STATISTICS AND PROBABILITY

Prepared by:

DR. AKRAM ALQESMAH

Contents

Chapter 1

Introduction to Statistics

1.1 Definition of Statistics

The science of statistics deals with the collection, analysis, interpretation, and presentation of data to make good decisions.

Statistics is divided into two types, **descriptive statistics** and **inferential statistics**.

Descriptive Statistics:

Descriptive statistics is concerned with collecting, analyzing and presenting the data without making decisions about them. Therefore, it consists of graphical and numerical procedures to summarize and process data. Often the descriptive statistics is a step before the inferential statistics.

Inferential Statistics:

Using the graphical and numerical results of the descriptive statistics about data to make predictions, forecasts and estimates to assist decision making.

1.2 Some Concepts in Statistics

Key Terms:

- **A population** is a collection of persons, things, or objects under study.
- **A sample** is a subset of the population. The process of selection of a sample is called **sampling**. A **random sample** is one in which each member of population has an equal chance to being included in it.
- **A statistic** is a number that represents a property of the sample.

- **A parameter** is a numerical characteristic of the whole population that can be estimated by a statistic.
- **A variable** is a characteristic or measurement that can be determined for each member of a population. Usually, variables notated by capital letters such as X and Y. Variables may be **numerical** or **categorical**.
  (i) **Numerical variables** take on values with equal units such as weight in pounds and time in hours. For example, let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable.
  (ii) **Categorical variables** place the person or thing into a category. For example, let Y be a person's martial status, then Y either be single, married, divorced and windowed. Thus, Y is a categorical variable.
- **Data** are the actual values of the variable. They may be numbers or they may be words.

Example 1.1. Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution:
The population is all first year students attending ABC College this term.
The sample could be the 100 first year students at the college who surveyed in the study.
The parameter is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term: the population mean.
The statistic is the average (mean) amount of money spent (excluding books) by first year college students in the sample.
The variable could be the amount of money spent (excluding books) by one first year student. Let $X =$ the amount of money spent (excluding books) by one first year student attending ABC College.
The data are the dollar amounts spent by the first year students. Examples of the data are 150, 200, and 225.

Example 1.2. Determine what the key terms refer to in the following study.
A study was conducted at a local college to analyze the average cumulative GPAs of students who graduated last year. Choose from the following the best describes of the key terms.

(a) all students who attended the college last year

(b) the cumulative GPA of one student who graduated from the college last year

(c) 3.65, 2.80, 1.50, 3.90

(d) a group of students who graduated from the college last year, randomly selected

(e) the average cumulative GPA of students who graduated from the college last year

(f) all students who graduated from the college last year

(g) the average cumulative GPA of students in the study who graduated from the college last year

Solution:

1. Population = (f)
2. Sample = (d)
3. Parameter = (e)
4. Statistic = (g)
5. Variable = (b)
6. Data = (c)

1.3 Data and Sampling

Data:

Data may come from a population or from a sample. Lowercase letters like x or y generally are used to represent data values. Most of data can be characterize into the following categories:

- **Qualitative data** are the data which can described by words or letters, for example, hair color, blood types, ethnic group, the car a person drives, and the street a person lives. Qualitative data divided into two types:
  - (i) **Qualitative ordered data** are qualitative data which described by ordered words or letters, for example level of incomes (high, medium, low) and grade average (excellent, very good, good, passable, weak).
  - (ii) **Qualitative named data** are qualitative data which described by words or letters without ordering, for example eyes colors, blood groups, etc.

- **Quantitative data** are the data which can described by numbers. Amount of money, pulse rate, weight, number of people living in your town, and number of students in a class are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

  (i) **Quantitative discrete data** are the results of counting. Therefore, it always takes positive integers. For example, the number of accidents in a week, the number of students in a college.

  (ii) **Quantitative continuous data** may include fractions, decimals, or irrational numbers. For example, the weight, lengths or times.

Example 1.3. determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete.

(a) the number of pairs of shoes you own

(b) the type of car you drive

(c) the distance from your home to the nearest grocery store

(d) the number of classes you take per school year

(e) the type of calculator you use

(f) weights of sumo wrestlers

(g) number of correct answers on a quiz

Solution:
Items a, d, and g are quantitative discrete; items c and f are quantitative continuous; items b and e are qualitative named data.

Sampling:

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing**. There are several different methods of random sampling. In this subsection, four methods of random sampling are considered which are **the simple random sample, the systematic sample, the stratified sample and cluster sample**.

(i) **Simple random sample:** In this type of sampling we label each member of the population by a number starting from one (i.e. $1, 2, 3, \cdots$). After typing the numbers in rows and columns, randomly select a number and starting from that number keep selecting the other numbers of the sample vertically or horizontally or diagonally.

Example 1.4. Suppose in some study there are 50 students. We need to select a simple random sample consists of 15 students. That can be done as in the following:

- label the 50 students by numbers starting from 1 to 50 and type these numbers in rows and columns as follows:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50

- suppose we randomly select the number 5 and keep selecting diagonally to 49 then keep selecting vertically to the number 9 and again keep selecting diagonally to the number 36 and finally keep selecting horizontally to the number 32. Then we obtain the sample

$$5, 16, 27, 38, 49, 39, 29, 19, 9, 18, 36, 35, 34, 33, 32$$

(ii) **Systematic sample:** Suppose we have a population of N members and we need to select a systematic random sample with n members. To do that, follow the following steps:

(a) number the population members from 1 to N

(b) choose a length $S = \dfrac{N}{n} = h$

(c) randomly select a number say z, then the sample of n members will be as

$$z, z + h, z + 2h, \cdots, z + (n - 1)h$$

Example 1.5. A study contains 200 members. We need to select a systematic random sample with 10 members. To do that, choose the length

$$S = \frac{N}{n} = \frac{200}{10} = 20,$$

then after numbering the population from 1 to 200, select a number randomly (say 7). Therefore, the systematic sample will be as

$$7, 27, 47, 67, 87, 107, 127, 147, 167, 187.$$

(iii) **Stratified sample:** To choose a stratified sample, divide the population into groups called strata and then take **a proportionate** number from each stratum.

