instead of rectangles, an ogive has a single point marking where the top right of the rectangle would be. It is usually easier to create this kind of graph from a frequency table.

Let's draw an Ogive graph for the set of data given before

Range of marks	**Frequency**	**Relative frequency**	**Cumulative relative frequency**
1–10	2	0.025	0.025
11–20	2	0.025	0.05
21–30	4	0.05	0.1
31–40	6	0.075	0.175
41–50	7	0.0875	0.2625
51–60	8	0.1	0.3625
61–70	15	0.1875	0.55
71–80	22	0.275	0.825
81–90	10	0.125	0.95
91–100	4	0.05	1
Total	80		



CHAPTER 2

Numerical representation of data

When summarizing and describing numerical variables you need to do more than just prepare the tables and charts discussed in the last Chapter. In reading this chapter, you can learn some of the descriptive measures that identify the properties of central tendency, variation, and shape.

2.1 MEASURES OF LOCATION

These are measures of the centre of a distribution. They are single values that give a description of the data. They are also referred to as measure of central tendency. Some of them are arithmetic mean, geometric mean, harmonic mean, mode, and median.

2.1.1 THE ARITHMETIC MEAN (A.M)

The arithmetic mean (average) of set of observation is the sum of the observation divided by the number of observation. Given a set of a numbers $x_1, x_2, ..., x_n$, the arithmetic mean denoted by $\overline{X}$ is defined by

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

Example. The ages of ten students are 16,20,19,21,18,20,17,22,20,17. The mean age is

$$\overline{X} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{16 + 20 + 19 + 21 + 18 + 20 + 17 + 22 + 20 + 17}{10} = 19$$

If the values $x_1, x_2, ..., x_n$ occur $f_1, f_2, ..., f_n$ times respectively, then

$$\overline{X} = \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$$

Example. Find the mean for the table below

Scores (x_i)	2	5	6	8
Frequency (f_i)	1	3	4	2

The average value is

$$\overline{X} = \frac{\sum_{i=1}^{4} f_ix_i}{\sum_{i=1}^{4} f_i} = \frac{1 \times 2 + 3 \times 5 + 4 \times 6 + 2 \times 8}{1 + 3 + 4 + 2} = \frac{57}{10} = 5.7$$

Calculation of mean from grouped data

If the items of a frequency distribution are classified in intervals, we make the assumption that every item in an interval has the mid-values of the interval and we use this midpoint for x_i.

Example. The table below shows the distribution of the waiting time for some customers in a certain petrol station.

Waiting time (in min)	1.5 – 1.9	2.0 – 2.4	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4	$\sum$
No. of customers	3	10	18	10	7	2	50
mid-values	1.7	2.2	2.7	3.2	3.7	4.2	
$f_i \times x_i$	5.1	22	48.6	32	25.9	8.4	142

The average waiting time of the customers is $\overline{X} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} = \frac{142}{50} = 2.84$.

ADVANTAGE OF MEAN

The mean is an average that considers all the observations in the data set. It is single and easy to compute and it is the most widely used average.

DISADVANTAGE OF MEAN

Its value is greatly affected by the extremely too large or too small observation.

2.1.2 THE HARMONIC MEAN (H.M)

The H.M of a set of numbers $x_1, x_2, ..., x_n$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers. It is used when dealing with the rates of the type per (such as kilometers per hour, Dinar per liter). The formula is expressed thus

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

If the values $x_1, x_2, ..., x_n$ occur $f_1, f_2, ..., f_n$ times respectively, then

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

Example. The harmonic mean of 2,4,8,11,4 is $HM = \frac{5}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{11}+\frac{1}{4}} = \frac{440}{107} = 4.112\,1$
The harmonic mean takes into account every value and extreme values have least effect. The formula breaks down when "0" is one of the observations.

2.1.3 THE GEOMETRIC MEAN(G.M)

The geometric mean is useful in finding the average change of percentages, ratios, indexes, or growth rates over time. It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the gross domestic product, which compound or build on each other. The geometric mean of a set of n positive numbers $x_1, x_2, ..., x_n$ is defined as the n^{th} root of the product of n values. The formula for the geometric mean is written

$$GM = \sqrt[n]{x_1 \times x_2 \times ... \times x_n}$$

If f_i is the frequency of x_i, then

$$GM = \sqrt[\Sigma_{i=1}^{n} f_i]{x_1^{f_1} \times x_2^{f_2} \times ... \times x_n^{f_n}}$$

Example. The geometric average of 5, 8, 12, 25 and 34 is

$$GM = \sqrt[5]{5 \times 8 \times 12 \times 25 \times 34} = 13.247$$

Note that GM calculate takes into account every value but It cannot be computed when "0" is on of the observation.

Relation between Arithmetic mean, Geometric and Harmonic

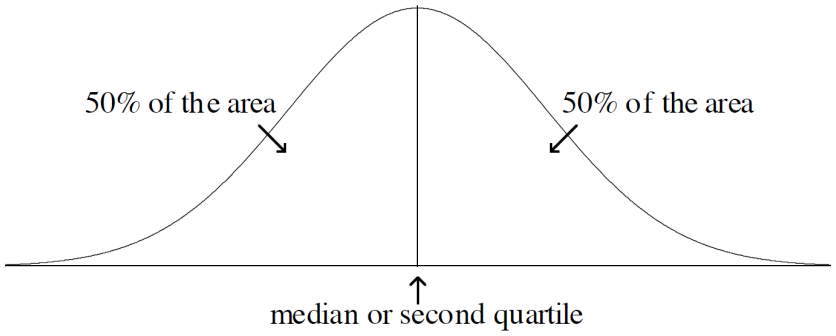
In general, the geometric mean for a set of data is always less than or equal to the corresponding arithmetic mean but greater than or equal to the harmonic mean.
That is,

$$HM \leq GM \leq AM$$

The equality signs hold only if all the observations are identical.

2.1.4 THE MEDIAN

This is the value of the variable that divides a distribution into two equal parts when the values are arranged in order of magnitude.



If there are n (odd) observation, the median $\widetilde{X}$ is the center of observation in the ordered list. The location of the median is $\left(\frac{n+1}{2}\right)^{th}$ item.

But if n is even, the median is the average of the two middle observations in the ordered list. i.e.

$$\widetilde{X} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)^{th}} + x_{\left(\frac{n}{2}+1\right)^{th}}\right)$$

Example. The values of a random variable x are given as 8,5,9,12,10,6 and 4. Find the median ?

In an sorted array: 4,5,6,8,9,10,12. $n = 7$ is odd, therefore, the median is

$$\widetilde{X} = x_{\left(\frac{7+1}{2}\right)^{th}} = x_{(4)^{th}} = 8$$

Example. The value of a random variable are given as 15,15,17,19,21,22,25 and 28. Since $n = 8$ is even, the median is given by

$$\widetilde{X} = \frac{1}{2}\left(x_{\left(\frac{8}{2}\right)^{th}} + x_{\left(\frac{8}{2}+1\right)^{th}}\right) = \frac{1}{2}\left(x_{(4)^{th}} + x_{(5)^{th}}\right) = \frac{1}{2}(19+21) = 20.$$

Calculation of Median from a grouped data

The formula for calculating the median from grouped data is defined as

$$\widetilde{X} = L_1 + \left(\frac{\frac{n}{2} - Cf_b}{f_m}\right) \times w$$

Where

L_1 =Lower class boundary of the median class

Cf_b =Cumulative frequency before the median class

f_m =Frequency of the median class

w =Class size or width

Example. The table below shows the height of 70 men randomly selected.

Height	118-126	127-135	136-144	145-153	154-162	163-171	172-180
f_i	8	10	14	18	9	7	4

Compute the median?

Height	Frequency	Cumulative frequency
118 – 126	8	8
127 – 135	10	18
136 – 144	14	32
145 – 153	18	50
154 – 162	9	59
163 – 171	7	66
172 – 180	4	70

We have $\frac{n}{2} = 35$. The sum of first three classes frequency is 32 which therefore means that the median lies in the fourth class and this is the median class. Then

$$\begin{aligned} L_1 &= 145 \\ Cf_b &= 32 \\ f_m &= 18 \\ w &= 8 \end{aligned}$$

$\widetilde{X} = L_1 + \left(\frac{\frac{n}{2} - Cf_b}{f_m}\right) \times w = 145 + \left(\frac{\frac{70}{2} - 32}{18}\right) \times 8 = 146.33$

The advantage of the median is that Its value is not affected by extreme values; thus it is a resistant measure of central tendency and It is a good measure of location in a skewed distribution, however, It does not take into consideration all the value of the variable.

2.1.5 THE MODE

The mode is the value of the data which occurs most frequently. A set of data may have no, one, two or more modes. A distribution is said to be uni-model, bimodal and multimodal if it has one, two and more than two modes respectively. As example,the mode of scores 2, 5, 2, 6, 7 is 2.

Calculation of mode from grouped data

From a grouped frequency distribution, the mode can be obtained from the formula

$$\widehat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times w$$

Where

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

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

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

w = Class size

Example. For the table below, find the mode.

Class	11 – 20	21 – 30	31 – 40	41 - 50	51 – 60	61 – 70
frequency	6	20	12	10	9	9

The modal class is the second class with $f_2 = 20$.

$$\begin{aligned} L_{mo} &= 21 \\ \Delta_1 &= 14 \\ \Delta_2 &= 8 \\ w &= 9 \end{aligned}$$

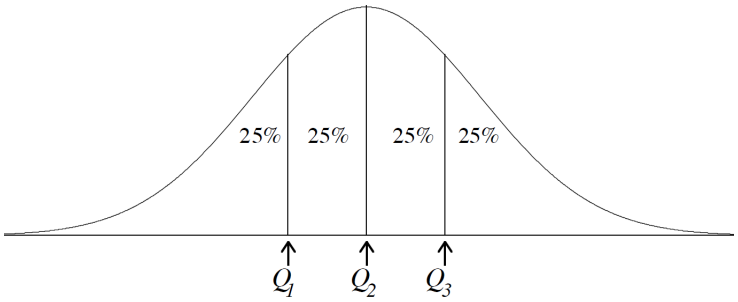
hence, $\widehat{X} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) \times w = 21 + \left(\frac{14}{14+8}\right) \times 9 = 26.727.$

2.2 MEASURES OF PARTITION

From the previous section, we've seen that the median is the value that divides a distribution into two equal parts. Also there are other quantity that divides a set of data (in an array) into different equal parts. Such data must have been arranged in order of magnitude. Some of the partition values are: the quartile, deciles and percentiles.

2.2.1 THE QUARTILES

Quartiles divide a set of data in an array into four equal parts.

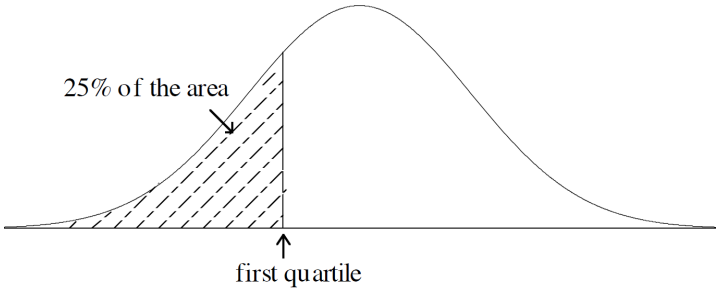


For ungrouped data, the distribution is first arranged in ascending order of magnitude.

Then

$$\begin{aligned}
&\text{First Quartiles :} && Q_1 = \left(\tfrac{n+1}{4}\right)^{th} \\
&\text{Second Quartile :} && Q_2 = 2 \times \left(\tfrac{n+1}{4}\right)^{th} = \text{median} \\
&\text{Third Quartile :} && Q_3 = 3 \times \left(\tfrac{n+1}{4}\right)^{th} \text{ member of the distribution}
\end{aligned}$$

if for $i = 1, 2, 3$; $i \times \left(\frac{n+1}{4}\right) \notin \mathbb{N}$, then $Q_i = \frac{x_{(k)} + x_{(k+1)}}{2}$ where k is the first integer befor $i \times \left(\frac{n+1}{4}\right)$, $x_{(k)}$ and $x_{(k+1)}$ are the $(k)^{th}$ and $(k+1)^{th}$ ordered observations.

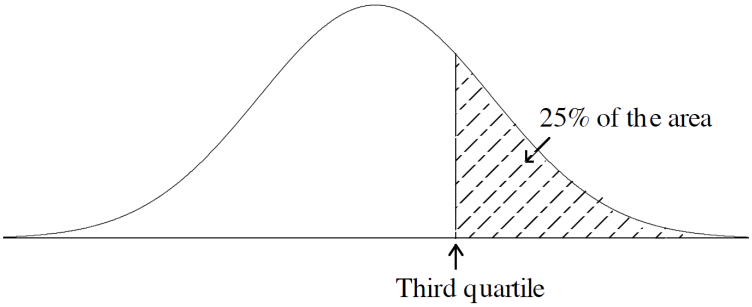


For a grouped data

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

Where $i = 1, 2, 3$ and refers to the quartile number,

L_{qi} :	Lower class boundary of the class counting the quartile
Cf_{bi} :	Cumulative frequency before the Q_i class
f_{qi} :	The frequency of the Q_i class
w :	Class size of the Q_i class.



2.2.2 DECILES

The values of the variable that divide the frequency of the distribution into ten equal parts are known as deciles and are denoted by $D_1, D_2, ..., D_9$. the fifth deciles is the median ($D_5 = Q_2 = \widetilde{X}$).

For ungrouped data, the distribution is first arranged in ascending order of magnitude. Then

$$\begin{aligned} D_1 &: \quad \left(\tfrac{n+1}{10}\right)^{th} \text{ member of the distribution} \\ &\vdots \qquad\qquad\qquad \vdots \\ D_5 &: \quad 5 \times \left(\tfrac{n+1}{10}\right)^{th} \text{ member of the distribution} \\ &\vdots \qquad\qquad\qquad \vdots \\ D_9 &: \quad 9 \times \left(\tfrac{n+1}{10}\right)^{th} \text{ member of the distribution} \end{aligned}$$

if for $i = 1, ..., 9$; $i \times \left(\frac{n+1}{10}\right) \notin \mathbb{N}$, then $D_i = \frac{x_{(k)} + x_{(k+1)}}{2}$ where k is the first integer befor $i \times \left(\frac{n+1}{10}\right)$, $x_{(k)}$ and $x_{(k+1)}$ are the $(k)^{th}$ and $(k+1)^{th}$ ordered observations.

For a grouped data

$$D_i = L_{Di} + \left(\frac{\frac{i \times n}{10} - Cf_{bi}}{f_{Di}}\right) \times w, \qquad i = 1, ..., 9$$

Where

L_{Di} : Lower class boundary of the class counting the decile

Cf_{bi} : Cumulative frequency before the D_i class

f_{Di} : The frequency of the D_i class

w : Class size of the D_i class.

2.2.3 PERCENTILE

The values of the variable that divide the frequency of the distribution into hundred equal parts are known as percentiles and are generally denoted by $P_1, ..., P_{99}$. The fiftieth percentile is the median ($P_{50} = D_5 = Q_2 = \widetilde{X}$).

For ungrouped data, the distribution is first arranged in ascending order of magnitude. Then

$$\begin{aligned} P_1 &: \quad \left(\tfrac{n+1}{100}\right)^{th} \text{ member of the distribution} \\ &\vdots \qquad\qquad\qquad\quad \vdots \\ P_{50} &: \quad 50 \times \left(\tfrac{n+1}{100}\right)^{th} \text{ member of the distribution} \\ &\vdots \qquad\qquad\qquad\quad \vdots \\ P_{99} &: \quad 99 \times \left(\tfrac{n+1}{100}\right)^{th} \text{ member of the distribution} \end{aligned}$$

if for $i = 1, ..., 99$; $i \times \left(\frac{n+1}{100}\right) \notin \mathbb{N}$, then $P_i = \frac{x_{(k)} + x_{(k+1)}}{2}$ where k is the first integer befor $i \times \left(\frac{n+1}{100}\right)$, $x_{(k)}$ and $x_{(k+1)}$ are the $(k)^{th}$ and $(k+1)^{th}$ ordered observations.

For a grouped data

$$P_i = L_{Pi} + \left(\frac{\frac{i \times n}{100} - Cf_{bi}}{f_{Pi}}\right) \times w, \qquad i = 1, ..., 99$$

Where

L_{Pi} : Lower class boundary of the class counting the percentile

Cf_{bi} : Cumulative frequency before the P_i class

f_{Pi} : The frequency of the P_i class

w : Class size of the P_i class.

