Statistics

Contents:

Chapter (1): Introduction to Statistics

Chapter (2): Measures of Central Tendency

Chapter (3): Measures of dispersion

Chapter (4): Correlation and Simple Regression

Chapter (5): Hypothesis Testing

CHAPTER (1)

Introduction to Statistics

Objectives of teaching statistics
Descriptive and inferential statistics
Type of the data
Histograms
Bar graphs
Frequency polygons
Pie

The Objectives of teaching statistics are:

لتشجيع وتطوير التفكير النقدي لدى الطلاب في التعامل مع بيانات العالم الحقيقي

1 To encourage and develop criticall thinking in students in handling real world data preferably of local origin.

جعل الطلاب يتعلمون من خلال إجراء التجارب وجمع البيانات ووصفها

2 To make the students learn by conducting experiments, collecting and describing data.

لتعزيز فهم الموضوع لغرض تحليل البيانات واستخلاص استنتاجات صحيحة.

3 To promote understanding of the subject for the purpose of analyzing data and drawing valid inferences.

لتزويد الطلاب بخلفية أساسية سليمة تمكنهم من متابعة الدراسات في الإحصاء على مستويات أعلى.

4 To provide the students sound basic background which would enable them to pursue studies in statics at higher levels.

إعداد الطلاب لتولي الوظائف الثابتة في مختلف المؤسسات الحكومية وشبه الحكومية والخاصة

5 To prepare students for taking up statical jobs in various government/sémi-government/private organization.

تعريف الطلاب بأدوات أنجوس الحالية مثل أجهزة الكمبيوتر والحزم

6 To expose students with present day tools of angus i.e. computers and packages.

Statistics: Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. It also involves model building. Suppose we wish to study household incomes in a certain neighborhood. We may decide to randomly select, say, 50 families end examine their household incomes. [illegible]. When we consider this [illegible] examples, we note that in the first case the population (the household incomes of all families in the neighborhood) really exists, [illegible]

لنفترض أننا نرغب في دراسة دخل الأسرة في حي معين

In either case we can
visualize the totality of the population values, of which our sample data are only a small part. Thus we define a population to be the set of all measurements or objects that are of interest and a sample to be a subset of that population. The population acts as the sampling frame from which a sample is selected. Now we introduce some basic notions commonly used in statistics.

- **Population:** A set of units (people, objects, transactions, events) that we are interested in studying (A population is the collection or set of all objects or measurements that are of interest to the collector). Suppose we want to study the salary of Sohag people, our population includes all those persons who work in Sohag. However as the population is so big that it is not practical and economical to collect salary data of all the working people, we always select randomly only a subset of the population and the data is sample.

- **Sample:** The sample is a subset of data selected from a population. The size of a sample is the number of elements in it.

- **Descriptive and Inferential statistics:**

The methods consisting mainly of organizing, summarizing, and presenting data in the form of tables, graphs, and charts are called ***descriptive statistics***. The methods of drawing inferences and making decisions about the population using the sample are called inferential statistics. Inferential statistics uses probability theory.

- ***A statistical inference*** is an estimate, a prediction, a decision, or a generalization about the population based on information contained in a sample. There are two main approaches:
  - Confidence intervals, where we estimate and specify our degree of certainty,
  - Hypothesis testing, where we evaluate a claim using relevant data.

- **Types of Data**

Quantitative data are observations measured on a numerical scale. The number of car accidents in different Egypt cities is ***quantitative data***.

Non numerical data that can only be classified into one of the groups of categories are said to be ***qualitative*** or ***categorical data***. The blood group of each person in a community as O, A, B, AB is qualitative data.

(٢-١-٤) **مركز الفترة وطول الفترة:**

نعرف مركز الفترة وطول الفترة كما يلي:

- **Class:** a category into which data can be classified. Generally the classes will be intervals of equal length. The center of each class is called a class mark. The end points of each class interval are called class boundaries. Usually, there are two ways of

choosing class boundaries. One way is to choose non overlapping class boundaries so that none of the data points will simultaneously fall in two classes. Another way is that for each class, except the last, the upper boundary is equal to the lower boundary of the subsequent class.

- **Lower Class Limit:** The least value that can belong to a class.
- **Upper Class Limit:** The greatest value that can belong to a class.
- **Class Width:** The difference between the upper (or lower) class limits.
- **Class Midpoint:** The middle value of each data class. To find the class midpoint, average the upper and lower class limits x_i , $x_i = \frac{upper + lower}{2}$.
- **Class Boundaries:** The numbers that separate classes without forming gaps between them.