Suppose the population has N members and it is divided into k groups with $N_1, N_2, \cdots, N_k$ members, namely $N = N_1 + N_2 + \cdots + N_k$. To choose a stratified sample with n members we should take a proportionate number from each group i.e. $n_1, n_2, \cdots, n_k$, respectively, such that $n = n_1 + n_2 + \cdots + n_k$. To find the numbers n_i where $i = 1, 2, \cdots, k$, use the following formula.

$$\frac{n}{N} = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_k}{N_k}.$$

Example 1.6. A study contains 9000 students divided into four groups with 2000, 3000, 2500 and 1500 students, respectively. How a researcher could select a stratified sample contains 100 students?

Solution: We have the number of members in the population

$$N = 2000 + 3000 + 2500 + 1500 = 9000,$$

where $N_1 = 2000$, $N_2 = 3000$, $N_3 = 2500$ and $N_4 = 1500$.

Also the number of members of the sample

$$n = 100 = n_1 + n_2 + n_3 + n_4,$$

where n_1 is selected from N_1, n_2 from N_2, n_3 from N_3 and n_4 from N_4.

Therefore, n_1, n_2, n_3 and n_4 are determined as follows:

$$\begin{aligned}
\frac{n}{N} &= \frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4} \\
\frac{100}{9000} &= \frac{n_1}{2000} = \frac{n_2}{3000} = \frac{n_3}{2500} = \frac{n_4}{1500} \\
\frac{100}{9000} &= \frac{n_1}{2000} \Leftrightarrow n_1 = \frac{100 \times 2000}{9000} = 22,22 \approx 22 \\
\frac{100}{9000} &= \frac{n_2}{3000} \Leftrightarrow n_2 = \frac{100 \times 3000}{9000} = 33.33 \approx 33 \\
\frac{100}{9000} &= \frac{n_3}{2500} \Leftrightarrow n_3 = \frac{100 \times 2500}{9000} = 27,7 \approx 28 \\
\frac{100}{9000} &= \frac{n_4}{1500} \Leftrightarrow n_4 = \frac{100 \times 1500}{9000} = 16.66 \approx 17.
\end{aligned}.$$

(iv) **Cluster sample:** To choose a cluster sample, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. This means that, the number of members in the sample will be determined after sampling.

For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

1.4 Frequency Distribution and Data

1.4.1 Types of Data

Data can be characterized into two types:

(i) **Raw Data** is an initial collection of information. For example, the collection of information $2, 5, 3, 8, 11, 7, 9, 2$ is a raw data.

(ii) **Grouped Data** is data classified in frequency distribution table. For example, the following table shows a grouped data

Class interval	**Frequency**
11 − 20	4
21 − 30	5
31− 40	3
41 − 50	6

Table 1.1: Example of grouped data

1.4.2 Tabulating of Numerical Data

In this subsection, we will learn how to create the frequency distribution table of a grouped data. To do that, follow the following steps:

(i) Find the range of data, $R = Max(x_i) - Min(x_i) + 1$.

(ii) Select the number of class intervals m (usually be between 5–15). We can use the following formula to determine the number of classes

$$m = 1 + 3.322\ \log(n),$$

where n is the number of data.

(iii) Determine the class length (l) as follows: $l = \dfrac{R}{m}$ (rounded to the greater integer number).

(iv) Determine class bounds (limits). We can start from the smallest value of the data as the lower bound of the first class or choose the lower bound of the first class less than the smallest value of data by one or two numbers. Thus, for each class there are two bounds: **Lower bound (LB)** and **Upper bound (UB)**, where

$$UB = LB + l - 1.$$

(v) Determine class midpoints (y_i) as $y_i = \dfrac{UB + LB}{2}$ for each class.

Example 1.7. Initiate the frequency distribution table of the following data:

$x_i = 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58, 61, 60, 62$

Solution: We have the number of data $n = 23$.

(i) Determine the range, $R = 62 - 12 + 1 = 51$.

(ii) Select the number of class intervals

$$m = 1 + 3.322\ \log(n) = 1 + 3.322\ \log(23) = 5.52 \approx 6.$$

(iii) Determine the class length, $l = \dfrac{R}{m} = \dfrac{51}{6} = 8.5 \approx 9$.

(iv) Choose the first class starting from 11. Then the classes will be as

$$11 - 19, 20 - 28, 29 - 37, 38 - 46, 47 - 55, 56 - 64.$$

Then the frequency distribution table is given by

Class interval	**Frequency** (f_i)	**Class midpoint** (y_i)
11 – 19	3	15
20 – 28	6	24
29 – 37	4	33
38 – 46	5	42
47 – 55	1	51
56 – 64	4	60

Table 1.2

1.4.3 Some Important Concepts

(i) **Relative Frequency** (Rf_i)**:** for each class the relative frequency is equal

$$\frac{f_i}{\sum_{i=1}^{m} f_i} = \frac{f_i}{n},$$

where m is the number of classes. Recall that $n = \sum_{i=1}^{m} f_i$ is the number of data.

(ii) **Real Bounds:** For each class,
Lower real bound = Lower bound - 0.5
Upper real bound = Upper bound + 0.5.

(iii) **Cumulative Frequency:** Two types of cumulative frequency are determined:

- **Increasing Cumulative frequency (CF ↑):** For each class CF ↑ is equal the sum of all data less than its upper bound (UB).
- **Decreasing Cumulative frequency (CF ↓):** For each class CF ↓ is equal the sum of all data greater than its lower bound (LB).

Example 1.8. In Example 1.7, the relative frequency, the real bounds and the cumulative frequencies are determined in the following table:

Class	f_i	y_i	Rf_i	**Real Bounds**	CF ↑ **Less than UB**	CF ↓ **Greater than LB**
11 − 19	3	15	0.130	10.5 − 19.5	3	23
20 − 28	6	24	0.261	19.5 − 28.5	9	20
29 − 37	4	33	0.174	28.5 − 37.5	13	14
38 − 46	5	42	0.217	37.5 − 46.5	18	10
47 − 55	1	51	0.044	46.5 − 55.5	19	5
56 − 64	4	60	0.174	55.5 − 64.5	23	4
Total	23		1.00			

Table 1.3

1.5 Exercises 1

1. Determine what the key terms refer to in the following study.

   We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

2. Determine what the key terms refer to in the following study.

   An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

3. Determine the type of data in the following studies.

   (i) The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book.

   (ii) The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3.

   (iii) The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has 10 machines, one gym has 22 machines, and the other gym has 20 machines.

   (iv) The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet.

   (v) The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack.

   (vi) The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white.

4. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

(a) cluster sampling

(b) stratified sampling

(c) simple random sampling

(d) systematic sampling.

5. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

   (a) simple random

   (b) stratified

   (c) cluster

   (d) systematic

6. A study was done on 25 members for some topic and get the following data:

$$61{,}53, 42, 47, 35, 10, 19, 17, 21, 53, 32, 44, 73,$$
$$85, 73, 44, 81, 49, 25, 17, 43, 35, 17, 47, 89.$$

   Create the frequency distribution table of these data.

7. Complete the following frequency distribution table by the center of classes (y_i), the real boundaries, relative frequency (Rf_i), and increasing and decreasing cumulative frequencies (CF ↑ and CF ↓).

Class	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$
f_i	5	7	4	8	4	5

Chapter 2

Measures of Central Tendency

The **Mean**, the **Median** and the **Mode** are single values that represent a set of values, and may be used to indicate the general size of the members in a set. These three measures are called the measures of central tendency because their values tend to be in the center of the values.

2.1 The Mean

The arithmetic mean value is found by adding together the values of the members of a set and dividing the sum by the number of members in the set.

1. **Un-Grouped Data:**

   - **Raw data:**

     For the raw data the mean (denoted by $\bar{x}$) is defined as

     $$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

     Example 2.1. Find the mean for the following data

     $$x_i = 11, 12, 13, 14, 15, 17.$$

     Solution:

     $$\bar{x} = \frac{\sum x_i}{n} = \frac{11 + 12 + 13 + 14 + 15 + 17}{6} = \frac{82}{6} = 13.6.$$

   - **Frequented data:**

     In this type of data we have values with frequencies. Thus, the mean is defined as

     $$\bar{x} = \frac{\sum x_i f_i}{\sum f_i},$$

     where f_i is the frequency of the value x_i for i.

Example 2.2. Find the mean for the following data

Value x_i	19	20	22	24	18	25
f_i	5	7	4	8	4	5

Solution:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{19 \times 5 + 20 \times 7 + 22 \times 4 + 24 \times 8 + 18 \times 4 + 25 \times 5}{5 + 7 + 4 + 8 + 4 + 5} = \frac{712}{33} = 21.6.$$

2. **Grouped Data:**

In this type of data we need to find the centers of the classes interval (class midpoints y_i) and then the mean is found by the following formula

$$\bar{y} = \frac{\sum y_i f_i}{\sum f_i},$$

where y_i are the centers of classes and f_i are their frequencies.

Example 2.3. Find the mean for the following data

Class	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69
f_i	5	7	4	8	4	5

Solution:

From the following table we get

Class	f_i	y_i	$y_i f_i$
10 – 19	5	14.5	72.5
20 – 29	7	24.5	171.5
30 – 39	4	34.5	138
40 – 49	8	44.5	356
50 – 59	4	54.5	218
60 – 69	5	64.5	322.5
$\sum$	33		1278.5

$$\bar{y} = \frac{\sum y_i f_i}{\sum f_i} = \frac{1278.5}{33} = 38.74.$$

2.2 The Median M_e

A median of a set of values is the value which is located in the meddle of the values when the values are in numerical order (smallest to largest).

1. **Un-Grouped Data:**

   - **Raw data:**

     (i) If the number of values is **odd**, then the order of the median is $\frac{n+1}{2}$.

     (ii) If the number of values is **even**, then the median has two orders which are $\frac{n}{2}$ and $\frac{n}{2}+1$.

     Example 2.4. Find the median for the following data

     $$x_i = 11, 12, 13, 14, 15, 17, 10, 7, 8, 16, 20.$$

     Solution:

     (a) Order the data as $7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 20$.

     (b) The number of the values $n = 11$ odd, then the order of the median is $\frac{n+1}{2} = \frac{11+1}{2} = 6$.

     Therefore, the median $M_e = 13$.

     Example 2.5. Find the median for the following data

     $$x_i = 23, 14, 25, 15, 17, 29, 30, 27, 38, 40.$$

     Solution:

     (a) Order the data as $14, 15, 17, 23, 25, 27, 29, 30, 38, 40$.

     (b) The number of the values $n = 10$ even, then the orders of the median are $\frac{n}{2} = \frac{10}{2} = 5$ and $\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6$.

     Therefore, we have two medians $M_e^1 = 25$ and $M_e^2 = 27$. Hence the median is

     $$M_e = \frac{M_e^1 + M_e^2}{2} = \frac{25+27}{2} = \frac{52}{2} = 26.$$

- **Frequented data:**

  Example 2.6. Find the median for the following data

Value x_i	19	20	22	24	18	25
f_i	5	7	4	8	4	5

Solution:

After ordering the values in the table, we need to find the increasing cumulative frequency to determine the location of the median as follows

Value x_i	18	19	20	22	24	25
f_i	4	5	7	4	8	5
$CF_i \uparrow$	4	9	16	20	28	33

The number of values $n = 33$ is odd, then the order of the median is

$$\frac{n+1}{2} = \frac{33+1}{2} = \frac{34}{2} = 17.$$

Hence from the table, the median $M_e = 22$.

2. **Grounded Data:**

In this type of data, by using the increasing cumulative frequency of the class intervals we determine the **median class** and based on this class we determine the values of the terms in the following formula:

$$M_e = LB + \frac{\frac{\sum f_i}{2} - F_i}{f_i} \times l,$$

where
LB is the lower bound of the median class
F_i is the increasing cumulative frequency of the class before the median class
f_i is the frequency of the median class
l is the class length.

Example 2.7. Find the median for the following data

Class	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$
f_i	5	7	4	8	4	5

Solution:

We need to determine the increasing cumulative frequency as follows

Class	f_i	$CF_i \uparrow$
$10-19$	5	5
$20-29$	7	12
$30-39$	4	16
$40-49$	8	24
$50-59$	4	28
$60-69$	5	33

The number of data is $n = 33$ is odd, then the order of the median class is given by

$$\frac{n+1}{2} = \frac{33+1}{2} = 17.$$

Therefore, the median class is (40–49) and then $LB = 40$, $F_i = 16$, $f_i = 8$ and $l = 10$. Hence,

$$\begin{aligned} M_e =& LB + \frac{\frac{\sum f_i}{2} - F_i}{f_i} \times l = 40 + \frac{\frac{33}{2} - 16}{8} \times 10 \\ =& 40 + \frac{16.5 - 16}{8} \times 10 = 40 + \frac{0.5 \times 10}{8} \\ =& 40 + \frac{5}{8} = 40 + 0.625 \\ M_e =& 40.625 \end{aligned}$$

2.3 The Mode M_0

The modal value, or mode, is the most commonly occurring value in a set. If two values occur with the same frequency, the set is **bi–modal** and if the occurring of all the values are same then the set has no mode.