Example. For the table below, find by calculation (using appropriate expression)

(i) Lower quartile, Q_1

(ii) Upper Quartile, Q_3

(iii) 6th Deciles, D_6

(iv) 45th percentile P_{45}

of the following distribution

Marks	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Frequency	8	10	14	26	20	16	4	2
Cf	8	18	32	58	78	94	98	100
			$\uparrow$	$\uparrow$	$\uparrow$			
			Q_1	P_{45}	Q_3, D_6			

(i) Lower quartile, $Q_1 = L_{q1} + \left(\frac{\frac{100}{4} - Cf_{b1}}{f_{q1}}\right) \times w = 40 + \left(\frac{25-18}{14}\right) \times 10 = 45$

(ii) Upper Quartile, $Q_3 = L_{q3} + \left(\frac{\frac{3\times 100}{4} - Cf_{b3}}{f_{q3}}\right) \times w = 60 + \left(\frac{75-58}{20}\right) \times 10 = 68.5$

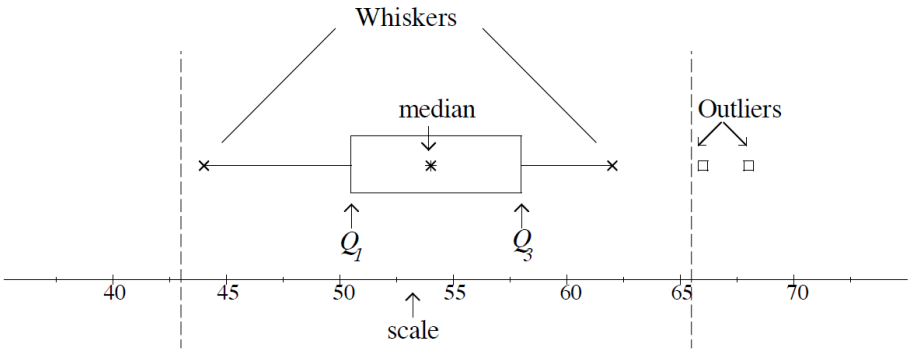
(iii) 6th Deciles, $D_6 = L_{D6} + \left(\frac{\frac{6\times 100}{10} - Cf_{b6}}{f_{D6}}\right) \times w = 60 + \left(\frac{60-58}{20}\right) \times 10 = 61$

(iv) 45th percentile $P_{45} = L_{P45} + \left(\frac{\frac{45\times 100}{100} - Cf_{b45}}{f_{P45}}\right) \times w = 50 + \left(\frac{45-32}{26}\right) \times 10 = 55$

2.2.4 The Box-plot

The box-plot is another way of representing a data set graphically. It is constructed using the quartiles, and gives a good indication of the spread of the data set and its symmetry (or lack of symmetry). It is a very useful method for comparing two or more data sets.

The box-plot consists of a scale, a box drawn between the first and third quartile, the median placed within the box, whiskers on both sides of the box and outliers (if any). The two dashed vertical lines in the figure are the lower and upper outlier thresholds and are not normally included in a box-plot.



The following data set was used to construct a box-plot :

57 46 61 66 48 59 55 56 60 49 44 53 68 57 55 54 49 50 52 54 62 59 51 52 53 54 47 53

Constructing a Box-plot

Step 1: Order the data and calculate the quartiles.

44 46 47 48 49 49 50

51 52 52 53 53 53 54

54 54 55 55 56 57 57

59 59 60 61 62 66 68

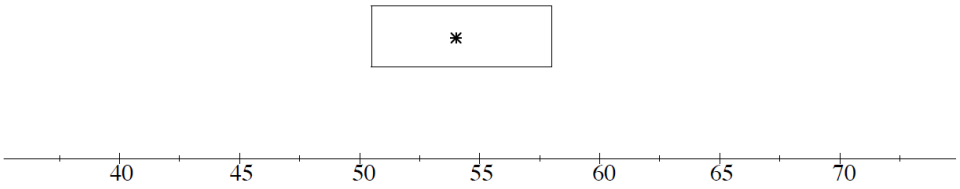
Now we calculate the median, the first quartile and the third quartile.

For these data, median = 54, the first quartile = 50.5 and the third quartile = 58.

With this information we can begin to construct the box-plot.

Step 2: Draw the scale and mark on the quartiles.

Mark the median at the correct place above the scale with a asterix, draw a box around this asterix with the left hand side of the box at the first quartile, 50.5, and the right hand side of the box at the third quartile, 58. This is illustrated in next figure



Step 3: Calculate the interquartile range and determine the position of the outlier thresholds.

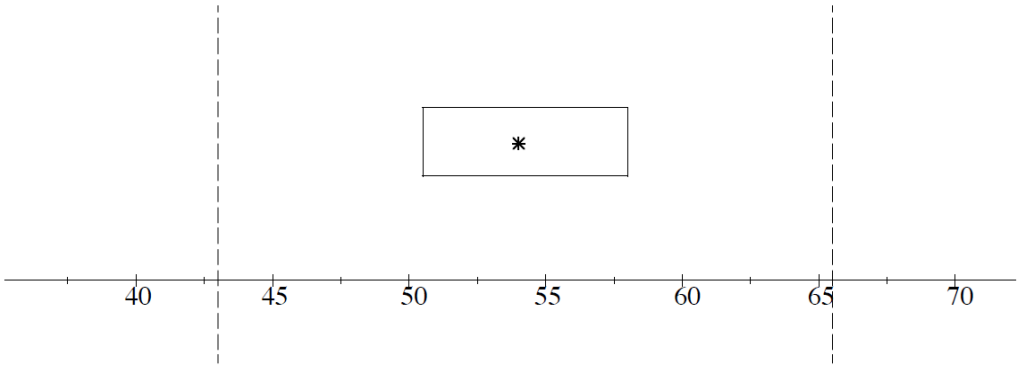
Interquartile range = third quartile - first quartile = 58 - 50.5 = 7.5.

The position of the lower outlier threshold is found by subtracting the interquartile range from the first quartile, 50.5 - 7.5 = 43.

The position of the upper outlier threshold is found by adding the interquartile range to the third quartile, 58 + 7.5 = 65.5.

(Some texts add or subtract 1.5 × interquartile range.)

We now add the outlier thresholds to our diagram. This is illustrated in the figure below.

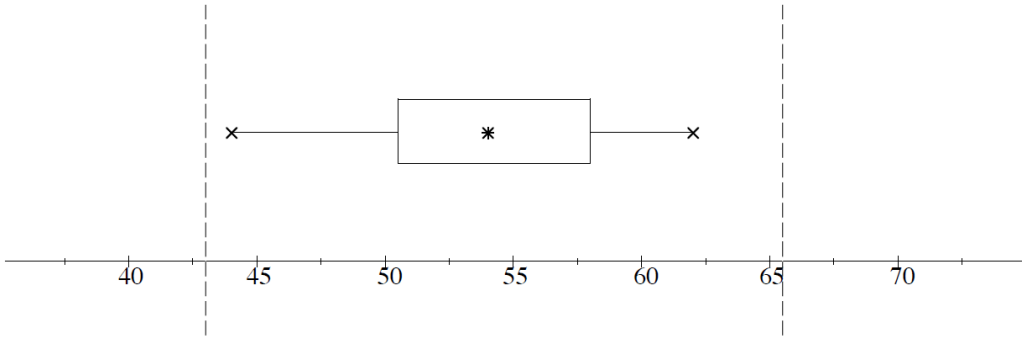


Step 4: Use the outlier thresholds to draw the whiskers.

To draw the left hand whisker, we need the smallest data value that lies inside the outlier thresholds. In this example, it is the value 44. This is drawn on our diagram with a small cross level with the asterix. A horizontal line is now drawn to the left hand side of the box.

To draw the right hand whisker, we find the largest data value that lies inside the outlier thresholds. In this example, the value is 62. This is drawn on the right hand side of the box with a small cross and connected to the box by a horizontal line.

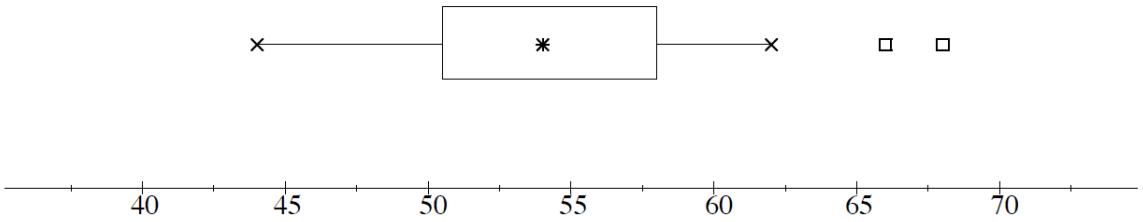
This is illustrated in the next figure.



Step 5: Determine the outliers and remove the outlier thresholds.

Values (if any) that lie outside the outlier thresholds are called outliers. In this example, 66 and 68 are outliers. These are placed on the diagram using a small square or circle. Finally, the outlier thresholds are removed.

The completed box-plot is illustrated in following figure.



2.3 MEASURES OF DISPERSION

Dispersion or variation is degree of scatter or variation of individual value of a variable about the central value such as the median or the mean. These include range, mean deviation, semi-interquartile range, variance, standard deviation and coefficient of variation.

2.3.1 THE RANGE

This is the simplest method of measuring dispersions. It is the difference between the largest and the smallest value in a set of data. It is commonly used in statistical quality control. However, the range may fail to discriminate if the distributions are of different types.

$$Range = L - S$$

2.3.2 SEMI – INTERQUARTILE RANGE

This is the half of the difference between the first (lower) and third quartiles (upper). It is good measure of spread for midrange and the quartiles.

$$S.I.R = \frac{Q_3 - Q_1}{2}$$

2.3.3 THE MEAN/ABSOLUTE DEVIATION

Mean deviation is the mean absolute deviation from the centre. A measure of the center could be the arithmetic mean or median.

Given a set $x_1, ..., x_n$, the mean deviation from the arithmetic mean is defined by:

$$AD = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{X} \right|$$

In a grouped data

$$AD = \frac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} f_i \left| x_i - \overline{X} \right|$$

Example. Below is the average of 6 heads household randomly selected from a country. 47, 45, 56, 60, 41, 54 .Find the (i) Range, (ii) Mean, (iii) Mean deviation from the mean.

The range of the data is $R = 60 - 41 = 19$, and the mean is $\overline{X} = 50.5$. The absolute

deviation from the mean is

$$\begin{aligned} AD &= \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{X}\right| \\ &= \frac{|47-50.5|+|45-50.5|+|56-50.5|+|60-50.5|+|41-50.5|+|54-50.5|}{6} \\ &= 6.166\,7 \end{aligned}$$

Example. The table below shown the frequency distribution of the scores of 42 students. the mean deviation from the mean ($\overline{X} = \frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=1}^{n} f_i x_i = \frac{1429}{42} = 34.024$) for the data is $AD = \frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=1}^{n} f_i\left|x_i - \overline{X}\right| = \frac{465.76}{42} = 11.09$.

Scores	midpoint x_i	f_i	$f_i x_i$	$\left\|x - \overline{X}\right\|$	$f_i\left\|x - \overline{X}\right\|$
0-9	4.5	2	9	29.52	59.04
10-19	14.5	5	72.5	19.52	97.6
20-29	24.5	8	196	9.52	76.16
30-39	34.5	12	414	0.48	5.76
40-49	44.5	9	400.5	10.48	94.32
50-59	54.5	5	272.5	20.48	102.4
60-69	64.5	1	64.5	30.48	30.48
		42	1429		465.76

2.3.4 THE STANDARD DEVIATION AND VARIANCE

The standard deviation, usually denoted by the Greek alphabet σ(small signal is for the population) is defined as the "*positive square root of the arithmetic mean of the squares of the deviation of the given observation from their arithmetic mean*". The variance of a set of observations is defined as "*the square of the standard deviation*" and is thus

given by σ^2.

Given $x_1, ..., x_n$ as a set of observations, then the standard deviation and variance are given by:

STD	**For population**	**For samples**
scattered data	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$
grouped data	$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{N} f_i}\sum_{i=1}^{N} f_i (x_i - \mu)^2}$	$s = \sqrt{\frac{1}{\left(\sum_{i=1}^{n} f_i\right)-1}\sum_{i=1}^{n} f_i \left(x_i - \overline{X}\right)^2}$

VAR	**For population**	**For samples**
scattered data	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$
grouped data	$\sigma^2 = \frac{1}{\sum_{i=1}^{N} f_i}\sum_{i=1}^{N} f_i (x_i - \mu)^2$	$s^2 = \frac{1}{\left(\sum_{i=1}^{n} f_i\right)-1}\sum_{i=1}^{n} f_i \left(x_i - \overline{X}\right)^2$

When we compute the variance, it is important to understand the unit of measure and what happens when the differences in the numerator are squared. When we calculate the variance, the unit of measure for the variance will be the one of the studied variable squared. There is a way out of this difficulty. By taking the square root of the population variance, we can transform it to the same unit of measurement used for the original data and this is the main importance of the standard deviation.

The formula for the population mean is $\mu = \sum x/N$. We just changed the symbols for the sample mean; that is, $\overline{X} = \sum x/n$. Unfortunately, the conversion from the population variance to the sample variance is not as direct. It requires a change in the denominator. Instead of substituting n (number in the sample) for N (number in the population), the denominator is $n - 1$.

Why is this change made in the denominator? Although the use of n is logical since x is used to estimate μ, it tends to underestimate the population variance, σ^2. The use of $(n-1)$ in the denominator provides the appropriate correction for this tendency.

Because the primary use of sample statistics like s^2 is to estimate population parameters like σ^2, $(n-1)$ is preferred to n in defining the sample variance. We will also use this convention when computing the sample standard deviation.

Chebyshev's Theorem

We have stressed that a small standard deviation for a set of values indicates that these values are located close to the mean. Conversely, a large standard deviation reveals that the observations are widely scattered about the mean. The Russian mathematician P. L. Chebyshev (1821–1894) developed a theorem that allows us to determine the minimum proportion of the values that lie within a specified number of standard deviations of the mean. For example, according to Chebyshev's theorem, at least three out of every four, or 75%, of the values must lie between the mean plus two standard deviations and the mean minus two standard deviations. This relationship applies regardless of the shape of the distribution. Further, at least eight of nine values, or 88.9%, will lie between plus three standard deviations and minus three standard deviations of the mean. At least 24 of 25 values, or 96%, will lie between plus and minus five standard deviations of the mean.

CHEBYSHEV'S THEOREM *For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least* $1 - 1/k^2$*, where k is any value greater than 1.*

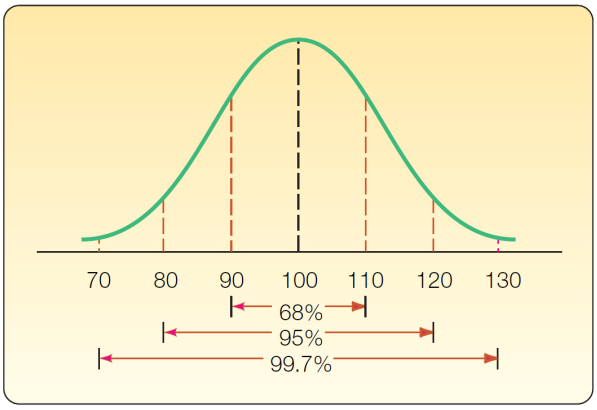
The Empirical Rule

Chebyshev's theorem applies to any set of values; that is, the distribution of values can have any shape. However, for a symmetrical, bell-shaped distribution such as the one in Chart below, we can be more precise in explaining the dispersion about the mean.

These relationships involving the standard deviation and the mean are described by the **Empirical Rule**, sometimes called the **Normal Rule**.

EMPIRICAL RULE *For a symmetrical, bell-shaped frequency distribution, approximately 68% of the observations will lie within plus and minus one standard deviation of the mean; about 95% of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7%) will lie within plus and minus three standard deviations of the mean.*

These relationships are portrayed graphically in the below plot for a bell-shaped distribution with a mean of 100 and a standard deviation of 10.