٢. تنظيم البيانات وعرضها

- **Range of Class:** The highest value – the lowest value.
- **A frequency table:** is a table that divides a data set into a suitable number of categories (classes). A frequency table is created by choosing a specific number of classes in which the data will be placed. Once the data are summarized in the form of a frequency table, a graphical representation can be given through ***bar graphs, pie charts***, and **histograms.**

- **Grouped data:** Data presented in the form of a frequency table are called grouped data.
- **The cumulative frequency:** Cumulative means the total of all frequencies. Cumulative totals can be used to determine how many scores are above or below a set level (It can be ascending or descending).

- **Relative frequency:** Is the percentage of data elements in that class. Let f_i denote the frequency of the class i and let n be sum of all frequencies. Then the relative frequency for the class i is defined as the ratio $\frac{f_i}{n}$.

- **Cumulative relative frequency:** The cumulative relative frequency (see the following table) for the class i is defined by $\sum \frac{f_i}{n} = 1$.

Class Boundaries	Frequency	Ascending cumulative frequency	Descending cumulative frequency
18.5 – 22.5	7	7	35
22.5 – 26.5	9	16	28
26.5 – 30.5	6	22	19
30.5 – 34.5	10	32	13
34.5 – 38.5	3	35	3

- **Construction of a frequency table**
  - Determine the maximum and minimum values of the observations.
  - The range, R = maximum value − minimum value.
  - The class width should be slightly larger than the ratio $\frac{R}{Number\ of\ sets}$.
  - The first interval should begin a little below the minimum value, and the last interval should end a little above the maximum value. The intervals are called class intervals and the boundaries are called class boundaries. The class limits are the smallest and the largest data values in the class. The class mark is the midpoint of a class.

None of the data values should fall on the boundaries of the classes.

Construct a table (frequency table) that lists the class intervals, a tabulation of the number of measurements in each class (tally), the frequency f_i of each class, and, if needed, a column with relative frequency, $\frac{f_i}{n}$, where n is the total number of observations.

- **Histograms:** A histogram is a graphical representation of the information in a frequency table using a bar graph with sides touching (Like a bar graph, except the data is continuous, so bars touch).

- **Frequency Polygon:** Connect the midpoints of each class to make a polygon (A frequency polygon is a line graph representation of the information in a frequency table).

- **Examples**

Example: 1

The data represent the statistics marks for 50 students. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Solution:

Step 1: Determine the classes

- Find the **highest value** and the **lowest value** and use them to find the **range**.

Range = highest value in the data set – lowest value=134-100=34

- Find the **class width** by dividing the range by the number of classes. Round the answer up to the next whole number if there is a remainder. The class width is the difference between the lower class limit of one class and the lower class limit of the next class.

Class width=Range/number of sets=34/7 ~ 5

- Use your lowest value as your starting point. Add the class width to the starting point to get the lower limit for the next class. Keep adding until there are 7 classes. Subtract 1 from the lower limit of the second class to get the upper limit of the first class.

- **Step 2: Find the class boundaries**
- Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit. In future calculations we will also need to find the midpoints x_i of each class.

- **Step 3:** Tally the data and find the numerical frequencies from the tallies.
- **Step 4:** Find the cumulative frequencies: The cumulative frequency for a class is the sum of the frequencies for that class and all previous classes. To find this value, add up all the frequencies that lead up to each class.

Now let us construct our frequency distribution:

Class Limits	Class Boundaries	Frequency (f)	Cumulative Frequency	Midpoints (x_i)
100-104	99.5-104.5	2	2	102
105-109	104.5-109.5	8	10	107
110-114	109.5-114.5	18	28	112
115-119	114.5-119.5	13	41	117
120-124	119.5-124.5	7	48	122
125-129	124.5-129.5	1	49	127
130-134	129.5-134.5	1	50	132
Sum		50		

Example: 2

If the following data (life of laptop computer batteries) are available

130, 145, 126, 146,164, 130, 132, 152, 145, 129, 133, 155, 140, 127, 139, 137, 131, 126, 145, 148, 125, 132, 126, 126, 126, 135, 131, 129, 147, 136, 129, 136, 156, 146, 130, 146, 132, 142, 132, 132.

a. Construct a frequency distribution and a histogram.

b. Construct a relative frequency distribution and cumulative relative frequency plot.

Solution:

Minimum point =125

Maximum point = 164

Range = 164 – 125 = 39.