1. **Un-Grouped Data:**

   - **Raw data:**

     Example 2.8. Find the mode for the following data

     $$x_i = 11, 12, 13, 14, 12, 17, 11, 13, 10, 12, 14.$$

     Solution: It can be seen that the most frequented value of these data is 12. Therefore, the mode $M_0 = 12$.

     In the following example, the mode of qualitative data is illustrated.

     Example 2.9. Find the mode of the following qualitative data

     red, green, blue, red, blue, green, red, brown.

     Solution: It can be seen that the most common color is red. Hence, the mode of these data is **red**.

   - **Frequented data:**

     Example 2.10. Find the mode for the following data

Value x_i	19	20	22	24	18	25
f_i	5	7	4	8	4	5

     Solution:

     From the table of data, it can be seen that the most frequented value of data is 24. Hence, $M_0 = 24$.

2. **Grouped Data:**

   In this type of data, by using the frequency of the class intervals we determine the **mode class** and based on this class we determine the values of the terms in the following formula:

$$M_0 = LB + \frac{d_1}{d_1 + d_2} \times l,$$

where
LB is the lower bound of the mode class
d_1 is the difference between the frequencies of the mode class and its precedent class, respectively
d_2 is the difference between the frequencies of the mode class and its following class, respectively
l is the class length.

Example 2.11. Find the mode of the following data

Class	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$
f_i	5	7	4	8	4	5

Solution:

From the table of data, it can be seen that the mode class is (40–49) because it has the largest frequency. Therefore, $LB = 40$, $d_1 = 8 - 4 = 4$, $d_2 = 8 - 4 = 4$ and $l = 10$. Hence,

$$\begin{aligned} M_0 =& LB + \frac{d_1}{d_1 + d_2} \times l = 40 + \frac{4}{4+4} \times 10 \\ =& 40 + \frac{4}{8} \times 10 = 40 + \frac{40}{8} \\ =& 40 + 5 = 45 \end{aligned}$$

2.4 Measures of Location (Quartiles, Deciles and Percentiles)

The locating measures are measures depending only on the location of the values between the data after ordering. In this types of measures only we need to determine the location of the value between the ordering data and then choose it from the data (for the un-grouped data) or calculate it by some formulas (for the grouped data).

2.4.1 Quartiles, Deciles and Percentiles

Let n be the number of data. Then

- The location of quartiles are given by

$$L_{Q_j} = \frac{j(n+1)}{4}, \qquad j = 1, 2, 3.$$

This means that the location of the first quartile is $L_{Q_1} = \dfrac{n+1}{4}$ and for the third quartile is $L_{Q_3} = \dfrac{3(n+1)}{4}$ (note that the second quartile is the median it self $Q_2 = M_e$).

- The location of deciles are given by

$$L_{D_j} = \frac{j(n+1)}{10},$$

(note that the fifth decile is the median it self $D_5 = M_e$).

- The location of percentiles are given by

$$L_{P_j} = \frac{j(n+1)}{100},$$

(note that the fiftieth percentile is the median it self $P_{50} = M_e$).

1. **Un-Grouped Data:**

   - **Raw data:**

     Example 2.12. Find Q_1, Q_3, D_2, D_8, and P_{60} for the following data

$$x_i = 15, 18, 21, 14, 25, 17, 11, 20, 10, 12, 23.$$

Solution: First, we need to order the data as follows

$$x_i = 10, 11, 12, 14, 15, 17, 18, 20, 21, 23, 25$$

Since the number of data $n = 11$, then

– the locations of Q_1 and Q_3 are given, respectively, by

$$L_{Q_1} = \frac{n+1}{4} = \frac{11+1}{4} = 3$$

$$L_{Q_3} = \frac{3(n+1)}{4} = \frac{3(11+1)}{4} = \frac{36}{4} = 9$$

Hence, $Q_1 = 12$ and $Q_3 = 21$.

– the locations of D_2 and D_8 are given, respectively, by

$$L_{D_2} = \frac{2(n+1)}{10} = \frac{2(11+1)}{10} = \frac{24}{10} = 2.4$$

$$L_{D_8} = \frac{8(n+1)}{10} = \frac{8(11+1)}{10} = \frac{96}{10} = 9.6$$

Hence,

$$D_2 = 11 + 0.4(12 - 11) = 11 + 0.4 = 11.4$$

and

$$D_8 = 21 + 0.6(23 - 21) = 21 + 1.2 = 22.2$$

– the location of P_{60} is given by

$$L_{P_{60}} = \frac{60(n+1)}{100} = \frac{6(11+1)}{10} = \frac{72}{10} = 7.2$$

Hence,

$$P_{60} = 18 + 0.2(20-18) = 18 + 0.4 = 18.4$$

- **Frequented data:**

Example 2.13. Find Q_1, Q_3, D_4 and P_{80} for the following data

Value x_i	19	20	22	24	18	25
f_i	5	7	4	8	4	5

Solution:

First we need to order the data in the table and add the increasing cumulative frequency as follows:

Value x_i	18	19	20	22	24	25
f_i	4	5	7	4	8	5
$CF_i\uparrow$	4	9	16	20	28	33

Since the number of data $n = 33$, then

– the locations of Q_1 and Q_3 are given, respectively, by

$$L_{Q_1} = \frac{n+1}{4} = \frac{33+1}{4} = 8.5$$

$$L_{Q_3} = \frac{3(n+1)}{4} = \frac{3(33+1)}{4} = \frac{102}{4} = 25.5$$

Hence, $Q_1 = 19$ and $Q_3 = 24$.

– the location of D_4 is given by

$$L_{D_4} = \frac{4(n+1)}{10} = \frac{4(33+1)}{10} = \frac{136}{10} = 13.6$$

Hence, $D_4 = 20$.

– the location of P_{60} is given by

$$L_{P_{80}} = \frac{80(n+1)}{100} = \frac{8(33+1)}{10} = \frac{272}{10} = 27.2$$

Hence, $P_{80} = 24$.

2. **Grouped data:**

In grouped data the quartiles, deciles, and percentiles are given by the following formulas:

- **Quartiles**:

$$Q_j = LB + \frac{\frac{j\sum f_i}{4} - F_i}{f_i} \times l,$$

where
LB is the lower bound of Q_j class
F_i is the increasing cumulative frequency of the class before the Q_j class
f_i is the frequency of the Q_j class
l is the class length.