2.3.5 COEFFICIENT OF VARIATION/DISPERSION

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

$$CV = \frac{\sigma}{\mu} \times 100$$

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

Example. The data below represents the age of 77 applicants in an achievement text for the post of Botanist in a large company. Compute the

(i) Mean

(ii) Standard deviation

(iii) Coefficient of variation.

Ages(in years)	x_i	f_i	$f_i x_i$	$f_i(x-\overline{X})^2$
50-54	52	1	52	402.40
55-59	57	2	114	453.61
60-64	62	10	620	1012.04
65-69	67	12	804	307.24
70-74	72	18	1296	0.07
75-79	76	25	1925	610.09
80-84	82	9	738	889.23
		77	5549	3674.68

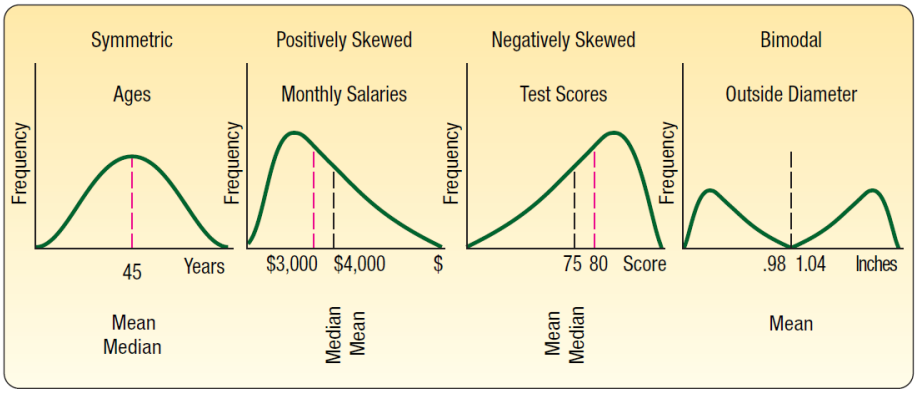
(i) $\overline{X} = \mu = \frac{5549}{77} = 72.065$, (ii) $\sigma = \sqrt{\frac{1}{\sum_{i=1}^{N} f_i} \sum_{i=1}^{N} f_i \left(x_i - \mu\right)^2} = \sqrt{\frac{3674.68}{77}} = 6.908\,2$

(iii) $CV = \frac{\sigma}{\mu} \times 100 = \frac{6.908\,2}{72.065} \times 100\% = 9.586\,1\%$.

2.3.6 SKEWNESS

Another characteristic of a distribution is the shape. There are four shapes commonly observed: symmetric, positively skewed, negatively skewed, and bimodal. In a symmetric distribution the mean and median are equal and the data values are evenly spread around these values. The shape of the distribution below the mean and median is a mirror image of distribution above the mean and median. A distribution of values is

skewed to the right or positively skewed if there is a single peak, but the values extend much farther to the right of the peak than to the left of the peak. In this case, the mean is larger than the median. In a negatively skewed distribution there is a single peak, but the observations extend farther to the left, in the negative direction, than to the right. In a negatively skewed distribution, the mean is smaller than the median. Positively skewed distributions are more common. Salaries often follow this pattern. Think of the salaries of those employed in a small company of about 100 people. The president and a few top executives would have very large salaries relative to the other workers and hence the distribution of salaries would exhibit positive skewness. A bimodal distribution will have two or more peaks. This is often the case when the values are from two or more populations. This information is summarized in the following figure.



There are several formulas in the statistical literature used to calculate skewness. The simplest, developed by Professor Karl Pearson (1857–1936), is based on the difference between the mean and the median.

PEARSON'S COEFFICIENT OF SKEWNESS

$$sk = \frac{3(\overline{X} - \widetilde{X})}{s}$$

Using this relationship, the coefficient of skewness can range from -3 up to 3. A value near -3, such as -2.57, indicates considerable negative skewness. A value such as 1.63 indicates moderate positive skewness. A value of 0, which will occur when the mean and median are equal, indicates the distribution is symmetrical and there is no skewness present.

MEASURE OF SKEWNESS LARGELY USED IN SOFTWARES

$$sk = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x - \overline{X}}{s} \right)^3$$

This formula offers an insight into skewness. The right-hand side of the formula is the difference between each value and the mean, divided by the standard deviation. That is the portion $\frac{x-\overline{X}}{s}$ of the formula. This idea is called **standardizing**. We will discuss the idea of standardizing a value in more detail later when we describe the normal probability distribution. At this point, observe that the result is to report the difference between each value and the mean in units of the standard deviation. If this difference is positive, the particular value is larger than the mean; if the value is negative, the standardized quantity is smaller than the mean. When we cube these values, we retain the information on the direction of the difference. Recall that in the formula for the standard deviation we squared the difference between each value and the mean, so that the result was all nonnegative values.

If the set of data values under consideration is symmetric, when we cube the standardized values and sum over all the values, the result would be near zero. If there are several large values, clearly separate from the others, the sum of the cubed differences

would be a large positive value. If there are several small values clearly separate from the others, the sum of the cubed differences will be negative.

CHAPTER 3

Two ways statistics

In a lot of statistical research, we are not interested in just one character but several at the same time. When we study two characters X and Y on a given population, it is generally because we want to know if there is a link between them and what is the intensity of the link.

Example of possible relationships between the following variables: height and age; diabetes and weight, cholesterol level and diet, ecological niche and population, sunshine and plant growth, toxin and metabolic reaction, survival and pollution, effects and doses...etc. The characters studied can be both qualitative and quantitative.

Correlation is a statistical technique to ascertain the association or relationship between two or more variables. Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables. A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts

a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

Uses of correlations:

1. Correlation analysis helps in deriving precisely the degree and the direction of such relationship.
2. The effect of correlation is to reduce the range of uncertainity of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
3. Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective
4. Economic theory and business studies show relationships between variables like price and quantity demanded, advertising expenditure and sales promotion measures etc.
5. The measure of coefficient of correlation is a relative measure of change.

Types of Correlation:

Correlation is described or classified in several different ways. Three of the most important are:

I. Positive and Negative

II. Simple, Partial and Multiple

III. Linear and non-linear

I. Positive, Negative and Zero Correlation.

Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

Positive Correlation: If both the variables vary in the same direction, correlation is said to be positive. It means if one variable is increasing, the other on an average is also increasing or if one variable is decreasing, the other on an average is also deceasing, then the correlation is said to be positive correlation. For example, the correlation between heights and weights of a group of persons is a positive correlation.

Negative Correlation: If both the variables vary in opposite direction, the correlation is said to be negative. If means if one variable increases, but the other variable decreases or if one variable decreases, but the other variable increases, then the correlation is said to be negative correlation. For example, the correlation between the price of a product and its demand is a negative correlation.

Zero Correlation: Actually it is not a type of correlation but still it is called as zero or no correlation. When we don't find any relationship between the variables then, it is said to be zero correlation. It means a change in value of one variable doesn't influence or change the value of other variable. For example, the correlation between weight of person and intelligence is a zero or no correlation

II. Simple, Partial and Multiple Correlation:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks scored by student and the attendance of student in class, it is a problem of simple correlation.

Partial Correlation: In case of partial correlation one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective

teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant.

Multiple Correlation: When three or more variables are studied, it is a case of multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc, it is a case of multiple correlation.

III. Linear and Non-linear Correlation:

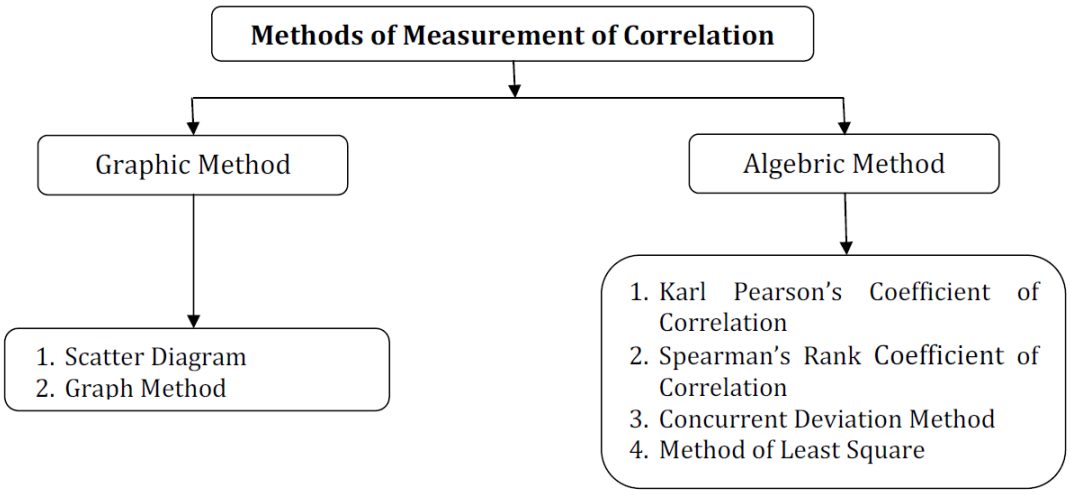
Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line. For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line. For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.

Methods of measurement of correlation:

Quantification of the relationship between variables is very essential to take the benefit of study of correlation. For this, we find there are various methods of measurement of correlation, which can be represented as given below



Among these methods we will discuss only Scatter Diagram, Karl Pearson's Coefficient of Correlation and Spearman's Rank Coefficient of Correlation.

3.1 Bivariate statistical distributions

We consider a population of N individuals measured simultaneously by the two characters X and Y which may be qualitative or quantitative, and which may not be of the same nature. The k modalities of X are denoted by $x_1, ..., x_j, ..., x_k$; the l modalities of Y are denoted by $y_1, ..., y_j, ..., y_l$.

3.1.1 Statistical table

The distribution of the N observations, or joint distribution, according to the modalities of X and Y is presented in the form of a double-entry table, called a **contingency** table or a double-entry table or a **cross table** or sometimes a correlation table (table of k rows and l columns).

X\Y	y_1	y_2	...	y_j	...	y_l	TOTAL
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	$n_{2.}$
.	.	.		.		.	.
.	.	.	....	.	....	.	.
.	.	.		.		.	.
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	$n_{i.}$
.	.	.		.		.	.
.	.	.	....	.	.....	.	.
.	.	.		.		.	.
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	$n_{k.}$
TOTAL	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.l}$	$n_{..} = N$

- The number n_{ij} indicates the number of times the modality x_i of the variable X and the modality y_j of the variable Y were observed simultaneously.

- The number $n_{i.}$, called the marginal number of X, represents the total number of observations of modality x_i of X, whatever the modality of Y.

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

- Similarly, the number $n_{.j}$, called the marginal number of Y, is the total number of observations of the modality y_j of Y, whatever the modality of X.

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

Obviously, we have

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

The joint distribution can also be dened by the frequencies

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

Example. Consider the following two-dimensional statistical series of the pair (X, Y)

X/Y	-2	0	2	3	$n_{i.}$
2	3	4	0	6	13
3	4	3	3	2	12
4	2	3	3	2	10
$n_{.j}$	9	10	6	10	35

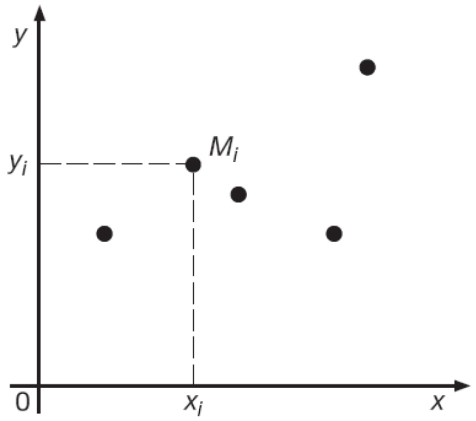
3.1.2 Graphical Representation

This is a very convenient graph for representing the simultaneous observations of two quantitative variables.

Scatter Diagram

If the observations of two statistical variables X and Y are known individually, we start by visualizing them by representing them in the form of a cloud of points: in a Cartesian coordinate system, each observation $(x_i; y_i)$ is represented by the point M_i of coordinates $(x_i; y_i)$, and the shape of the cloud gives information on the type of a possible link.

This is graphic method of measurement of correlation. It is a diagrammatic representation of bivariate data to ascertain the relationship between two variables. Under this



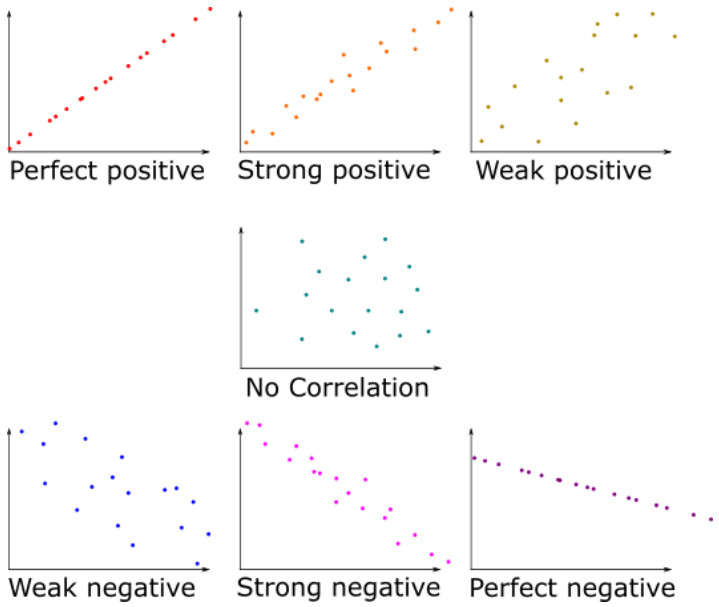
method the given data are plotted on a graph paper in the form of dot. i.e. for each pair of X and Y values we put dots and thus obtain as many points as the number of observations. Usually an independent variable is shown on the X-axis whereas the dependent variable is shown on the Y-axis. Once the values are plotted on the graph it reveals the type of the correlation between variable X and Y. A scatter diagram reveals whether the movements in one series are associated with those in the other series.

- Perfect Positive Correlation: In this case, the points will form on a straight line falling from the lower left hand corner to the upper right hand corner.

- Perfect Negative Correlation: In this case, the points will form on a straight line rising from the upper left hand corner to the lower right hand corner.

- High Degree of Positive Correlation: In this case, the plotted points fall in a narrow band, wherein points show a rising tendency from the lower left hand corner to the upper right hand corner.

- High Degree of Negative Correlation: In this case, the plotted points fall in a narrow band, wherein points show a declining tendency from upper left hand



corner to the lower right hand corner.

- Low Degree of Positive Correlation: If the points are widely scattered over the diagrams, wherein points are rising from the left hand corner to the upper right hand corner.

- Low Degree of Negative Correlation: If the points are widely scattered over the diagrams, wherein points are declining from the upper left hand corner to the lower right hand corner.

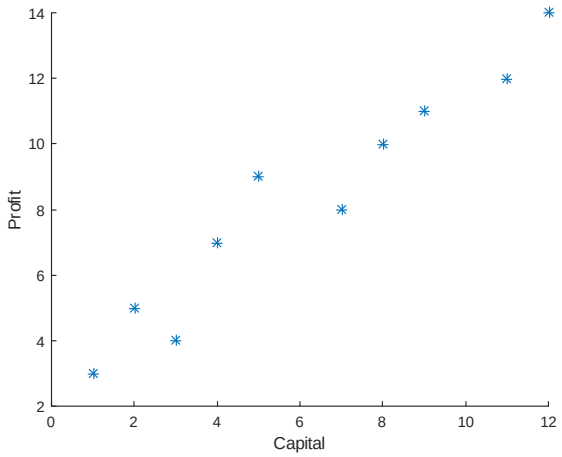
- Zero (No) Correlation: When plotted points are scattered over the graph haphazardly, then it indicate that there is no correlation or zero correlation between two variables.

Example. Given the following pairs of values

Capital Employed (M)	1	2	3	4	5	7	8	9	11	12
Profit (M)	3	5	4	7	9	8	10	11	12	14

(a) Draw a scatter diagram

(b) Do you think that there is any correlation between profits and capital employed? Is it positive or negative? Is it high or low?



From the observation of scatter diagram we can say that the variables are positively correlated. In the diagram the points trend toward upward rising from the lower left hand corner to the upper right hand corner, hence it is positive correlation. Plotted points are in narrow band which indicates that it is a case of high degree of positive correlation.