Number of data points n = 40.

Number of sets close to $\sqrt{n}$ ~7

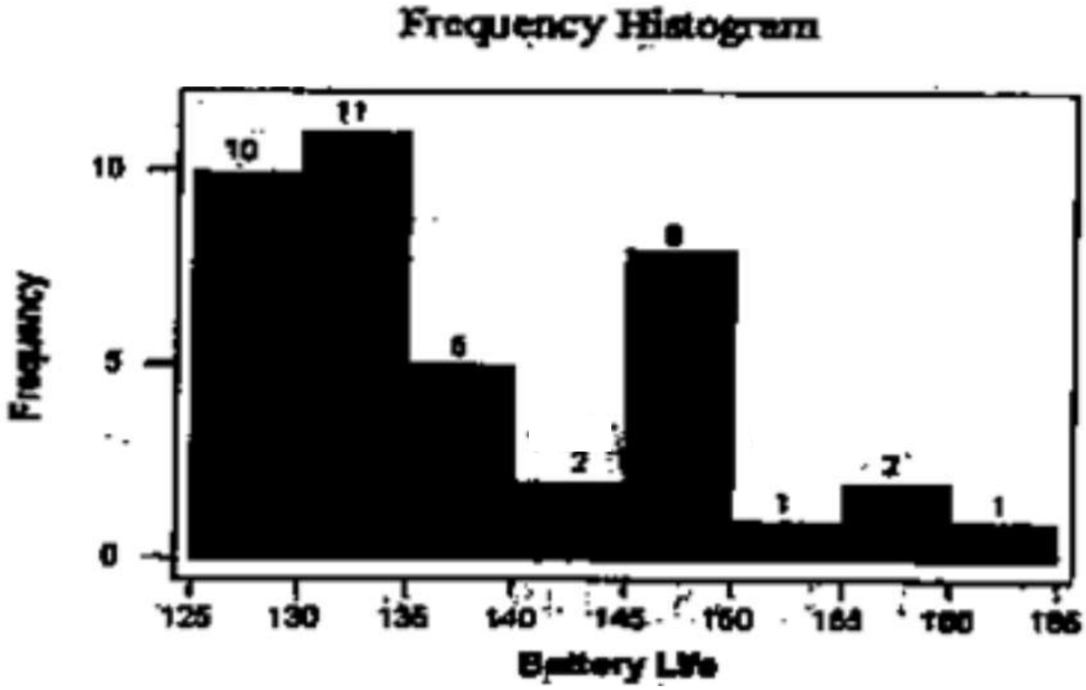
The class width (L) may be determined as L = Range/7=39/7=5.57~6.

Class Interval	Tally	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
125-129	~~IIII~~ ~~IIII~~	10	10	0.25	0.250
130-134	~~IIII~~ ~~IIII~~ I	11	21	0.275	0.525
135-139	~~IIII~~	5	26	0.125	0.650
140-144	II	2	28	0.05	0.700
145-149	~~IIII~~ III	8	36	0.20	0.900
150-154	I	1	37	0.025	0.925