- **Deciles**:

$$D_j = LB + \frac{\frac{j\sum f_i}{10} - F_i}{f_i} \times l,$$

where
LB is the lower bound of D_j class
F_i is the increasing cumulative frequency of the class before the D_j class
f_i is the frequency of the D_j class
l is the class length.

- **Percentiles**:

$$P_j = LB + \frac{\frac{j\sum f_i}{100} - F_i}{f_i} \times l,$$

where
LB is the lower bound of P_j class
F_i is the increasing cumulative frequency of the class before the P_j class
f_i is the frequency of the P_j class
l is the class length.

Example 2.14. Find Q_1, Q_3, D_6 and P_{40} of the following data

Class	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$
f_i	5	7	4	8	4	5

Solution:

We need to determine the increasing cumulative frequency as follows

Class	f_i	$CF_i \uparrow$
10 – 19	5	5
20 – 29	7	12
30 – 39	4	16
40 – 49	8	24
50 – 59	4	28
60 – 69	5	33

The number of data is $n = 33$, then

- for Q_1, the location of Q_1 is
$$L_{Q_1} = \frac{n+1}{4} = \frac{33+1}{4} = 8.5$$
Therefore, the Q_1 class is $(20 - 29)$. Thus, $LB = 20$, $F_i = 5$, $f_i = 7$ and $l = 10$. Hence,
$$\begin{aligned} Q_1 &= LB + \frac{\frac{\sum f_i}{4} - F_i}{f_i} \times l \\ &= 20 + \frac{\frac{33}{4} - 5}{7} \times 10 = 20 + \frac{8.25 - 5}{7} \times 10 \\ &= 20 + \frac{32.5}{7} = 20 + 4.64 = 24.64. \end{aligned}$$

- for Q_3, the location of Q_3 is
$$L_{Q_3} = \frac{3(n+1)}{4} = \frac{3(33+1)}{4} = 25.5$$
Therefore, the Q_3 class is $(50 - 59)$. Thus, $LB = 50$, $F_i = 24$, $f_i = 4$ and $l = 10$. Hence,
$$\begin{aligned} Q_3 &= LB + \frac{\frac{3\sum f_i}{4} - F_i}{f_i} \times l \\ &= 50 + \frac{\frac{3\times 33}{4} - 24}{4} \times 10 = 50 + \frac{24.75 - 24}{4} \times 10 \\ &= 50 + \frac{7.5}{4} = 50 + 1.875 = 51.875. \end{aligned}$$

- for D_6, the location of D_6 is

$$L_{D_6} = \frac{6(n+1)}{10} = \frac{6(33+1)}{10} = 20.4$$

Therefore, the D_6 class is $(40-49)$. Thus, $LB = 40$, $F_i = 16$, $f_i = 8$ and $l = 10$. Hence,

$$\begin{aligned} D_6 &= LB + \frac{\frac{6\sum f_i}{10} - F_i}{f_i} \times l \\ &= 40 + \frac{\frac{6\times 33}{10} - 16}{8} \times 10 = 40 + \frac{19.8-16}{8} \times 10 \\ &= 40 + \frac{38}{8} = 40 + 4.75 = 44.75. \end{aligned}$$

- for P_{40}, the location of P_{40} is

$$L_{P_{40}} = \frac{40(n+1)}{100} = \frac{40(33+1)}{100} = 13.6$$

Therefore, the P_{40} class is $(30-39)$. Thus, $LB = 30$, $F_i = 12$, $f_i = 4$ and $l = 10$. Hence,

$$\begin{aligned} P_{40} &= LB + \frac{\frac{40\sum f_i}{100} - F_i}{f_i} \times l \\ &= 30 + \frac{\frac{40\times 33}{100} - 12}{4} \times 10 = 30 + \frac{13.2-12}{4} \times 10 \\ &= 30 + \frac{12}{4} = 30 + 3 = 33. \end{aligned}$$

Chapter 3

Measures of Variation

The measures of variation shows how the data are spread or scattered around the mean. Many types of measures of variation are defined such as the **Range, Quartile Deviation, Mean Deviation, Variance and the Standard Deviation**. In this section, the Mean Deviation, Variance and the Standard Deviation are presented.

Note That: Whenever the values of measures of variation close to zero this means that the data are close to their mean.

3.1 Mean Deviation MD

The Mean deviation (denoted by MD) is defined as the sum of the absolute values of the difference between the data and their arithmetic mean divided by the number of data, namely

$$MD = \frac{\sum |x_i - \bar{x}|}{n},$$

where n is the number of data.

1. **Un-Grouped Data:**

   - **Raw data:**

     Example 3.1. Find the mean deviation for the following data

     $$x_i = 11, 12, 13, 14, 15, 17.$$

     Solution: First, we need to find the mean of the data:

     $$\bar{x} = \frac{\sum x_i}{n} = \frac{11 + 12 + 13 + 14 + 15 + 17}{6} = 13.6$$

Now, create the following table to find the deviation between the data and their mean.

x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$
11	-2.6	2.6
12	-1.6	1.6
13	-0.6	0.6
14	0.4	0.4
15	1.4	1.4
17	3.4	3.4
$\sum$		10

Therefore, the mean deviation is given by

$$MD = \frac{\sum |x_i - \bar{x}|}{n} = \frac{10}{6} = 1.6$$

- **Frequented data:**

In this type of data the mean deviation is given by

$$MD = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i}$$

Example 3.2. Find the mean deviation for the following data

Value x_i	18	19	20	22	24	25
f_i	4	5	7	4	8	5

Solution:

First, we find the mean of the data:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{712}{33} = 21.6.$$

Value x_i	18	19	20	22	24	25	$\sum$
f_i	4	5	7	4	8	5	33
$x_i f_i$	72	95	140	88	192	125	712
$\lvert x_i - \bar{x} \rvert$	3.6	2.6	1.6	0.4	2.4	3.4	
$f_i \lvert x_i - \bar{x} \rvert$	14.4	13	11.2	1.6	19.2	17	76.4

Therefore, the mean deviation is given by

$$MD = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} = \frac{76.4}{33} = 2.31$$

2. **Grouped Data:**

In this type of data the mean deviation is given by

$$MD = \frac{\sum f_i |y_i - \bar{y}|}{\sum f_i},$$

where y_i are the midpoints of the class intervals and $\bar{y}$ is the mean.