3.1.3 Marginal distributions

The marginal distribution is determined by isolating the first and last columns of the contingency table. The first column contains the modalities x_i and the last, the corresponding frequencies. That is to say on the margin of the contingency table, we can extract the data only with respect to X and only with respect to Y.

The k pairs $(x_i; n_i)$ form the marginal distribution of the variable X.

The l pairs $(y_j; n_j)$ form the marginal distribution of the variable Y.

Marginal distributions can also be given as frequencies

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

Moreover, we have

$$\sum_{i=1}^{k} f_{i.} = \sum_{j=1}^{l} f_{.j} = 1.$$

These two distributions can be presented in the form of statistical tables

X	Marginal frequency	marginal relative frequency
x_1	$n_{1.}$	$f_{1.} = \frac{n_{1.}}{N}$
x_2	$n_{2.}$	$f_{2.} = \frac{n_{2.}}{N}$
$\vdots$	$\vdots$	$\vdots$
x_k	$n_{k.}$	$f_{k.} = \frac{n_{k.}}{N}$
Total	N	1

Y	Marginal frequency	marginal relative frequency
y_1	$n_{.1}$	$f_{.1} = \frac{n_{.1}}{N}$
y_2	$n_{.2}$	$f_{.2} = \frac{n_{.2}}{N}$
$\vdots$	$\vdots$	$\vdots$
y_l	$n_{.l}$	$f_{.l} = \frac{n_{.l}}{N}$
Total	N	1

3.1.4 Numerical description

Having a joint distribution, we can deduce the marginal distributions which allow us to study each variable separately by graphically representing its distribution and, if it is a quantitative variable, by calculating its central tendency and dispersion characteristics.

Characteristic of marginals

The marginal means of the variables X and Y are:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} n_{i.}x_i = \sum_{i=1}^{k} f_{i.}x_i \qquad \text{and} \qquad \overline{Y} = \frac{1}{N}\sum_{j=1}^{l} n_{.j}y_j = \sum_{j=1}^{l} f_{.j}y_j.$$

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

$$Var(X) = \overline{X^2} - \overline{X}^2 = \sum_{i=1}^{k} f_{i.} x_i^2 - \left(\sum_{i=1}^{k} f_{i.} x_i \right)^2$$

and

$$Var(Y) = \overline{Y^2} - \overline{Y}^2 = \sum_{j=1}^{l} f_{.j} y_j^2 - \left(\sum_{j=1}^{l} f_{.j} y_j \right)^2$$

The marginal standard deviations of X and Y are given by the squared roots of their variances.

Example. Reconsider the following two-dimensional statistical series of the pair (X, Y)

X/Y	-2	0	2	3	$n_{i.}$
2	3	4	0	6	13
3	4	3	3	2	12
4	2	3	3	2	10
$n_{.j}$	9	10	6	10	35

$\overline{X} = \frac{1}{N} \sum_{i=1}^{k} n_{i.} x_i = \frac{1}{35} (2 \times 13 + 3 \times 12 + 4 \times 10) = 2.914\,3.$

and

$\overline{Y} = \frac{1}{N} \sum_{j=1}^{l} n_{.j} y_j = \frac{1}{35} (-2 \times 9 + 0 \times 10 + 2 \times 6 + 3 \times 10) = 0.685\,71.$

The marginal variances of the variables X and Y are

$Var(X) = \overline{X^2} - \overline{X}^2 = \frac{1}{35} (2^2 \times 13 + 3^2 \times 12 + 4^2 \times 10) - (2.914\,3)^2 = 0.649\,71$

and

$Var(Y) = \overline{Y^2} - \overline{Y}^2 = \frac{1}{35} \left((-2)^2 \times 9 + 0^2 \times 10 + 2^2 \times 6 + 3^2 \times 10 \right) - (0.685\,71)^2 = 3.$

$815\,5.$

Conditional distributions

The distribution of the variable Y knowing the variable X being equal to x_i, is called the conditional distribution of Y for $X = x_i$:

$Y \mid X = x_i$	y_1	$\cdots$	y_j	$\cdots$	y_l	Total
Frequency	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i.}$

This distribution of ni observations, satisfying the condition X= xi, is presented in the form of conditional relative frequencies:

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} \qquad \text{and} \qquad \sum_{j=1}^{l} f_{j/i} = 1.$$

There are k conditional distributions of Y for $i = 1, ..., k$.

When the variable Y is quantitative, we can calculate for each value x_i its conditional mean $\overline{Y}_i$ and its conditional variance

$$\overline{Y}_i = \sum_{j=1}^{l} f_{j/i} y_j \qquad \text{and} \qquad Var(Y \mid X = x_i) = \sum_{j=1}^{l} f_{j/i} \left(y_j - \overline{Y}_i\right)^2$$

The k modalities of X inducing a partition of the observations into k subgroups, the mean can be expressed as a weighted sum of the k means $\overline{Y}_i$

$$\overline{Y} = \sum_{i=1}^{k} f_{i.} \overline{Y}_i$$

Symmetrically, we have l conditional distributions of X and we define the conditional relative frequencies $f_{i/j}$

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} \qquad \text{and} \qquad \sum_{i=1}^{k} f_{i/j} = 1.$$

$X \mid Y = y_j$	x_1	$\cdots$	x_i	$\cdots$	x_k	Total
Frequency	$f_{1/j}$	$\cdots$	$f_{i/j}$	$\cdots$	$f_{k/j}$	1

When the variable X is quantitative, we can calculate for each value y_j its conditional mean $\overline{X}_j$ and its conditional variance

$$\overline{X}_j = \sum_{i=1}^{k} f_{i/j} x_i \qquad \text{and} \qquad Var(X \mid Y = y_j) = \sum_{i=1}^{k} f_{i/j} \left(x_i - \overline{X}_j\right)^2$$

We have the following relationship between the mean $\overline{X}$ and the l conditional means $\overline{X}_j$

$$\overline{X} = \sum_{j=1}^{l} f_{.j} \overline{X}_j.$$

Example. Retake the previous example. thus, to determine the conditional mean of X when $Y = 2$, it suffices to observe the behavior of X relative to the column $Y = 2$.

X	n_{i2}
2	0
3	3
4	3
$n_{.2}$	6

$\overline{X}_2 = \frac{1}{6}(0 \times 2 + 3 \times 3 + 4 \times 3) = 3.5$

3.1.5 Covariance between two statistical variables

The covariance is equal to the average of the deviations of the pairs $(x_i; y_i)$ of X and Y with respect to the point $(\overline{X}; \overline{Y})$

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - \overline{X}.\overline{Y}$$

In the case of grouped data in a contingency table (weighted covariance) is given by

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l}\left(x_i - \overline{X}\right)\left(y_j - \overline{Y}\right) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \overline{X}.\overline{Y}$$

The covariance indicates the direction of the relationship between the variables X and Y.

Thus, the following cases can be distinguished:

- If $Cov(X;Y) > 0$, then we can say that the relationship between the two variables is positive.

In this case, these two variables vary in the same direction.

- If $Cov(X;Y) < 0$; then we can say that the relationship between the two variables is negative.

In this case, these two variables vary in opposite directions.

- If $Cov(X;Y) = 0$, then we can say that there is no relationship between the two variables.

In this case, the variations of one do not lead to the variation of the other.

Covariance properties

1. $Cov(X;Y) = Cov(Y;X)$.
2. $Cov(X;X) = Var(X)$.
3. $Var(X+Y) = var(X) + var(Y) + 2cov(X;Y)$.
4. $\forall a;b;x_0;y_0 \in \mathbb{R} : Cov(aX+x_0;bY+y_0) = abCov(X;Y) \Longrightarrow Var(aX+bY+c) = a^2Var(X) + b^2Var(Y) + 2abCov(X;Y)$.
5. $|Cov(X;Y)| \leq \sqrt{Var(X)Var(Y)}$.

The magnitude of the covariance is not very informative since it is affected by the magnitude of both X and Y . However, the sign of the covariance tells us something useful about the relationship between X and Y.

Consider the following conditions:

- $x_i > \overline{X}$ and $y_j > \overline{Y}$ then $(x_i - \overline{X})(y_j - \overline{Y})$ will be positive.
- $x_i < \overline{X}$ and $y_j < \overline{Y}$ then $(x_i - \overline{X})(y_j - \overline{Y})$ will be positive.
- $x_i > \overline{X}$ and $y_j < \overline{Y}$ then $(x_i - \overline{X})(y_j - \overline{Y})$ will be negative.
- $x_i < \overline{X}$ and $y_j > \overline{Y}$ then $(x_i - \overline{X})(y_j - \overline{Y})$ will be negative.

Since $Cov(X, Y)$ depends on the magnitude of X and Y we would prefer to have a measure of association that is not affected by changes in the scales of the variables.

The most common measure of linear association is correlation where the magnitude of the correlation measures the strength of the linear association and the sign determines if it is a positive or negative relationship.

3.1.6 Karl Pearson's Coefficient of linear Correlation

Karl Pearson's method of calculating coefficient of correlation is based on the covariance of the two variables in a series. This method is widely used in practice and the coefficient of correlation is denoted by the symbol "ρ". If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship of correlation.

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

$$-1 \leq \rho(X, Y) \leq 1$$

Remarque 3.1.1 *The link between two numerical variables can be studied by the correlation coefficient. Nevertheless, it should be kept in mind that the Pearson correlation coefficient only measures linear relationships, and its value is in no way a reflection of the existence of a causal link between the two variables. Given random variables X and Y*

$$X \text{ and } Y \text{ are independent } \Longrightarrow Cov(X,Y) = \rho(X,Y) = 0$$

$$Cov(X,Y) = \rho(X,Y) = 0 \not\Rightarrow X \text{ and } Y \text{ are independent}$$

Properties of the linear correlation coefficient:

1. The correlation coefficient is always between -1 and +1.
2. If $\rho = +1$ then the points are all on the same increasing line, the perfect positive linear correlation.
3. If $\rho = -1$ then the points are all on the same decreasing line, the perfect negative linear correlation.
4. If $\rho = 0$ then there is no linear relationship between the variables X and Y.
5. We have for all $a, b, x_0, y_0 \in \mathbb{R}$:

$$\rho(aX + x_0, bY + y_0) = \frac{Cov(aX + x_0, bY + y_0)}{\sigma_{aX+x_0}.\sigma_{bY+y_0}} = \frac{abCov(X,Y)}{|ab|\,\sigma_X.\sigma_Y}$$
$$= \begin{cases} \rho(X,Y) & \text{if } ab > 0 \\ -\rho(X,Y) & \text{if } ab < 0 \end{cases}$$

Example. From following information find the correlation coefficient between advertisement expenses and sales volume using Karl Pearson's coefficient of correlation method.

Firm	1	2	3	4	5	6	7	8	9	10
Advertisement Exp.	11	13	14	16	16	15	15	14	13	13
Sales Volume	50	50	55	60	65	65	65	60	60	50

Let us assume that advertisement expenses are variable X and sales volume are variable Y.

Firm	x_i	y_i	$\left(x_i-\overline{X}\right)^2$	$\left(y_i-\overline{Y}\right)^2$	$\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right)$
1	11	50	9	64	24
2	13	50	1	64	8
3	14	55	0	9	0
4	16	60	4	4	4
5	16	65	4	49	14
6	15	65	1	49	7
7	15	65	1	49	7
8	14	60	0	4	0
9	13	60	1	4	-2
10	13	50	1	64	8
$\sum$	**140**	**580**	**22**	**360**	**70**

$$\overline{X}=\frac{140}{10}=10 \quad \text{and} \quad \overline{Y}=\frac{580}{10}=58$$

$$\rho(X,Y)=\frac{\sum\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right)}{\sqrt{\sum\left(x_i-\overline{X}\right)^2}\times\sqrt{\left(y_i-\overline{Y}\right)^2}}=\frac{\mathbf{70}}{\sqrt{22}\times\sqrt{360}}=0.786\,57.$$

Interpretation: From the above calculation it is very clear that there is high degree of positive correlation i.e.$\rho = 0.7866$, between the two variables. i.e. Increase in advertisement expenses leads to increased sales volume.

Example. Find the correlation coefficient between age and playing habits of the following students using Karl Pearson's coefficient of correlation method.

Age	15	16	17	18	19	20
Number of students	250	200	150	120	100	80
Regular Players	200	150	90	48	30	12

To find the correlation between age and playing habits of the students, we need to compute the percentages of students who are having the playing habit.

Percentage of playing habits = (No. of Regular Players /Total No. of Students)$\times$100

Now, let us assume that ages of the students are variable X and percentages of playing habits are variable Y.

	Age (X)	% of playing habits (Y)	$\left(x_i - \overline{X}\right)^2$	$\left(y_i - \overline{Y}\right)^2$	$\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$
	15	$200/250 \times 100 = 80$	6.25	900	-75
	16	75	2.25	625	-37.5
	17	60	0.25	100	-5
	18	40	0.25	100	-5
	19	30	2.25	400	-30
	20	15	6.25	1225	-87.5
$\sum$	**105**	**300**	**17.5**	**3350**	**-240**

$$\overline{X} = \frac{105}{6} = 17.5 \quad \text{and} \quad \overline{Y} = \frac{300}{6} = 50$$

$$\rho(X,Y) = \frac{\sum \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum \left(x_i - \overline{X}\right)^2} \times \sqrt{\left(y_i - \overline{Y}\right)^2}} = \frac{-240}{\sqrt{17.5} \times \sqrt{3350}} = -0.991\,22$$

Interpretation: From the above calculation it is very clear that there is high degree of negative correlation i.e. $\rho(X,Y) = -0.9912$, between the two variables of age and playing habits. i.e. Playing habits among students decreases when their age increases.

3.1.7 Spearman's Rank Coefficient of Correlation:

When quantification of variables becomes difficult such leadership ability, knowledge of person etc, then this method of rank correlation is useful which was developed by British psychologist Charles Edward Spearman in 1904. In this method ranks are allotted to each element either in ascending or descending order.

The correlation coefficient between these allotted two series of ranks is popularly called as "Spearman's Rank Correlation" and denoted by "r_s". It is defined as the Pearson correlation coefficient between the rank variables. For a sample of size N, the N raw scores x_i, y_i are converted to ranks $R(x_i), R(y_i)$ and r_s is computed as

$$r_s = \rho(R(X), R(Y)) = \frac{Cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

where ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,$Cov(R(X), R(Y))$ is the covariance of the rank variables, $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the rank variables. Only if all N ranks are distinct

integers, it can be computed using the popular formula

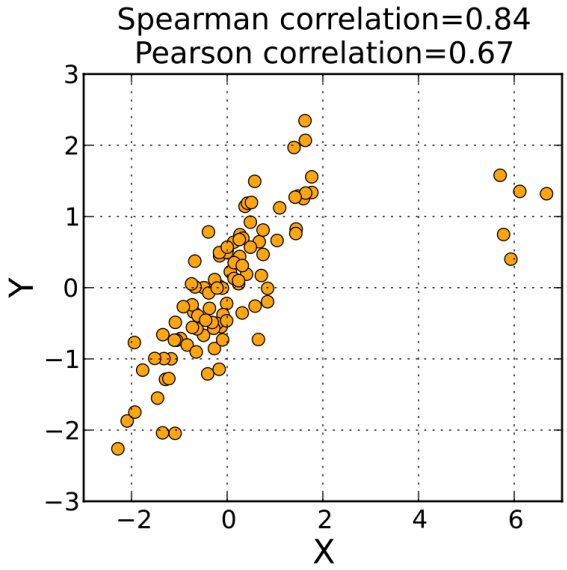
$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

where, $d_i = R(x_i) - R(y_i)$ is the difference between the two ranks of each observation, N = Number of pairs of ranks.

Important Inference to keep in mind: The Spearman correlation can evaluate a monotonic relationship between two variables — Continous or Ordinal and it is based on the ranked values for each variable rather than the raw data.

In case of tie in ranks or equal ranks

In some cases it may be possible that it becomes necessary to assign same rank to two or more elements or individual or entries. In such situation, it is customary to give each individual or entry an average rank. For example, if two individuals are ranked equal to 5th place, then both of them are allotted with common rank $(5 + 6)/2 = 5.5$ and if three are ranked in 5th place, then they are given the rank of $(5 + 6 + 7)/3 = 6$. It means where two or more individuals are to be ranked equal, the rank assigned for the purpose of calculating coefficient of correlation is the average of the ranks with these individual or items or entries would have got had they differed slightly with each other. Where equal ranks are assigned to some entries, an adjustment factor is to be added to the value of $6 \sum d_i^2$ in the above formula for calculating the rank coefficient correlation. This adjustment factor is to be added for every repetition of rank. Adjustment factor $= \frac{m \times (m^2 - 1)}{12}$ where, m = number of items whose rank are common. For example, if a particular rank repeated two times then m=2 and if it repeats three times then $m = 3$ and so on.