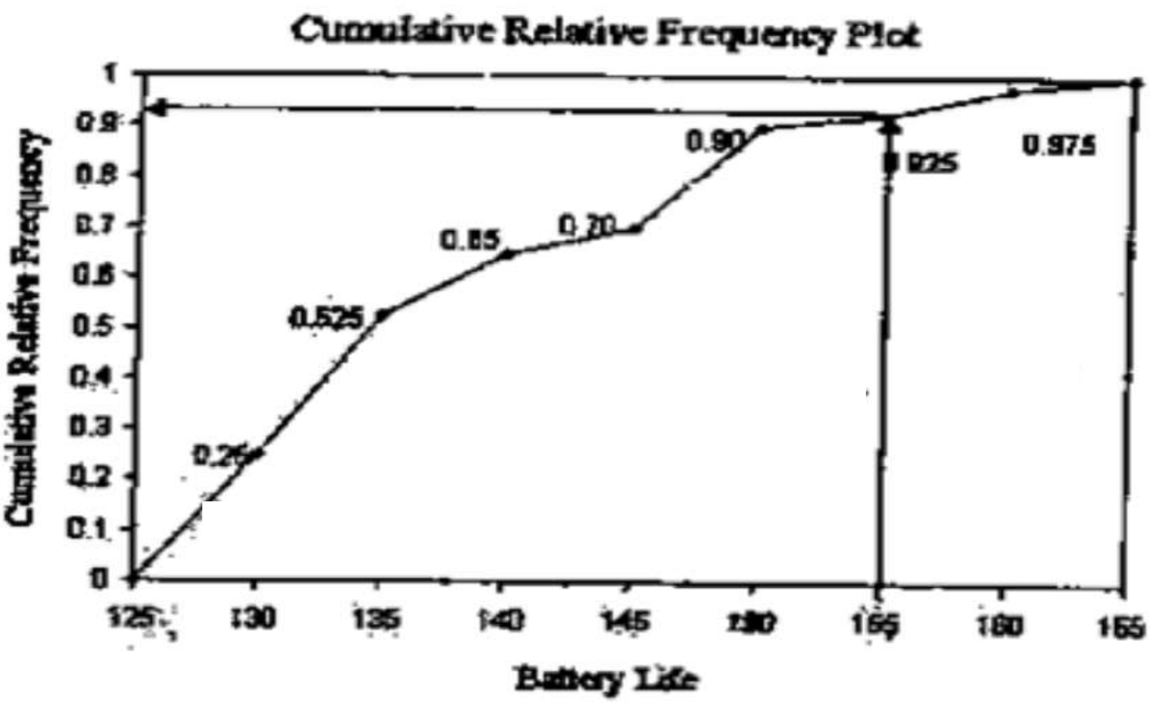
To simplify calculations, we may increase number of sets to 8 and modify L to 5. If we start the first class at 125, its upper bound would be 129, and all other classes are determined accordingly. **One can write sets as:**

125-130, 130-135, 135-140,140-145,145-150, 150-155, 155-160, 160-165.

The following **histogram** is a graphical depiction on the frequencies above. It shows that most of the data are clustered around 135, with few points above 150.



A cumulative relative frequency plot may be used to calculate various probabilities. For example, in the plot below, we see that the probability of a battery life of less than 150 is 0.925.



Example: 3

Suppose that 20 statistics students' scores on an exam are as follows:

97, 92, 88, 75, 83, 67, 89, 55, 72, 78, 81, 91, 57, 63, 67, 74, 87, 84, 98, 46

Construct a frequency table with classes 40-49, 50-59, 60-69 etc.

Solution:

Class	Frequency (f)	Cumulative Frequency	Relative Frequency (f/n)
40-49	1	20	0.20
50-59	2	19	0.30
60-69	3	17	0.20
70-79	4	14	0.15
80-89	6	10	0.10
90-99	4	4	0.05

Note that: the sum of the frequency column is equal to 20, the number of test scores.

Example: 4

Construct a frequency distribution for the data below.

1	2	6	7	12	13	2	6	9	5
18	7	3	15	15	4	17	1	14	5
4	16	4	5	8	6	5	18	5	2

Solution:

Class	Frequency (f)
1-3	6
4-6	11
7-9	4
10-12	1
13-15	4
16-18	4

Example: 5

Construct a frequency distribution for the data below.

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

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

Solution:

Class	Frequency (f)	Percent
A	5	20
B	7	28
O	9	36
AB	4	16
Sum	25	100

Example: 6

Construct a grouped frequency distribution for the following data using 6 classes.

91	78	93	57	75	52	99	80	73	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Solution:

Class Limits	Class Boundaries	Frequency (f)	Cumulative Frequency
50-59	49.5-59.5	2	2
60-69	59.5-69.5	13	15
70-79	69.5-79.5	16	31
80-89	79.5-89.5	7	38
90-99	89.5-99.5	7	45
100-109	99.5-109.5	5	50
Sum		50	