Example 3.3. Find the mean deviation of the following data

Class	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69
f_i	5	7	4	8	4	5

Solution:

First, we find the mean of the data, then we calculate the absolute difference between the mean and the midpoints of the classes, see the following table:

Class	f_i	y_i	$y_i f_i$	$\lvert y_i - \bar{y}\rvert$	$f_i\lvert y_i - \bar{y}\rvert$
10 − 19	5	14.5	72.5	24.24	121.2
20 − 29	7	24.5	171.5	14.24	99.68
30 − 39	4	34.5	138	4.24	16.96
40 − 49	8	44.5	356	5.76	46.08
50 − 59	4	54.5	218	15.76	63.04
60 − 69	5	64.5	322.5	25.76	128.8
$\sum$	33		1278.5		475.16

Thus,

$$\bar{y} = \frac{\sum y_i f_i}{\sum f_i} = \frac{1278.5}{33} = 38.74$$

Hence,

$$MD = \frac{\sum f_i |y_i - \bar{y}|}{\sum f_i} = \frac{475.16}{33} = 14.4$$

3.2 Variance and Standard Deviation

The variance is defined as the mean of the squares of the deviations of data from the arithmetic mean. Actually in general, if the variance is computed for the data of population, then it is denoted by σ^2 and defined by

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N},$$

where N is the number of data in the population.

On the other hand, if the variance is computed for the elements of the sample, then it is denoted by s^2 and defined by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1},$$

where n is the size of the sample.

For simplicity, the variance in our study here is defined as

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n},$$

with skipping subtracting 1 from the number of data.

The standard deviation is defined as the square root of the variance of the data, namely:

$$s = \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

1. **Un-Grouped Data:**

- **Raw data:**

Example 3.4. Find the variance and the standard deviation of the following data

$$x_i = 11, 12, 13, 14, 15, 17.$$

Solution: First, we need to find the mean of the data:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{11 + 12 + 13 + 14 + 15 + 17}{6} = 13.6$$

Now, create the following table to find the deviation between the data and their mean.

x_i	11	12	13	14	15	17	$\sum$
$x_i - \bar{x}$	-2.6	-1.6	-0.6	0.4	1.4	3.4	
$(x_i - \bar{x})^2$	6.76	2.56	0.36	0.16	1.96	11.56	23.36

Therefore, the variance is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{23.36}{6} = 3.89,$$

and hence, the standard deviation is given by $s = \sqrt{3.89} = 1.97$.

- **Frequented data:**

In this type of data the variance is given by

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$$

Example 3.5. Find the variance and the standard deviation of the following data

Value x_i	18	19	20	22	24	25
f_i	4	5	7	4	8	5

Solution:

First, we find the mean of the data:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{712}{33} = 21.6.$$

Value x_i	18	19	20	22	24	25	$\sum$
f_i	4	5	7	4	8	5	33
$x_i f_i$	72	95	140	88	192	125	712
$x_i - \bar{x}$	−3.6	−2.6	−1.6	0.4	2.4	3.4	
$(x_i - \bar{x})^2$	12.96	6.76	2.56	0.16	5.76	11.56	
$f_i(x_i - \bar{x})^2$	51.84	33.8	17.92	0.64	46.08	57.8	208.08

Therefore, the variance is given by

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{208.08}{33} = 6.30.$$

Hence, the standard deviation is given by $s = \sqrt{6.30} = 2.51$.

2. **Grouped Data:**

In this type of data the variance is given by

$$s^2 = \frac{\sum f_i (y_i - \bar{y})^2}{\sum f_i},$$

where y_i are the midpoints of the class intervals and $\bar{y}$ is the mean.

Example 3.6. Find the variance and the standard deviation of the following data

Class	10 − 19	20 − 29	30 − 39	40 − 49	50 − 59	60 − 69
f_i	5	7	4	8	4	5

Solution:

First, we find the mean of the data, then we calculate the difference between the mean and the midpoints of the classes, next compute the squares of the deviations , see the following table:

Class	f_i	y_i	$y_i f_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$f_i(y_i - \bar{y})^2$
10 − 19	5	14.5	72.5	−24.24	587.57	2937.85
20 − 29	7	24.5	171.5	−14.24	202.77	1419.39
30 − 39	4	34.5	138	−4.24	17.97	71.88
40 − 49	8	44.5	356	5.76	33.17	265.36
50 − 59	4	54.5	218	15.76	248.37	993.48
60 − 69	5	64.5	322.5	25.76	663.57	3317.85
$\sum$	33		1278.5			9005.81

Thus,

$$\bar{y} = \frac{\sum y_i f_i}{\sum f_i} = \frac{1278.5}{33} = 38.74$$

Therefore,

$$s^2 = \frac{\sum f_i (y_i - \bar{y})^2}{\sum f_i} = \frac{9005.81}{33} = 272.9.$$

Hence, the standard deviation is given by $s = \sqrt{272.9} = 16.519.$

3.2.1 Shortest Method to Calculate Standard Deviation

We know that

$$\begin{aligned}
s^2 =& \frac{\sum (x_i - \bar{x})^2}{n} = \frac{1}{n} \sum \left(x_i^2 - 2x_i\bar{x} + (\bar{x})^2 \right) \\
=& \frac{1}{n} \left[\sum x_i^2 - 2\bar{x} \sum x_i + \sum (\bar{x})^2 \right] \\
=& \left[\frac{\sum x_i^2}{n} - 2\bar{x} \frac{\sum x_i}{n} + \frac{n(\bar{x})^2}{n} \right] = \frac{\sum x_i^2}{n} - 2(\bar{x})^2 + (\bar{x})^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2 \\
=& \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n} \right)^2.
\end{aligned}$$

Therefore, the standard deviation is given by

$$s = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

Also for the grouped data:

$$s^2 = \frac{\sum f_i y_i^2}{\sum f_i} - \left(\frac{\sum f_i y_i}{\sum f_i}\right)^2,$$

and hence,

$$s = \sqrt{\frac{\sum f_i y_i^2}{\sum f_i} - \left(\frac{\sum f_i y_i}{\sum f_i}\right)^2},$$

where y_i are the midpoints of the class intervals.

Example 3.7. Resolve Example 3.6 by using the shortest method of calculating the standard deviation.