Hence the above formula can be re-written as follows:

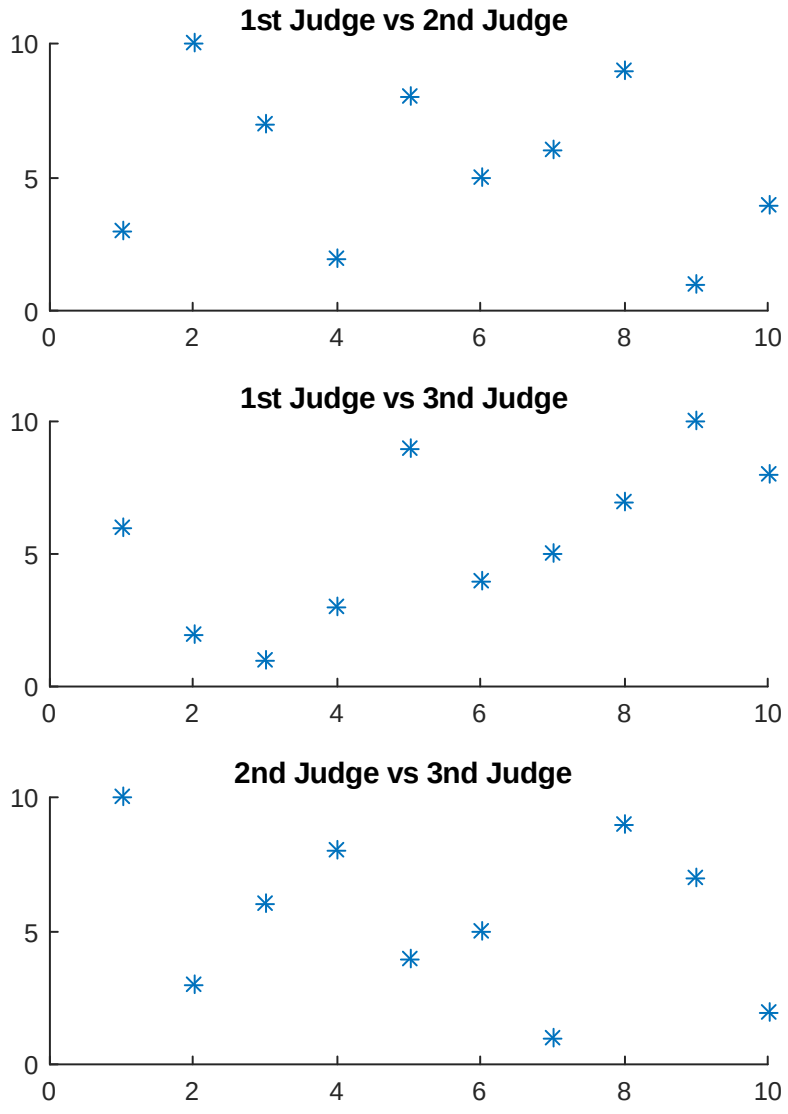
$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{m_1 \times (m_1^2 - 1)}{12} + \frac{m_2 \times (m_2^2 - 1)}{12} + ...\right]}{N(N^2 - 1)}$$

The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's r_s limits the outlier to the value of its rank.

Example. Ten competitors in a skill contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pairs of judges has the nearest approach to common tastes.


1st Judge vs 2nd Judge
1st Judge vs 3nd Judge
2nd Judge vs 3nd Judge
0
2
4
6
8
10
0
5
10

In order to find out which pair of judges has the nearest approach to common tastes, we compare rank correlation between the judgements of

1. 1st Judge and 2nd Judge
2. 2nd Judge and 3rd Judge
3. 1st Judge and 3rd Judge

Note $R1$ =Rank by 1st Judge, $R2$ =Rank by 2nd Judge, $R3$ =Rank by 3rd Judge.

$R1$	$R2$	$R3$	$d^2=(R1-R2)^2$	$d^2=(R2-R3)^2$	$d^2=(R1-R3)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
			$\sum d^2=200$	$\sum d^2=214$	$\sum d^2=60$

1. 1st Judge and 2nd Judge: $r_s=1-\frac{6\sum d_i^2}{N(N^2-1)}=1-\frac{6\times 200}{10\times(10^2-1)}=-0.212\,12$
2. 2nd Judge and 3rd Judge: $r_s=1-\frac{6\sum d_i^2}{N(N^2-1)}=1-\frac{6\times 214}{10\times(10^2-1)}=-0.296\,97$
3. 1st Judge and 3rd Judge: $r_s=1-\frac{6\sum d_i^2}{N(N^2-1)}=1-\frac{6\times 60}{10\times(10^2-1)}=0.636\,36.$

Interpretation: From the above calculation it can be observed that coefficient of correlation is only positive in the judgement of the first and third judges. Therefore, it can be concluded that first and third judges have the nearest approach to common

tastes.

Example. From the following data, compute the rank correlation.

X	82	68	75	61	68	73	85	68
Y	81	71	71	68	62	69	80	70

Calculation of Spearman's Rank Coefficient of Correlation

X	Y	R_1	R_2	$d^2 = (R_1 - R_2)^2$
82	81	2	1	1
68	71	6	3.5	6.25
75	71	3	3.5	0.25
61	68	8	7	1
68	62	6	8	4
73	69	4	6	4
85	80	1	2	1
68	70	6	5	1
				$\sum d^2 = 18.5$

In the problem we find there are repetitions of ranks. Value of $X = 68$ repeated 3 times and value of $Y = 71$ repeated 2 times. Therefore we need to compute adjustment factor to be added to the value of $\sum d^2$. We have

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

where m_1 is for the value X repeated three times, $m_1 = 3$ and for value Y repeated two times, $m_2 = 2$. Thus

$$r_s = 1 - \frac{6 \times \left[18.5 + \frac{3 \times (3^2 - 1)}{12} + \frac{2 \times (2^2 - 1)}{12}\right]}{8 \times (8^2 - 1)} = 0.75.$$

Spearman's Rank Coefficient of Correlation = 0.75, which indicates there is high degree of positive correlation.

3.2 Regression Analysis

Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences which involves identifying and evaluating the relationship between a dependent variable and one or more independent variables, which are also called predictor or explanatory variables. It is particularly useful for assess and adjusting for confounding. Model of the relationship is hypothesized and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables. Linear regression explores relationships that can be readily described by straight lines or their generalization to many dimensions. A surprisingly large number of problems can be solved by linear regression, and even more by means of transformation of the original variables that result in linear relationships among the transformed variables.
When there is a single continuous dependent variable and a single independent variable, the analysis is called a **simple linear regression analysis**. This analysis assumes that there is a linear association between the two variables. **Multiple regression** is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.
I*ndependent variables* are characteristics that can be measured directly; these variables are also called predictor or explanatory variables used to predict or to explain the

behavior of the dependent variable.

Dependent variable is a characteristic whose value depends on the values of independent variables.

Objectives of Regression Analysis

Regression analysis used to explain variability in dependent variable by means of one or more of independent or control variables and to analyze relationships among variables to answer; the question of how much dependent variable changes with changes in each of the independent's variables, and to forecast or predict the value of dependent variable based on the values of the independent's variables.

The primary objective of regression is to develop a linear relationship between a response variable and explanatory variables for the purposes of prediction, assumes that a functional linear relationship exists, and alternative approaches (functional regression) are superior.

3.2.1 Assumption of Regression Analysis

The regression model is based on the following assumptions.

- The relationship between independent variable and dependent is linear.
- The expected value of the error term is zero
- The variance of the error term is constant for all the values of the independent variable, the assumption of homoscedasticity.
- There is no autocorrelation.
- The independent variable is uncorrelated with the error term.

- The error term is normally distributed.
- On an average difference between the observed value and the predicted value is zero.
- On an average the estimated values of errors and values of independent variables are not related to each other.
- The squared differences between the observed value and the predicted value are similar.
- There is some variation in independent variable. If there are more than one variable in the equation, then two variables should not be perfectly correlated.

3.2.2 Simple Regression Model

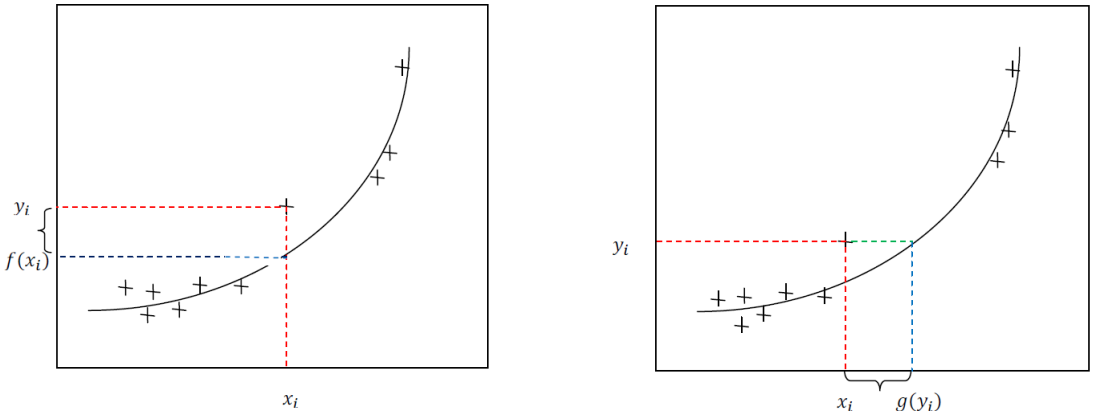
Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable y, based on the value of an independent variable x.

- One variable, denoted (x), is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted (y), is regarded as the response, outcome, or dependent variable.

We choose a mathematical function whose graphical representation approaches the shape of the point cloud as closely as possible. The problem is to determine the coefficients of the function which minimize the sum of the squares of the deviations between

the points representative of the observations and the curve representative of the function.

In the least squares method, the deviations are measured parallel to the axes, that is to say vertically or horizontally in an orthogonal frame.



Differences measured vertically (Y in X) Differences measured horizontally (X in Y)

The fitted curve of Y on X

Let (x_i, y_i),$i = 1, \ldots, n$,n observations of the pair (X, Y). We are looking for a function $f : X \longrightarrow Y = f(X)$. Let $f(x_i)$ be the value taken by the function f when $X = x_i$. The vertical difference between the point (x_i, y_i) of the series and the curve is noted $\varepsilon_i = y_i - f(x_i)$. The gap ε_i is called residual gap or residual. The sum of the squared deviations calculated for each point of the cloud is equal to $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$.

We call The fitted curve of Y on X and we note $C_{Y/X}$, the representative curve of the function such that the sum $\sum_{i=1}^{n} (y_i - f(x_i))^2$ be minimal.

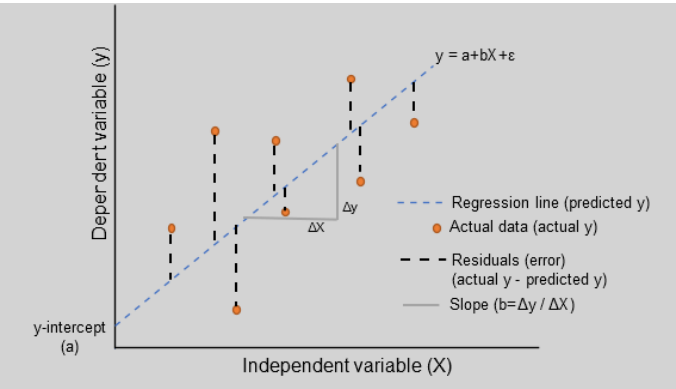
The fitted curve of X on Y

We are looking for a function $g : Y \longrightarrow X = g(Y)$. Let $g(y_i)$ the value taken by the function g when $Y = y_i$. The horizontal difference between the point (x_i, y_i) of the series and the curve is written $\varepsilon_i^{'} = x_i - g(y_i)$. The sum that must be made minimal is in this case is $\sum_{i=1}^{n} \varepsilon_i^{'2} = \sum_{i=1}^{n} (x_i - g(y_i))^2$.

Remarque 3.2.1 *The least squares method provides two curves of fit for a given statistical series. The two curves are all the closer to each other as the dispersion of the points is low.*

Linear fitting

When the shape of the point cloud leads to retaining the adjustment of a straight line, we say that we are making a linear fitting.



The fitted line of Y on X

The fitted line has an equation of the form $y = ax+b$. We put $\widehat{y_i} = ax_i+b$ the estimated value of y_i by the linear model. The difference between the point (x_i, y_i) of the cloud and the line is then $\varepsilon_i = y_i - ax_i - b$.

We call the fitted line of Y on X or the regression line of Y on X , and we note it $D_{Y/X}$, the line of equation $y = ax + b$ such that the sum $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ is minimal.
To find the optimal line in the sense of least squares, it is necessary to find the coefficients a (slope) and b (intercept).

Note

$$\phi(a, b) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

To find the optimal line, we must solve the system of partial derivatives of the first order

$$\begin{cases} \frac{\partial \phi(a,b)}{\partial a} = 0 \\ \frac{\partial \phi(a,b)}{\partial b} = 0 \end{cases}$$

which gives

$$\begin{cases} \sum_{i=1}^{n} -2x_i (y_i - ax_i - b) = 0 \\ \sum_{i=1}^{n} -2 (y_i - ax_i - b) = 0 \end{cases} \Longleftrightarrow \begin{cases} \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - nb = 0 \end{cases}$$

The solution to this system of equations is given by

$$a = \frac{Cov(X, Y)}{Var(X)}$$
$$b = \overline{Y} - a\overline{X}$$

The fitted line of X on Y

The fitted line has an equation of the form $x = a'y + b'$. We put $\widehat{x}_i = a'y_i + b'$ the estimated value of x_i by the linear model. The difference between the point (x_i, y_i) of the cloud and the line is then $\varepsilon_i^{'} = x_i - a'y_i - b'$.

We call the fitted line of X on Y or the regression line of X on Y , and we note it

$D_{X/Y}$, the line of equation $x = a'y + b'$ such that the sum $\sum_{i=1}^{n} (x_i - \widehat{x}_i)^2$ is minimal. To find the optimal line in the sense of least squares, it is necessary to find the coefficients a (slope) and b (intercept).

Note

$$\psi(a', b') = \sum_{i=1}^{n} (x_i - \widehat{x}_i)^2 = \sum_{i=1}^{n} (x_i - a'y_i - b')^2$$

The values of a' and b' are given by

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

$$b = \overline{X} - a'\overline{Y}$$

Remarque 3.2.2

- *The two regression lines $D_{Y/X}$ and $D_{X/Y}$ are generally distinct. They intersect at the center of gravity $G(\overline{X}, \overline{Y})$ of the cloud. The two slopes a and a' always have the same sign, that of the covariance.*
- *To draw the two regression lines on the same graph, we write $D_{X/Y}$ as follows :*

$$D_{X/Y} : y = \frac{x}{a'} - \frac{b'}{a'}$$

- *When the two variables and are independent, the two regression lines $D_{Y/X}$ and $D_{X/Y}$ are perpendicular and therefore $aa' = 0$.*
- *When the variables X and Y are functionally related, the two regression lines coincide and therefore $aa' = a\frac{1}{a} = 1$*
- *In general , $0 \leq aa' \leq 1$.*

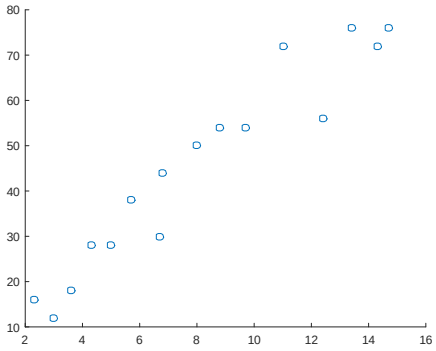
Example. (Ungrouped Data) Research is carried out on the speed of propagation of

nerve impulses in a nerve fiber. We denote by X the diameter in microns of nerve fibers and Y the speed in meters per second of the nerve impulse in the fiber of diameter X. The results are given in the table below

X	2.3	3	3.6	4.3	5	5.7	6.7	6.8	8	8.8	9.7	11	12.4	13.4	14.3	14.7
Y	16	12	18	28	28	38	30	44	50	54	54	72	56	76	72	76

We give $\sum_{i=1}^{n} x_i = 129.7$, $\sum_{i=1}^{n} y_i = 724$, $\sum_{i=1}^{n} x_i^2 = 1304.79$, $\sum_{i=1}^{n} y_i^2 = 39960$, $\sum_{i=1}^{n} x_i y_i = 7165.4$.