Example: 10

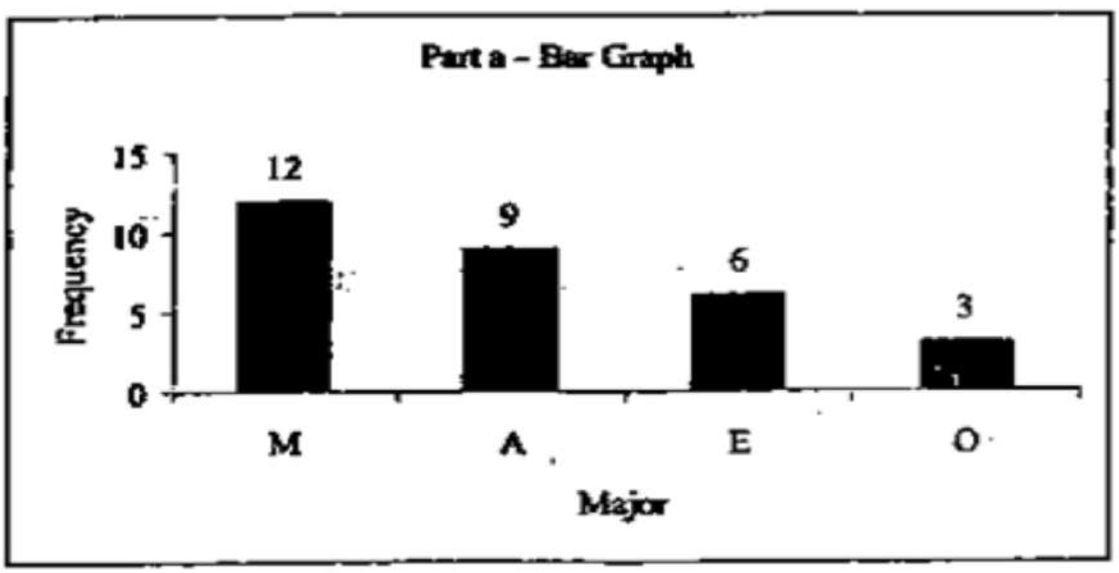
Thirty students in the Sohag faculty of science were asked what their majors were. The following represents their responses (M = Mathematics; A = Analysis;
E = Electronics; O = Others).

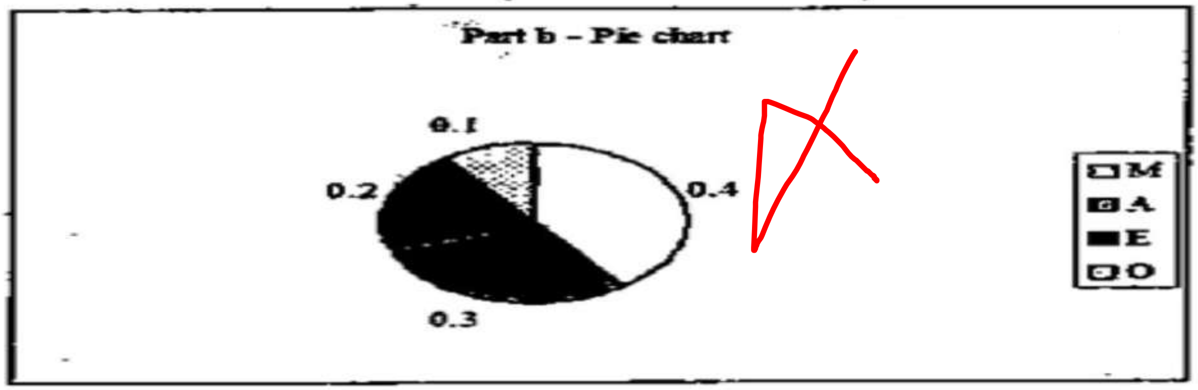
A	M	M	A	M	M	E	M	O	A
E	E	M	A	O	E	M	A	M	A
M	A	O	A	M	E	E	M	A	M

a. Construct a frequency distribution and a bar graph.
b. Construct a relative frequency distribution and a pie chart.

Solution:

Major	Frequency	Relative Frequency
M	12	0.4
A	9	0.3
E	6	0.2
O	3	0.1
Total	30	1.0




Part b – Pie chart
0.1
0.2
0.4
0.3
M
A
E
O

Exercises

1- Find class boundaries and midpoints of the following data:

Class	0–4	5–9	10–14	15–19	20–24
Frequency	5	14	15	10	6

2- Construct ascending cumulative frequency table of the following data:

Class Interval	25-29	30-34	35-39	40-44	45-49	50-54
Frequency	5	8	10	13	8	6

3- Construct descending cumulative frequency table of the following data:

Class Interval	25-29	30-34	35-39	40-44	45-49	50-54
Frequency	5	8	10	13	8	6

4- Find graphical representation of the following data by using frequency curve :

Class Interval	82-89	90-97	98-105	106-113	114-121	122-129
Frequency	2	7	12	15	10	4

5- What is the types of data?

CHAPTER (2)

Measures of Central Tendency

Contents.

In the previous section we looked at some graphical and tabular techniques for describing a data set. We shall now consider some numerical characteristics of a set of measurements. Suppose that we have a sample with values $x_1, x_2, x_3, \ldots, x_n$. There are many characteristics associated with this data set, for example, the central tendency and variability. The most commonly used measures are mean, geometric mean, harmonic mean, mode, median, quartiles, deciles and percentiles.