Solution:

To calculate the variance and the standard deviation for the data in Example 3.6, we can create the following table:

Class	f_i	y_i	$y_i f_i$	y_i^2	$f_i y_i^2$
10 – 19	5	14.5	72.5	210.25	1051.25
20 – 29	7	24.5	171.5	600.25	4201.75
30 – 39	4	34.5	138	1190.25	4761
40 – 49	8	44.5	356	1980.25	15842
50 – 59	4	54.5	218	2970.25	11881
60 – 69	5	64.5	322.5	4160.25	20801.25
$\sum$	33		1278.5		58538.25

Therefore,

$$s^2 = \frac{\sum f_i y_i^2}{\sum f_i} - \left(\frac{\sum f_i y_i}{\sum f_i}\right)^2 = \frac{58538.25}{33} - \left(\frac{1278.5}{33}\right)^2$$
$$= 1773.88 - (38.74)^2 = 1773.88 - 1500.78 = 273.1$$

Hence, the standard deviation is given as follows:

$$s = \sqrt{273.1} = 16.52$$

3.3 Exercises 2 & 3

1. Find the mean, median and the mode of the following data:

   (i) 60 50 80 70 90 100 50

   (ii) 25 29 32 35 37 41 42 45

2. In Question 1, find Q_1, Q_3, D_6 and P_{30}.

3. In the following table, find the mean, median, mode, Q_1, Q_3, D_6 and P_{40}

Class	17	20	25	27	29
f_i	3	7	6	4	4

4. In Question 1, calculate the mean deviation, variance and the standard deviation.

5. The following table presents the monthly income for a sample of families in a city (by thousands)

Class	62 – 65	66 – 69	70 – 73	74 – 77	78 – 81	82 – 85	86 – 89
f_i	3	8	20	21	14	10	4

   Find:

   (i) The mean and the mode.

   (ii) The median, first and third quartiles.

   (iii) D_9, P_{20}.

6. In Question 5, calculate the mean deviation and the standard deviation.

7. Calculate the standard deviation of the data in the following table by using the shortest method of finding the standard deviation:

Class	15 – 17	18 – 20	21 – 23	24 – 26	27 – 29
f_i	3	7	21	4	4

Chapter 4

Correlation and Regression

In this section we will study how to describe the relationship between two variables X and Y (if the study consists of two variables or more).

4.1 Correlation

A correlation is a relationship between two variables X and Y, this relationship may be **Linear** or **Non-Linear**. The correlation between two variables X and Y can be described as either strong or weak, and as either positive or negative. The data of the two variables can be represented as ordered pairs (x, y), where x is the independent variable and y is the dependent variable.

Types of Correlation

1. **Positive Linear Correlation**: There is a positive linear correlation when the variable on the x-axis increases as the variable on the y-axis increases. This is shown by an upwards sloping straight line.

2. **Negative Linear Correlation**: There is a negative linear correlation when one variable increases as the other variable decreases. This is shown by a downwards sloping straight line.

3. **Non-linear Correlation**: There is a non-linear correlation when there is a relationship between variables but the relationship is not linear (straight).

4. **No Correlation**: There is no correlation when there is no pattern that can be detected between the variables.

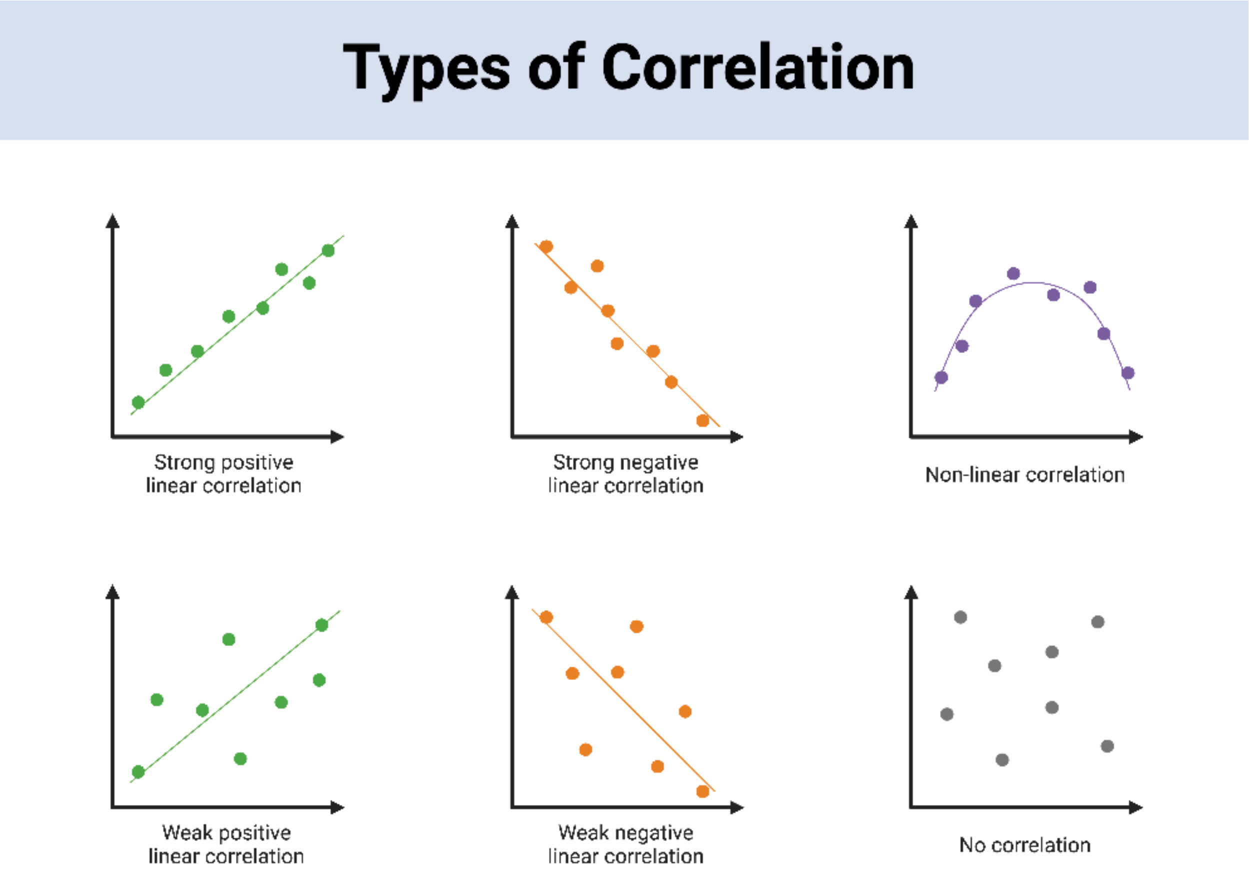

Figure 4.1: Scattered diagrams

In this section we study the **Linear correlation** which described by relative measures represent the strength and the direction of a relationship between two variables these measures are called the correlation coefficients.