1. Represent the scatterplot associated with this double statistical series.



2. Calculate the means of each of the variables X and Y.

$$\overline{X} = \frac{1}{16}\sum_{i=1}^{n} x_i = \frac{129.7}{16} = 8.1063$$

$$\overline{Y} = \frac{1}{16}\sum_{i=1}^{n} y_i = \frac{724}{16} = 45.25$$

3. Calculate the variances of each of the variables.

$$\sigma_X^2 = \frac{1}{16}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1304.79}{16} - (8.106\,3)^2 = 15.837$$
$$\sigma_Y^2 = \frac{1}{16}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2 = \frac{39960}{16} - (45.25)^2 = 449.94$$

4. Calculate the covariance between X and Y

$$Cov(X,Y) = \frac{1}{16}\sum_{i=1}^{n} x_i y_i - \overline{X}.\overline{Y} = \frac{7165.4}{16} - 8.106\,3 \times 45.25 = 81.027$$

5. Find the equations of the regression lines of Y on X and of X on Y.

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{81.027}{15.837} = 5.116\,3$$
$$b = \overline{Y} - a\overline{X} = 45.25 - 5.116\,3 \times 8.106\,3 = 3.775\,7$$

which gives

$$D_{Y/X} : y = 5.116\,3x + 3.775\,7.$$

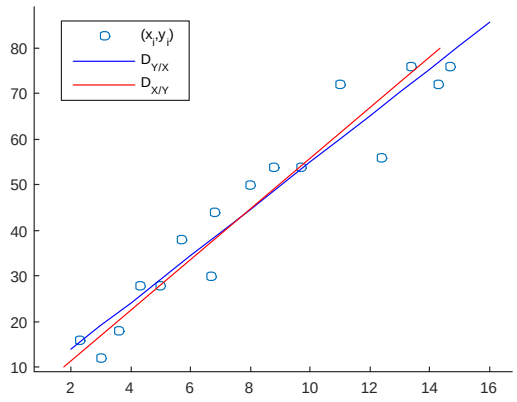
Also, for $D_{X/Y}$ we have

$$a' = \frac{Cov(X,Y)}{Var(Y)} = \frac{81.027}{449.94} = 0.180\,08$$
$$b = \overline{X} - a'\overline{Y} = 8.106\,3 - 0.180\,08 \times 45.25 = -0.042\,32$$

and

$$D_{X/Y} : x = 0.180\,08y - 0.042\,32.$$

6. Draw the two regression lines



If we look for the diameter of a nerve fiber through which the nerve impulse would propagate at the speed of $100m/s$ then we use the regression lines $D_{X/Y}$ to predict its value

$$x = 0.180\,08 \times 100 - 0.042\,32 = 17.966 \text{ microns}$$

and to determine the speed of the nerve impulse through a nerve fiber with a diameter of 18 microns, we use the regression lines $D_{Y/X}$

$$y = 5.116\,3 \times 18 + 3.775\,7 = 95.869m/s$$

Example. (Grouped Data) The table below represents the distribution of 28 students according to the annual number of absences (X) and the final grade (Y).

$X\backslash Y$	[0,5[	[5,10[	[10,15[	[15,20[
0	2	3	3	0
1	0	1	2	3
2	0	0	1	1
3	4	3	0	0
4	1	0	4	0

Give the equation of the regression lines of Y on X and that of X on Y.

$x_i \backslash y_i$	2.5	7.5	12.5	17.5	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
0	2| $n_{11}x_1y_1 = 0$	3| 0	3| 0	0| 0	**8**	0	0
1	0| 0	1| 7.5	2| 25	3| 52.5	**6**	6	6
2	0| 0	0| 0	1| 25	1| 35	**2**	4	8
3	4| 30	3| 67.5	0| 0	0| 0	**7**	21	63
4	1| 10	0| 0	4| 200	0| 0	**5**	20	80
$n_{.j}$	**7**	**7**	**10**	**4**	**28**	$\sum = 51$	$\sum = 157$
$n_{.j}y_j$	17.5	52.5	125	70	$\sum = 265$		
$n_{.j}y_j^2$	43.75	393.75	1562.5	1225	$\sum = 3225$		

$$\overline{X} = \frac{1}{28}\sum n_{i.}x_i = 1.8214 \qquad \overline{Y} = \frac{1}{28}\sum n_{.j}y_j = 9.4642$$

$$\sigma_X^2 = \frac{1}{28}\sum n_{i.}x_i^2 - \overline{X}^2 = 2.289541 \quad \sigma_Y^2 = \frac{1}{28}\sum n_{.j}y_j^2 - \overline{Y}^2 = 25.60587$$

$$Cov(X,Y) = \frac{1}{28}\sum\sum n_{ij}x_iy_j - \overline{X}.\overline{Y} = -1.077806.$$

We obtain $a = \frac{Cov(X,Y)}{\sigma_X^2} = -0.470752$ and $b = \overline{Y} - a\overline{X} = 10.321731$, thus

$$D_{Y/X} : y = -0.470752x + 10.321731.$$

Also $a' = \frac{Cov(X,Y)}{\sigma_Y^2} = -0.04209215$ and $b' = \overline{X} - a'\overline{Y} = 2.219801$,

$$D_{X/Y} : x = -0.04209215y + 2.219801.$$

Residual variance and explained variance by a regression line

Fitted line of Y on X.

The quality of the adjustment is even better than the sum of the squares of the residuals (or deviations) between the line and the points of the series is small.

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

$$V_R(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

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

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

Properties.

- $Var(Y) = V_R(Y) + V_E(Y)$.
- $V_E(Y) = \frac{Cov(X,Y)^2}{Var(X)} = Var(Y)\rho(X,Y)^2$.
- $V_R(Y) = Var(Y)(1 - \rho(X,Y)^2)$.

Coefficient of determination

We call the coefficient of determination the quantity R^2 defined by

$$R^2 = \frac{V_E(Y)}{Var(Y)} = \frac{V_E(X)}{Var(X)}$$

The coefficient of determination R^2 summarizes the part of the variance explained by the regression lines. It is a unitless number, between 0 and 1. It can be expressed as a percentage (percentage that the explained variance represents in relation to the total variance).

From $Var(Y) = V_R(Y) + V_E(Y)$, we have

$$R^2 = 1 - \frac{V_R(Y)}{Var(Y)} = 1 - \frac{V_R(X)}{Var(X)}.$$

R^2 is close to 1 if the points of the cloud are little dispersed around the regression line, it is close to 0 otherwise.

Example. from the previous example of nerve impulse propagation speed in a nerve fiber and diameter of a nerve fiber, we compute the coefficient of correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{81.02969}{\sqrt{15.83809}\sqrt{449.9375}} = 0.9598794.$$

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

$$V_R(Y) = Var(Y)(1 - \rho(X,Y)^2) = 35.37928,$$
$$V_E(Y) = Var(Y)\rho(X,Y)^2 = 414.5582.$$

then we calculate the coefficient of determination

$$R^2 = \frac{V_E(Y)}{Var(Y)} = \frac{414.5582}{449.9375} = 0.921\,37$$

The regression lines explain 92.13% of the variance (of X and Y) which gives that the goodness of fit is very good.

CHAPTER 4

Probability

Often in life we are confronted by our own ignorance. Whether we are pondering tonight's traffic jam, tomorrow's weather, next week's stock prices, an upcoming election, or where we left our hat, often we do not know an outcome with certainty. Instead, we are forced to guess, to estimate, to hedge our bets.

Probability is the science of uncertainty. It provides precise mathematical rules for understanding and analyzing our own ignorance. It does not tell us tomorrow's weather or next week's stock prices; rather, it gives us a framework for working with our limited knowledge and for making sensible decisions based on what we do and do not know.

To say there is a 40% chance of rain tomorrow is not to know tomorrow's weather. Rather, it is to know what we do not know about tomorrow's weather.

In this chapter, we will develop a more precise understanding of what it means to say there is a 40% chance of rain tomorrow. We will learn how to work with ideas of randomness, probability, expected value, prediction, estimation, etc., in ways that are

sensible and mathematically clear.

At the beginning of the twentieth century, Russians such as Andrei Andreyevich Markov, Andrey Nikolayevich Kolmogorov, and Pafnuty L. Chebyshev (and American Norbert Wiener) developed a more formal mathematical theory of probability. In the 1950s, Americans William Feller and Joe Doob wrote important books about the mathematics of probability theory. They popularized the subject in the western world, both as an important area of pure mathematics and as having important applications in physics, chemistry, and later in computer science, economics, and finance. Probability theory also plays a key role in many important applications of science and technology. For example, the design of a nuclear reactor must be such that the escape of radioactivity into the environment is an extremely rare event. Of course, we would like to say that it is categorically impossible for this to ever happen, but reactors are complicated systems, built up from many interconnected subsystems, each of which we know will fail to function properly at some time. Furthermore, we can never definitely say that a natural event like an earthquake cannot occur that would damage the reactor sufficiently to allow an emission. The best we can do is try to quantify our uncertainty concerning the failures of reactor components or the occurrence of natural events that would lead to such an event. This is where probability enters the picture. Using probability as a tool to deal with the uncertainties, the reactor can be designed to ensure that an unacceptable emission has an extremely small probability — say, once in a billion years—of occurring.

In this Chapter we introduce the basic concepts underlying probability theory. We begin with the sample space, which is the set of possible outcomes.

4.1 Sample Spaces and Events

The **sample space** Ω is the set of possible outcomes of an experiment. Points ω in Ω are called **sample outcomes**, **realizations**, or **elements**. Subsets of Ω are called **Events**.

Example. If we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$.

Example. Let ω be the outcome of a measurement of some physical quantity, for example, temperature. Then $\Omega = \mathbb{R}$. One could argue that taking $\Omega = \mathbb{R}$ is not accurate since temperature has a lower bound. But there is usually no harm in taking the sample space to be larger than needed. The event that the measurement is larger than 10 but less than or equal to 23 is $A =]10, 23]$.

Example. If we toss a coin forever, then the sample space is the infinite set

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, ...) : \omega_i \in \{H, T\}\}.$$

Let E be the event that the first head appears on the third toss. Then

$$E = \{(\omega_1, \omega_2, \omega_3, \omega_4, ..., \omega_i, ...) : \omega_1 = \omega_2 = T, \omega_3 = H, \ \forall i \geq 4, \omega_i \in \{H, T\}\}$$

Given an event A, let $A^c = \{\omega \in \Omega : \omega \notin A\}$ denote the **complement of A**. Informally, A^c can be read as "**not A**." The complement of Ω is the empty set $\emptyset$. The union of events A and B is defined

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{ both}\}$$

which can be thought of as "A or B." If $A_1, A_2, ...$ is a sequence of sets then

$$\cup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one } i\}.$$

The intersection of A and B is

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$

read "A and B." Sometimes we write $A \cap B$ as AB or (A, B). If $A_1, A_2, ...$ is a sequence of sets then

$$\cap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}.$$

The set difference is defined by

$$A - B = \{\omega : \omega \in A, \omega \notin B\}.$$

. If every element of A is also contained in B we write $A \subset B$ or, equivalently,$B \supset A$. If A is a finite set, let $|A|$ denote the number of elements in A. See the following table for a summary.

Summary of Terminology

Ω	sample space
ω	outcome (point or element)
A	event (subset of Ω)
A^c	complement of A (not A)
$A \bigcup B$	union (A or B)
$A \bigcap B$ or AB	intersection (A and B)
$A - B$	set difference (ω in A but not in B)
$A \subset B$	set inclusion
$\emptyset$	null event (always false)
Ω	true event (always true)

The complement of $A \cup B$, namely, the set of elements that are in neither A nor B. So

we immediately have

$$(A \cup B)^c = A^c \cap B^c.$$

Similarly, we can show that

$$(A \cap B)^c = A^c \cup B^c,$$

namely, the subset of elements that are not in both A and $B$is given by the set of elements not in A or not in B.

We say that $A_1, A_2, ...$ are **disjoint or aremutually exclusive** if $A_i \cap A_j = \emptyset$ whenever $i \neq j$. For example, $A_1 = [0, 1[, A_2 = [1, 2[, A_3 = [2, 3[, ...$ are disjoint. A **partition** of Ω is a sequence of disjoint sets $A_1, A_2, ...$ such that $\cup_{i=1}^{\infty} A_i = \Omega$. Given an event A, define the indicator function of A by

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$$

A sequence of sets $A_1, A_2, ...$ is **monotone increasing** if $A_1 \subset A_2 \subset \cdots$ and we define $\lim_{n \longrightarrow \infty} A_n = \cup_{i=1}^{\infty} A_i$. A sequence of sets $A_1, A_2, ...$ is **monotone decreasing** if $A_1 \supset A_2 \supset \cdots$ and then we define $\lim_{n \longrightarrow \infty} A_n = \cap_{i=1}^{\infty} A_i$. In either case, we will write $A_n \to A$.

Example. Let $\Omega = \mathbb{R}$ and let $A_i = [0, \frac{1}{i}[$ for $i = 1, 2, ...$ Then $\cup_{i=1}^{\infty} A_i = [0, 1]$ and $\cap_{i=1}^{\infty} A_i = \{0\}$. If instead we define $A_i =]0, \frac{1}{i}[$ then $\cup_{i=1}^{\infty} A_i =]0, 1[$ and $\cap_{i=1}^{\infty} A_i = \emptyset$.

4.2 Probability

We will assign a real number $P(A)$ to every event A, called the probability of A. We also call P a **probability distribution** or a **probability measure**. To qualify as a probability, P must satisfy three axioms:

Axiom 1 : $\forall A,\ P(A) \geq 0$

Axiom 2 : $P(\Omega) = 1$

Axiom 3 : If $A_1, A_2, ...$ are disjoint then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

There are many interpretations of $P(A)$. The two common interpretations are frequencies and degrees of beliefs. In the frequency interpretation, $P(A)$ is the long run proportion of times that A is true in repetitions. For example, if we say that the probability of heads is $1/2$, we mean that if we flip the coin many times then the proportion of times we get heads tends to $1/2$ as the number of tosses increases. An infinitely long, unpredictable sequence of tosses whose limiting proportion tends to a constant is an idealization, much like the idea of a straight line in geometry. The degree-of-belief interpretation is that $P(A)$ measures an observer's strength of belief that A is true. In either interpretation, we require that Axioms 1 to 3 hold. The difference in interpretation will not matter much until we deal with statistical inference. There, the differing interpretations lead to two schools of inference: the frequentist and the Bayesian schools.

One can derive many properties of P from the axioms, such as

$$P(\emptyset) = 0$$

$$A \subset B \Longrightarrow P(A) \leq P(B)$$

$$0 \leq P(A) \leq 1$$

$$P(A^c) = 1 - P(A)$$

$$A \cap B = \emptyset \Longrightarrow P(A \cup B) = P(A) + P(B)$$

We have the following basic theorem that allows us to decompose the calculation of the probability of B into the sum of the probabilities of the sets $A_i \cap B$. Often these are

easier to compute.

Theorem 4.2.1 (Law of total probability, unconditioned version) *Let $A_1, A_2, ...$ be events that form a partition of the sample space S. Let B be any event. Then*

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots.$$

Proof. The events $(A_1 \cap B), (A_2 \cap B), ...$are disjoint, and their union is B. Hence, the result follows immediately from the additivity property. ■

Suppose now that A and B are two events such that A contains B (in symbols, $A \supseteq B$). In words, all outcomes in B are also in A. Intuitively, A is a "larger" event than B, so we would expect its probability to be larger. We have the following result.

Theorem 4.2.2 *Let A and B be two events with $A \supseteq B$. Then*

$$P(A) = P(B) + P(A \cap B^c).$$

Proof. We can write $A = B \cup (A \cap B^c)$, where B and $A \cap B^c$ are disjoint. Hence, $P(A) = P(B) + P(A \cap B^c)$ by additivity. ■

A less obvious property is given in the following theorem.

Theorem 4.2.3 (Principle of inclusion–exclusion, two-event version) *For any events A and B,*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof. Write $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$ and note that these events are disjoint. Hence, making repeated use of the fact that P is additive for disjoint events,

we see that

$$\begin{aligned}
P(A \cup B) &= P((A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)) \\
&= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \\
&= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) + P(A \cap B) - P(A \cap B) \\
&= P(\underbrace{(A \cap B^c) \cup (A \cap B)}_{=A}) + P(\underbrace{(A^c \cap B) \cup (A \cap B)}_{=B}) - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B).
\end{aligned}$$

■

Example. Two coin tosses. Let H_1 be the event that heads occurs on toss 1 and let H_2 be the event that heads occurs on toss 2. If all outcomes are equally likely, then $P(H_1 \cup H_2) = P(H_1) + P(H_2) - P(H_1 \cap H_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$.

Theorem 4.2.4 (Continuity of Probabilities) *If $A_n \longrightarrow A$ then*

$$P(A_n) \longrightarrow P(A) \text{ as } n \longrightarrow \infty.$$

Proof. Suppose that A_n is monotone increasing so that $A_1 \subset A_2 \subset \cdots$. Let $A = \lim_{n \longrightarrow \infty} A_n = \cup_{i=1}^{\infty} A_i$. Define $B_1 = A_1, B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}, B_3 = \{\omega \in \Omega : \omega \in A_3, \omega \notin A_1, \omega \notin A_2\}, ...$It can be shown that $B_1, B_2, ...$ are disjoint, $\forall n, A_n = \cup_{i=1}^{n} A_i = \cup_{i=1}^{n} B_i$ and $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$. From Axiom 3,

$$P(A_n) = P(\cup_{i=1}^{n} B_i) = \sum_{i=1}^{n} P(B_i)$$

and hence, using Axiom 3 again,

$$\lim_{n\longrightarrow\infty} P(A_n) = \lim_{n\longrightarrow\infty} \sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i) = P(\cup_{i=1}^{\infty} B) = P(A)$$

■

Sometimes we do not need to evaluate the probability content of a union; we need only know it is bounded above by the sum of the probabilities of the individual events. This is called subadditivity.

Theorem 4.2.5 (Subadditivity) *Let $A_1, A_2, ...$ be a finite or countably infinite sequence of events, not necessarily disjoint. Then*

$$P(A_1 \cup A_2 \cup \cdots) \leq P(A1) + P(A2) + \cdots$$

4.3 Probability on Finite Sample Spaces

Suppose that the sample space $\Omega = \{\omega_1, ..., \omega_n\}$ is finite. For example, if we toss a die twice, then Ω has 36 elements: $\Omega = \{(i, j); i, j \in \{1, ...6\}\}$. If each outcome is equally likely, then $P(A) = \frac{|A|}{36}$ where $|A|$ denotes the number of elements in A. The probability that the sum of the dice is 11 is $\frac{2}{36}$ since there are two outcomes that correspond to this event.

If Ω is finite and if each outcome is equally likely, then

$$P(A) = \frac{|A|}{|\Omega|}$$

which is called the **uniform probability distribution**.

Example. Suppose now that we flip three different fair coins. The outcome can be

written as a sequence of three letters, with each letter being H (for heads) or T (for tails). Thus, $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.Here $|\Omega| = 8$, and each of the events is equally likely. Hence, $P(\{HHH\}) = 1/8, P(\{HHH, TTT\}) = 2/8 = 1/4$, etc. Note also that, by additivity, we have, for example, that P(exactly two heads) $= P(\{HHT, HTH, THH\}) = 1/8 + 1/8 + 1/8 = 3/8$, etc.

To compute probabilities, we need to count the number of points in an event A. Methods for counting points are called combinatorial methods. We needn't delve into these in any great detail. We will, however, need a few facts from counting theory that will be useful later. Given n objects, the number of ways of ordering these objects is $n! = n(n-1)(n-2)...3.2.1$.For convenience, we define $0! = 1$. We also define

$$C_n^k = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

read "n choose k", which is the number of distinct ways of choosing k objects from n. For example, if we have a class of 20 people and we want to select a committee of 3 students, then there are

$$\left(\begin{array}{c} 20 \\ 3 \end{array} \right) = \frac{20!}{3!17!} = \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = 1140$$

possible committees. We note the following properties:

$$\left(\begin{array}{c} n \\ 0 \end{array} \right) = \left(\begin{array}{c} n \\ n \end{array} \right) = 1 \quad \text{and} \quad \left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right)$$

4.4 Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2} \times \frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. The formal definition of independence is as follows:

Définition 4.4.1 *Two events A and B are independent if*

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

and we write $A \amalg B$. A set of events $\{A_i : i \in I\}$ is independent if

$$P(\cap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$$

for every finite subset J of I.

Independence can arise in two distinct ways. Sometimes, we explicitly assume that two events are independent. For example, in tossing a coin twice, we usually assume the tosses are independent which reflects the fact that the coin has no memory of the first toss. In other instances, we derive independence by verifying that $P(A \cap B) = P(A)P(B)$ holds. For example, in tossing a fair die, let $A = \{2, 4, 6\}$ and let $B = \{1, 2, 3, 4\}$. Then, $A \cap B = \{2, 4\}$, $P(A \cap B) = 2/6 = P(A)P(B) = (1/2) \times (2/3)$ and so A and B are independent. In this case, we didn't assume that A and B are independent — it just turned out that they were.

Suppose that A and B are disjoint events, each with positive probability. Can they be independent? No. This follows since $P(A)P(B) > 0$ yet $P(A \cap B) = P(\emptyset) = 0$.

Example. Toss a fair coin 10 times. Let $A =$"at least one head." Let T_j be the event that tails occurs on the j^{th} toss. Then

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - P(\text{all tails}) \\ &= 1 - P(T_1 \cap T_2 \cap \cdots \cap T_{10}) \\ &= 1 - P(T_1)P(T_2)\cdots P(T_{10}) \text{ using independence} \\ &= 1 - \left(\frac{1}{2}\right)^{10} \approx 0.999 \end{aligned}$$

Example. Two people take turns trying to sink a basketball into a net. Person 1 succeeds with probability 1/3 while person 2 succeeds with probability 1/4. What is the probability that person 1 succeeds before person 2?

Let E denote the event of interest. Let A_j be the event that the first success is by person 1 and that it occurs on trial number j. Note that $A_1, A_2, ...$ are disjoint and that $E = \cup_{j=1}^{\infty} A_j$. Hence,

$$P(E) = \sum_{j=1}^{\infty} P(A_j)$$

Now, $P(A_1) = 1/3$. A_2 occurs if we have the sequence person 1 misses, person 2 misses, person 1 succeeds. This has probability $P(A2) = (2/3)(3/4)(1/3) = (1/2)(1/3)$. Following this logic we see that $P(A_j) = (1/2)^{j-1}(1/3)$. Hence,

$$P(E) = \sum_{j=1}^{\infty} \left(\frac{1}{2}\right)^{j-1} \frac{1}{3} = \frac{1}{3} \sum_{j=1}^{\infty} \left(\frac{1}{2}\right)^{j-1} = \frac{1}{3} \frac{1}{1 - \frac{1}{2}} = \frac{2}{3}.$$

Here we used that fact that, if $0 < r < 1$ then $\sum_{j=k}^{\infty} r^j = \frac{r^k}{1-r}$.

4.5 Conditional Probability

Assuming that $P(B) > 0$, we define the conditional probability of A given that B has occurred as follows:

Définition 4.5.1 *If P(B) > 0 then the conditional probability of A given B is*

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

Think of $P(A|B)$ as the fraction of times A occurs among those in which B occurs. For any fixed B such that $P(B) > 0$, $P(\cdot|B)$ is a probability (i.e., it satisfies the three axioms of probability). In particular, $P(A|B) \geq 0$, $P(\Omega|B) = 1$ and if $A1, A2, ...$ are disjoint then $P(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$. But it is in general **not** true that $P(A|B \cup C) = P(A|B) + P(A|C)$. The rules of probability apply to events on the left of the bar.

In general it is not the case that $P(A|B) = P(B|A)$. People get this confused all the time. For example, the probability of spots given you have measles is 1 but the probability that you have measles given that you have spots is not 1. In this case, the difference between $P(A|B)$ and $P(B|A)$ is obvious but there are cases where it is less obvious.

Example. A medical test for a disease D has outcomes + and −. The probabilities are:

	D	D^c
+	0.009	0.099
−	0.001	0.891

From the definition of conditional probability,

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + .001} = 0.9$$

and

$$P(-|Dc) = \frac{P(- \cap D^c)}{P(D^c)} = \frac{0.891}{0.891 + 0.099} \approx 0.9.$$

Apparently, the test is fairly accurate. Sick people yield a positive 90 percent of the time and healthy people yield a negative about 90 percent of the time.

Suppose you go for a test and get a positive. What is the probability you have the disease? Most people answer 0.9. The correct answer is

$$P(D|+) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} \approx 0.08.$$

The lesson here is that you need to compute the answer numerically. Don't trust your intuition.

Proposition 4.5.1 (Compound probability formula) *Let n events $A_1, ..., A_n$ be such that $P(A_1 \cap ... \cap A_n) \neq 0$. Then*

$$P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n|A_1 \cap A_2... \cap A_{n-1}).$$

Example. An urn initially contains 7 black balls and 3 white balls. We successively draw 3 balls: if we draw a black one, we remove it, if we draw a white one, we remove it, and we add a black one instead. What is the probability of drawing 3 blanks in a row?

We denote by $B_i$the event "The i-th ball drawn is white". The desired probability is

$$P(B_1 \cap B_2 \cap B_3) = P(B_1)P(B_2|B_1)P(B_3|B_1 \cap B_2).$$

Clearly, $P(B_1) = 3/10$. Now, if B_1 is made, before the 2nd draw, the urn consists of 8 black and 2 white balls. We therefore have: $P(B_2|B_1) = 2/10$. If B_1 and B_2 are made, before the 3rd draw, the urn consists of 9 black balls and 1 white. We deduce $P(B_3|B_1 \cap B_2) = 1/10$. Finally $P(B_1 \cap B_2 \cap B_3) = 6/1000 = 3/500$.

4.6 Bayes' Theorem

First, we need a preliminary result. This formula makes it possible to calculate the probability of an event B by breaking it down according to a complete system of events (Indeed, B is equal to the disjoint union of $B \cap A_n$)

Theorem 4.6.1 (The Law of Total Probability) *Let $A_1, ..., A_k$ be a partition of Ω. Then, for any event B,*

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

Proof. Define $C_j = B \cap A_j$ and note that $C_1, ..., C_k$ are disjoint and that $B = \cup_{j=1}^{k} C_j$. Hence,

$$P(B) = P(\cup_{j=1}^{k} C_j) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j)$$

since $P(B \cap A_j) = P(B|A_j)P(A_j)$ from the definition of conditional probability. ■

Theorem 4.6.2 (Bayes' Theorem) *Let $A_1, ..., A_k$ be a partition of Ω such that $P(A_i) >$*

0 *for each* i. *If* $P(B) > 0$ *then, for each* $i = 1, ..., k$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} \ .$$

Proof. We apply the definition of conditional probability twice, followed by the law of total probability:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(Ai)}{P(B)} = \frac{P(B|Ai)P(Ai)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}.$$

■

Example. We divide emails into three categories: A_1 = "spam," A_2 ="low priority" and A_3 = "high priority." From previous experience we find that $P(A_1) = 0.7$, $P(A_2) = 0.2$ and $P(A_3) = 0.1$. Of course, $0.7 + 0.2 + 0.1 = 1$. Let B be the event that the email contains the word "free". From previous experience, $P(B|A_1) = 0.9, P(B|A_2) = 0.01, P(B|A_3) = 0.01$. (Note: $0.9+0.01+0.01 \neq 1$). We receive an email with the word "free". What is the probability that it is spam?

Bayes' theorem yields,

$$P(A_1|B) = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (.001 \times 0.1)} = 0.995.$$

4.7 Random Variables and Distribution

In previous sections, we discussed the probability model as the central object of study in the theory of probability. This required defining a probability measure P on a class of subsets of the sample space Ω. It turns out that there are simpler ways of presenting a particular probability assignment than this — ways that are much more convenient

to work with than P. This section is concerned with the definitions of random variables, distribution functions, probability functions, and the development of the concepts necessary for carrying out calculations for a probability model using these entities. This section also discusses the concept of the conditional distribution of one random variable, given the values of others. Conditional distributions of random variables provide the framework for discussing what it means to say that variables are related, which is important in many applications of probability and statistics.

4.7.1 Random Variables

The previous sections explained how to construct probability models, including a sample space Ω and a probability measure P. Once we have a probability model, we may define *random variables* for that probability model.

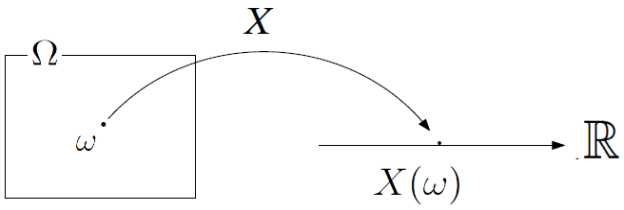
Intuitively, a random variable assigns a numerical value to each possible outcome in the sample space. For example, if the sample space is {rain, snow, clear}, then we might define a random variable X such that $X = 3$ if it rains, $X = 6$ if it snows, and $X = -2.7$ if it is clear.

More formally, we have the following definition.

Définition 4.7.1 *A random variable is a function from the sample space Ω to the set $\mathbb{R}$ of all real numbers that assigns a real number $X(\omega)$ to each outcome ω.*

$$X : \Omega \longrightarrow \mathbb{R}$$

$$\omega \longmapsto X(\omega) = x$$



The random variable described above could be written formally as X :{rain, snow, clear} $\to \mathbb{R}$ by $X(\text{rain})= 3$, $X(\text{snow})= 6$, and $X(\text{clear}) = -2.7$. We will return to this example below.

We now present several further examples. The point is, we can define random variables any way we like, as long as they are functions from the sample space to $\mathbb{R}$.

Example. If the sample space corresponds to flipping three different coins, then we could let X be the total number of heads showing, let Y be the total number of tails showing, let $Z = 0$ if there is exactly one head, and otherwise $Z = 17$, etc.

Example. If the sample space corresponds to rolling two fair dice, then we could let X be the square of the number showing on the first die, let Y be the square of the number showing on the second die, let Z be the sum of the two numbers showing, let W be thesquare of the sum of the two numbers showing, let R be the sum of the squares of the two numbers showing, etc.

Example. (Constants as Random Variables) As a special case, every constant value c is also a random variable, by saying that $c(\omega) = c$ for all $\omega \in \Omega$. Thus, 5 is a random variable, as is 3 or -21.6.

Example. (Indicator Functions) One special kind of random variable is worth mentioning. If A is any event, then we can define the *indicator function* of A, written $\mathbb{I}_A$, to be the random variable

$$\mathbb{I}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

which is equal to 1 on A, and is equal to 0 on A^c.

Given random variables X and Y, we can perform the usual arithmetic operations on them. Thus, for example, $Z = X^2$ is another random variable, defined by $Z(\omega) = X^2(\omega) = (X(\omega))^2 = X(\omega) \times X(\omega)$. Similarly, if $W = XY^3$, then $W(\omega) = X(\omega) \times Y(\omega) \times Y(\omega) \times Y(\omega)$, etc. Also, if $Z = X + Y$, then $Z(\omega = X(\omega) + Y(\omega)$, etc.

Example. Consider rolling a fair six-sided die, so that $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let X be the number showing, so that $X(\omega) = \omega$ for $\omega \in \Omega$. Let Y be three more than the number showing, so that $Y(\omega) = \omega+3$. Let $Z = X^2+Y$. Then $Z(\omega) = X(\omega)^2+Y(\omega) = \omega^2 + \omega + 3$. So $Z(1) = 5, Z(2) = 9$, etc.

We write $X = Y$ to mean that $X(\omega) = Y(\omega)$ for all $\omega \in \Omega$. Similarly, we write $X \leq Y$ to mean that $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, and $X \geq Y$ to mean that $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$. For example, we write $X \leq c$ to mean that $X(\omega) \leq c$ for all $\omega \in \Omega$.