4.1.1 Correlation Coefficients

The correlation coefficient is a measure of the strength and the direction of a linear relationship between two variables X and Y. The range of the correlation coefficient is between -1 and 1 described as:

(i) If the value of the correlation coefficient is close to 1 (or -1), then the variables X and Y have a strong positive (or negative) linear correlation, respectively.

(ii) If there is no linear correlation or a weak linear correlation between X and Y, this implies that the value of the correlation coefficient is close to 0.

Two types of linear coefficients of correlation are presented in this section which are Pearson's coefficient of correlation and Spearman's coefficient of correlation.

Pearson's Coefficient of Correlation

This coefficient of correlation is used only for quantitative data, denoted by r and defined by the following formula:

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Example 4.1. Calculate the Pearson's coefficient of correlation for the following data and describe the strength and direction of the relationship between the variables X and Y.

X	1	2	3	4	5
Y	-3	-1	0	1	2

Solution: Create the following table, where number of data $n = 5$:

X	Y	XY	X^2	Y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum$	$\sum$	$\sum$	$\sum$	$\sum$
15	-1	9	55	15

Therefore,

$$\begin{aligned} r =& \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \\ =& \frac{5(9) - (15)(-1)}{\sqrt{5(55) - (15)^2}\,\sqrt{5(15) - (-1)^2}} = \frac{45 + 15}{\sqrt{275 - 225}\,\sqrt{75 - 1}} \\ =& \frac{60}{\sqrt{50}\,\sqrt{74}} = 0.986. \end{aligned}$$

Hence, there is a positive strong linear correlation between X and Y.

Spearman's Coefficient of Correlation

This type of correlation depending on the rank of data so it is called Spearman's rank correlation coefficient. This correlation coefficient is a measure of the strength and direction for the ordinal data. The formula of the Spearman's coefficient of correlation is given as:

$$r = 1 - \frac{6\sum D^2}{n(n^2-1)},$$

where $D = rank(x) - rank(y)$ and n is the number of data.

To calculate the Spearman's correlation coefficient, follow the following steps:

(i) Assign a rank for each value of data for X and Y (if they are not given).

(ii) Calculate the difference D of ranks of X and ranks of Y and make them in a separate column.

(iii) Square the differences D and make them in a separate column.

(iv) Apply the formula to get the rank correlation.

Example 4.2. Calculate the Spearman's coefficient of correlation for the following data and describe the strength and direction of the relationship between the variables X and Y.

X	excellent	good	v.good	passable	v.good	weak
Y	excellent	v.good	passable	v.good	weak	excellent

Solution: Suppose the ranks of these grades are

Data	weak	passable	good	v.good	excellent
Rank	1	2	3	4	5

Thus, we can make the table of data as follows:

X	excellent	good	v.good	passable	v.good	weak	$\sum$
Y	excellent	v.good	passable	v.good	weak	excellent	
Rank (X)	5	3	4	2	4	1	
Rank (Y)	5	4	2	4	1	5	
D	0	-1	2	-2	3	-4	
D^2	0	1	4	4	9	16	34

Therefore,

$$r = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(34)}{6(36-1)} = 0.028.$$

Hence, there is no correlation or a very weak positive linear correlation between X and Y.

4.2 Regression

Regression analysis is the method used for estimating the unknown values of one variable corresponding to the known value of another variable.

If the scatter diagram indicates some relationship between two variables x and y, then the dots of the scatter diagram concentrated round a curve. This curve is called the **curve of regression**. In this section, we will focus on the **Linear Regression** as follows.

4.2.1 Linear Regression

When the curve of regression is a straight line, it is called a line of regression. A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

Equations of Linear Regression

There exists two types of line regression equations which are:

(i) Equation of line of regression of y on x, which defined as $y = bx + a$, where

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \ ,$$

and $a = \overline{y} - b\overline{x}$.

(ii) Equation of line of regression of x on y, which defined as $x = by + a$, where

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2} \quad ,$$

and $a = \overline{x} - b\overline{y}$.

Example 4.3. Find the equations of the line of regression (of y on x and x on y) for the following data.

X	1	2	3	4	5
Y	-3	-1	0	1	2

Solution: Create the following table, where number of data $n = 5$:

X	Y	XY	X^2	Y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum$	$\sum$	$\sum$	$\sum$	$\sum$
15	-1	9	55	15

First, we will find $\overline{y}$ and $\overline{x}$ as follows:

$$\overline{y} = \frac{\sum y_i}{n} = \frac{-1}{5} = -0.2\,,$$

$$\overline{x} = \frac{\sum x_i}{n} = \frac{15}{5} = 3.$$

Therefore,

(i) For the regression line of y on x, $\big(y = bx + a\big)$, we have

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{(5)(9) - (15)(-1)}{(5)(55) - (15)^2} = \frac{45 + 15}{275 - 225} = \frac{60}{50} = 1.2$$

and

$$a = \overline{y} - b\overline{x} = -0.2 - (1.2)(3) = -3.8$$

Hence, the equation of regression line of y on x is given by

$$y = 1.2x - 3.8$$

(ii) For the regression line of x on y, $\big(x = by + a\big)$, we have

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2} = \frac{(5)(9) - (15)(-1)}{(5)(15) - (-1)^2} = \frac{45 + 15}{75 - 1} = \frac{60}{74} = 0.81$$

and

$$a = \overline{x} - b\overline{y} = 3 - (0.81)(-0.2) = 3.162$$

Hence, the equation of regression line of x on y is given by

$$x = 0.81y + 3.162$$

4.3 Exercises 4

1. Calculate the Pearson's coefficient of correlation and find the regressions lines of the data in the following table

X	4	6	8	10	12
Y	2	3	4	6	10

2. Find the coefficient of correlation and obtain the regressions lines of the data in the following table

X	1	2	3	4	5
Y	2	5	3	8	7

3. Find the coefficient of correlation then determine the regressions lines of the data in the following table

X	6	2	10	4	8
Y	9	11	5	8	7