Example. Again consider rolling a fair six-sided die, with $\Omega = \{1, 2, 3, 4, 5, 6\}$. For $\omega \in \Omega$, let $X(\omega) = \omega$, and let $Y = X + \mathbb{I}_{\{6\}}$. This means that

$$Y(\omega) = X(\omega) + \mathbb{I}_{\{6\}}(\omega) = \begin{cases} \omega & \omega \leq 5 \\ 7 & \omega = 6 \end{cases}$$

If S is infinite, then a random variable X can take on infinitely many different values.

Example. If $\Omega = \{1, 2, 3, ...\}$, if X is defined by $X(\omega) = \omega^2$, then we always have $X \geq 1$. But there is no largest value of $X(\omega)$ because the value $X(\omega)$ increases without bound as $\omega \to \infty$. We shall call such a random variable an *unbounded random variable*. This is pretty much the same as a statistical variable except that in the case of a statistical variable one evaluates a realized behavior (average, etc) whereas in the case of random variables one assumes a future behavior (In this case, we speak of expectation

rather than average for example) or theoretical.

Random variables are used to model the outcome of a non-deterministic mechanism.

Définition 4.7.2 *A random variable (or r.v.) is a map $X : \Omega \to \mathbb{R}$. If $X(\Omega)$ is at most* countable*, we say that X is a* ***discrete*** *r.v. otherwise we say that it is* ***continuous****.*

Finally, suppose X is a random variable. We know that different states ω occur with different probabilities. It follows that $X(\omega)$ also takes different values with different probabilities. These probabilities are called the *distribution of* X; we consider them next.

4.7.2 Distributions of Random Variables

Because random variables are defined to be functions of the outcome ω, and because the outcome ω is assumed to be random (i.e., to take on different values with different probabilities), it follows that the value of a random variable will itself be random (as the name implies).

Specifically, if X is a random variable, then what is the probability that X will equal some particular value x? Well, $X = x$ precisely when the outcome ω is chosen such that $X(\omega) = x$.

A random variable is totally defined by its law of probability and it is characterized by:

– the set of values it can take (its domain of definition);

– the probabilities attributed to each of the potentially taken values $P(X = x)$.

In this case, the law of the random variable is the law of probability on the set of possible values of X which assigns the probability $P(X = x)$ to the singleton $\{x\}$.

Example. Let us again consider the random variable of the first example above, where $\Omega = \{\text{rain, snow, clear}\}$, and X is defined by $X(\text{rain}) = 3$, $X(\text{snow}) = 6$, and

$X(\text{clear}) = -2.7$. Suppose further that the probability measure P is such that $P(\text{rain}) = 0.4$, $P(\text{snow}) = 0.15$, and $P(\text{clear}) = 0.45$. Then clearly, $X = 3$ only when it rains, $X = 6$ only when it snows, and $X = -2.7$ only when it is clear. Thus, $P(X = 3) = P(\text{rain}) = 0.4$, $P(X = 6) = P(\text{snow}) = 0.15$, and $P(X = -2.7) = P(\text{clear}) = 0.45$. Also, $P(X = 17) = 0$, and in fact $P(X = x) = P(\emptyset) = 0$ for all $x \notin \{3, 6, -2.7\}$. We can also compute that

$$P(X \in \{3, 6\}) = P(X = 3) + P(X = 6) = 0.4 + 0.15 = 0.55,$$

while

$$P(X < 5) = P(X = 3) + P(X = -2.7) = 0.4 + 0.45 = 0.85,$$

etc.

4.7.3 Distribution function

Définition 4.7.3 (cumulative distribution function) *The distribution function (CDF) of an r.v. X is the map F_X of $\mathbb{R}$ in $[0, 1]$ defined by*

$$F_X(x) = P(X \leq x) = P(X^{-1}(]-\infty, x])$$

We are often interested in the cumulative probability. For example in the case of probabilities over $\mathbb{N}$:

$$P(X \leq n) = P(X = 0 \text{ or } X = 1 \text{ or} \cdots \text{ or } X = n).$$

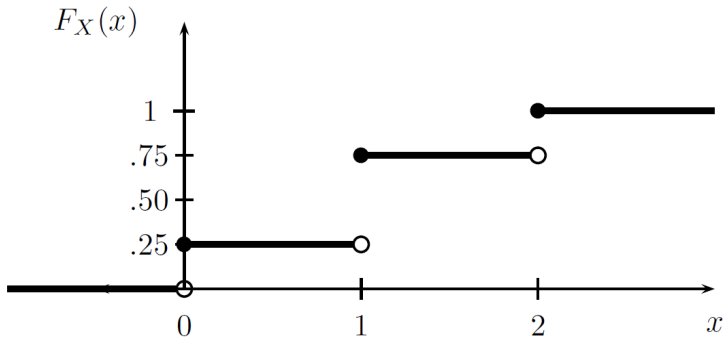
Since the events are mutually exclusive, we obtain

$$P(X \leq n) = \sum_{j=0}^{n} P(X = j).$$

We will see later that the CDF effectively contains all the information about the random variable. Sometimes we write the CDF as F instead of F_X.

Example. Flip a fair coin twice and let X be the number of heads. Then $P(X = 0) = P(X = 2) = 1/4$ and $P(X = 1) = 1/2$. The distribution function is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$



The CDF is shown in the above figure. Although this example is simple, study it carefully. CDF's can be very confusing. Notice that the function is right continuous, non-decreasing, and that it is defined for all x, even though the random variable only takes values $0, 1$, and 2. Do you see why $F_X(1.4) = 0.75$?

The following result shows that the CDF completely determines the distribution of a random variable.

Theorem 4.7.1 *Let X have CDF F and let Y have CDF G. If $F(x) = G(x)$ for all x, then $P(X \in A) = P(Y \in A)$ for all A.*

In fact, we say thet the two random variables X and Y are **equal in distribution** — written $X \stackrel{d}{=} Y$ — if $F_X(x) = F_Y(x)$ for all x and this does not mean that X and Y are equal. Rather, it means that all probability statements about X and Y will be the same. For example, suppose that $P(X = 1) = P(X = -1) = 1/2$. Let $Y = -X$. Then $P(Y = 1) = P(Y = -1) = 1/2$ and so $X \stackrel{d}{=} Y$. But X and Y are not equal. In fact, $P(X = Y) = 0$.

Theorem 4.7.2 *A function F mapping the real line to $[0, 1]$ is a CDF for some probability P if and only if F satisfies the following three conditions:*
(i) F is non-decreasing: $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$.
(ii) F is normalized: $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
(iii) F is right-continuous: $F(x) = F(x^+)$ for all x, where $F(x^+) = \lim_{y \stackrel{>}{\to} x} F(y)$.

Proof. *(i)* Since $x_1 < x_2$, then we have $]-\infty, x_1] \subset]-\infty, x_2]$ which gives

$$P(X \in]-\infty, x_1]) \leq P(X \in]-\infty, x_2]) \Longleftrightarrow F(x_1) \leq F(x_2)$$

(ii) Obvious
(iii) Suppose that F is a CDF. Let x be a real number and let $y_1, y_2, ...$ be a sequence of real numbers such that $y_1 > y_2 > \cdots$ and $y_i \xrightarrow[i\to\infty]{>} x$. Let $A_i =]-\infty, y_i]$ and let $A =]-\infty, x]$. Note that $A = \cap_{i=1}^{\infty} A_i$ and also note that $A_1 \supset A_2 \supset \cdots$. Because the events are monotone, $\lim_{i\longrightarrow\infty} P(A_i) = P(\cap_{i=1}^{\infty} A_i)$. Thus,

$$F(x) = P(A) = P(\cap_{i=1}^{\infty} Ai) = \lim_{i\longrightarrow\infty} P(A_i) = \lim_{i\longrightarrow\infty} F(y_i) = F(x^+).$$

■

Lemma 4.7.1 *Let F be the CDF for a random variable X. Then:*

(i) $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

(ii) $P(X > x) = 1 - F(x)$.

Proof. *(i)* Take $x_1 < x_2$, then

$$\begin{aligned} P(X \leq x_2) = P(X \in]-\infty, x_2]) &= P(\{X \in]-\infty, x_1]\} \cup \{X \in]x_1, x_2]\}) \\ &= P(X \in]-\infty, x_1]) + P(X \in]x_1, x_2]) \\ &= P(X \leq x_1) + P(x_1 < X \leq x_2) \\ &\Longrightarrow P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1) \end{aligned}$$

(ii) We replace x_2 in *(i)* by ∞. ■

4.7.4 Expectation and variance of a random variable

Average value of a random variable: Expectation

The intuitive idea of expectation has its origins in games of chance. Consider the following game: a die is rolled. Suppose that for a bet of 1 euro, we win 1 euro if the result obtained is even, 2 euros if the result is 1 or 3, and we lose 3 euros if the result is 5. Is it interesting to play this game? What is the average gain?

Let X be the random variable corresponding to the number of euros won or lost. The law of X is

x	-3	1	2
$P(X = x)$	1/6	1/2	1/3

The expected win, denoted $E[X]$, is then $E[X] = -3 \times 1/6 + 1 \times 1/2 + 2 \times 1/3 = 2/3$. The player therefore earns on average 2/3 euros for a bet of 1 euro.

An *expectation operator* is a function that assigns to each random variable X a real number $E[X]$ called the expectation or expected value of X.

Every expectation operator satisfies the following axioms.

Axiom E1 (Additivity). If X and Y are random variables, then X + Y is also a random variable, and

$$E[X + Y] = E[X] + E[Y].$$

Axiom E2 (Homogeneity). If X is random variable and a is a real number, then aX is also a random variable, and

$$E[aX] = aE[X].$$

These properties agree with either of the informal intuitions about expectations. Prices are additive and homogeneous. The price of a gallon of milk and a box of cereal is the sum of the prices of the two items separately. Also the price of three boxes of cereal is three times the price of one box. (The notion of expectation as fair price doesn't allow for volume discounts.)

Axiom E3 (Positivity). If X is random variable, then $X \geq 0$ implies $E[X] \geq 0$.

The expression $X \geq 0$, written out in more detail, means

$$X(\omega) \geq 0, \omega \in \Omega,$$

where Ω is the sample space. That is, X is always nonnegative. This axiom corresponds to intuition about prices, since goods always have nonnegative value and prices are also

nonnegative.

Définition 4.7.4 *The expectation of a random variable X is denoted $E[X]$. It represents the average value taken by the variable X.*

1. *If X is a discrete variable with values in $D = \{x_1, ..., x_n\}$, its expectation is*

$$E[X] = x_1 P(X = x_1) + ... + x_n P(X = x_n) = \sum_{i=1}^{n} x_i P(X = x_i)$$

2. *If X is a discrete variable with values in the infinite set $D = \{xi : i \geq 1\}$, when the sum is well defined, its expectation is*

$$E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i)$$

When a variable X verifies $E[X] = 0$, we say that the variable is **centered**.

Proposition 4.7.1 *1. the expectation is linear: let a and $b \in \mathbb{R}$ and two random variables X and Y with finite expectation then*

$$E[aX + bY] = aE[X] + bE[Y].$$

2. *If $X \leq Y$ then $E[X] \leq E[Y]$.*

The Multiplicativity Non-Property

One might suppose that there is a property analogous to the additivity property, except with multiplication instead of addition

$$E[XY] = E[X]E[Y], \text{ Uncorrelated } X \text{ and } Y \text{ only!}$$

As the editorial comment says, this property does not hold in general. We will later see that when it does hold we have a special name for this situation: we say the variables X and Y are uncorrelated.

Taking a Function Outside an Expectation

Suppose g is a linear function defined by

$$g(x) = a + bx, x \in \mathbb{R},$$

where a and b are real numbers. Then

$$E[g(X)] = g(E[X]), \text{ Linear } g \text{ only!}$$

The reason for the editorial comment is that property does not hold for general functions g, only for linear functions. Sometime you will be tempted to use it for a nonlinear function g. Don't! Remember that it is a "non-property."

For example, you may be asked to calculate $E[1/X]$ for some random variable X. The "non-property", if it were true, would allow to take the function outside the expectation and the answer would be $1/E[X]$, but it isn't true, and, in general

$$E[\frac{1}{X}] \neq \frac{1}{E[X]}.$$

4.7.5 Moments

If k is a positive integer, then the real number

$$\alpha_k = E[X^k]$$

is called the $k-th$ moment of the random variable X.

If p is a positive real number, then the real number

$$\beta_p = E[|X|^p]$$

is called the $p-th$ absolute moment of the random variable X.

If k is a positive integer and $\mu = E[X]$, then the real number

$$\mu_k = E[(X-\mu)^k]$$

is called the $k-th$ central moment of the random variable X.

That's not the whole story on moments. We can define lots more, but all moments are special cases of one of the two following concepts.

If k is a positive real number and a is any real number, then the real number $E[(X-a)^k]$ is called the $k-th$ moment about the point a of the random variable X. We introduce no special symbol for this concept. Note that the $k-th$ ordinary moment is the special case $a = 0$ and the $k-th$ central moment is the case $a = \mu$.

If p is a positive real number and a is any real number, then the real number $E[|X-a|^k]$ is called the $p-th$ absolute moment about the point a of the random variable X. Note that the $p-th$ absolute moment is the special case $a = 0$.

The first ordinary moment of a random variable X is also called the mean of X. It is commonly denoted by the Greek letter μ. Note that α_1, μ, and $E[X]$ are different notations for the same thing. We will use them all throughout the course.

When there are several random variables under discussion, we denote the mean of each using the same Greek letter μ, but add the variable as a subscript to distinguish them: $\mu_X = E[X], \mu_Y = E[Y]$, and so forth.

Theorem 4.7.3 *Suppose a real-valued random variable X is symmetric about the point a. If the mean of X exists, it is equal to a. Every higher odd integer central moment of X that exists is zero*

In notation, the two assertions of the theorem are

$$E[X] = \mu = a$$

and

$$\mu_{2k+1} = E[(X - \mu)^{2k+1}] = 0, \text{ for any positive integer } k.$$

The proof is left as an exercise.

Second Moments and Variances

The preceding section says all that can be said in general about first moments. As we shall now see, second moments are much more complicated.

The most important second moment is the second central moment, which also has a special name. It is called the *variance* and is often denoted σ^2. We will see the reason for the square presently. We also use the notation $var(X)$ for the variance of X. So

$$\sigma^2 = \mu_2 = var(X) = E[(X - \mu)^2].$$

As we did with means, when there are several random variables under discussion, we denote the variance of each using the same Greek letter σ, but add the variable as a subscript to distinguish them: $\sigma_X^2 = var(X)$, $\sigma_Y^2 = var(Y)$, and so forth.

Note that variance is just an expectation like any other, the expectation of the random variable $(X - \mu)^2$.

Corollary 4.7.1 *If X is a random variable having first and second moments, then*

$$var(X) = E[X^2] - E[X]^2.$$

There are various ways of restating the corollary in symbols, for example

$$\sigma_X^2 = E[X^2] - \mu_X^2$$

and

$$\mu_2 = \alpha_2 - \alpha_1^2$$

As always, mathematics is invariant under changes of notation. The important thing is the concepts symbolized rather than the symbols themselves.

Theorem 4.7.4 *Suppose X is a random variable having first and second moments and a and b are real numbers, then*

$$var(a + bX) = b^2 var(X)$$

Note that the right hand side does not involve the constant part a of the linear transformation $a + bX$. Also note that the b comes out squared.

The nonnegative square root of the variance is called the *standard deviation*. Conversely, the variance is the square of the standard deviation. The symbol commonly used for the standard deviation is σ. That's why the variance is usually denoted σ^2.

As with the mean and variance, we use subscripts to distinguish variables σ_X , σ_Y , and so forth.

It might have just occurred to you to ask why anyone would want two such closely re-

lated concepts. Won't one do? In fact more than one introductory (freshman level) statistics textbook does just that, speaking only of standard deviations, never of variances. But for theoretical probability and statistics, this will not do. Standard deviations are almost useless for theoretical purposes. The square root introduces nasty complications into simple situations. So for theoretical purposes variance is the preferred concept.

In contrast, for all practical purposes standard deviation is the preferred concept, as evidenced by the fact that introductory statistics textbooks that choose to use only one of the two concepts invariably choose standard deviation. The reason has to do with units of measurement and measurement scales. Suppose we have a random variable X whose units of measurement are inches, for example, the height of a student in the class. What are the units of $E[X]$, $var(X)$, and σ_X, assuming these quantities exist?