■ *Mean for ungrouped data*

$$Mean = \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i}^{n} x_i}{n}$$

Example: 1

Find the mean for the set data: 13, 17, 12, 11, and 17

Solution:

$$\bar{x} = \frac{13+17+12+11+17}{5} = \frac{70}{5} = 14$$

■ Mean for grouped data

$$Mean = \bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i}^{n} x_i f_i}{\sum_{i}^{n} f_i}$$

Example: 2

a) Find the mean of the set of data: 2, 4, 7, 8, and 9

b) Find the mean from the set of grouped data

Mark	10	20	30	40	50
Frequency	5	10	5	20	15

Solution

a)

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{2+4+7+8+9}{5} = \frac{30}{5} = 6.$$

b)

Mark (x)	10	20	30	40	50	Sum
Frequency (f_i)	6	8	7	15	14	50
$x_i f_i$	60	160	210	600	700	1730

$$\therefore \; \bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{1730}{50} = 34.6$$

Example: 3

Calculate the arithmetic mean of the following data

Sets	40-50	50-	60-	70-	80-	90- 100
f_i	18	10	5	22	25	20

Solution:

Sets	f_i	x_i	$x_i f_i$
40-50	18	(40+50)/2=45	810
50-	10	55	550
60-	5	65	325
70-	22	75	1650
80-	25	85	2125
90-100	20	95	1900
	$\Sigma f_i = 100$		$\Sigma x_i f_i = 7360$

$$\Rightarrow \bar{x} = \frac{7360}{100} = 73.6$$

Example: 4

Calculate the arithmetic mean of the following data.

Class	f_i	Class	f_i	Class	f_i	Class	f_i
0 – 10	122	40 – 50	311	80 – 90	180	120 – 130	106
10 – 20	180	50 – 60	278	90 – 100	175	130 – 140	99
20 – 30	256	60 – 70	250	100 – 110	143	140 – 150	97
30 – 40	350	70 – 80	211	110 – 120	120	150 – 160	75

Solution:

Class	x_i	f_i	$x_i f_i$	Class	x_i	f_i	$x_i f_i$
0 – 10	5	122	610	80 – 90	85	180	15300
10 – 20	15	180	2700	90– 100	95	175	16625
20 – 30	25	256	6400	100 – 110	105	143	15015
30 – 40	35	350	12250	110 – 120	115	120	13800
40 – 50	45	311	13995	120 – 130	125	106	13250
50 – 60	55	278	15290	130 – 140	135	99	13365
60 – 70	65	250	16250	140 – 150	145	97	14065
70 – 80	75	211	15825	150 – 160	155	75	11625

$$\Rightarrow \sum_{i=1}^{16} f_i = 2953, \ \sum_{i=1}^{16} x_i f_i = 196385,$$

so that:

$$\bar{x} = \frac{\sum_{i=1}^{16} x_i f_i}{\sum_{i=1}^{16} f_i} = \frac{196365}{2953} = 66.4968.$$

Example: 5

Calculate the arithmetic mean of the following data

Wight	32-34	34-36	36-38	38-40	40-42	42-44
Students	4	7	13	10	5	1

Solution:

Wight	f_i	x_i	$x_i f_i$
32-34	4	(32+34)÷2=33	4×33=132
35-37	7	36	7×36=252
38-40	13	39	13×39=507
41-43	10	42	10×42=420
44-46	5	45	5×45=225
47-49	1	48	1×48=48
Sum	40		1584

$$\Rightarrow \quad \bar{x} = \frac{\sum_{i=1}^{6} x_i f_i}{\sum_{i=1}^{6} f_i} = \frac{1584}{40} = 39.6\ k.g$$

Advantages and disadvantages of the mean

Advantages:

(i) All values in the distribution are used in its calculation.

(ii) Its method of calculation is simple and most people understand the meaning of its result.

(iii) Its result can easily be used in further analysis.

Disadvantages:

(i) Its result can be easily distorted by extreme values

(ii) In case of open end classes, mean can be calculated only if their class marks are determined. If such classes contain a large proportion of the values, then the mean may be subjected to substantial error.

■ Geometric mean for ungrouped data

Geometric mean is defined as the *n*-th root of the product of *n* observations.

Geometric mean:

$$G.M = \sqrt[n]{x_1 \times x_2 \times x_3 \ldots \times x_n},$$

where:

n = Number of observations.

We can see that:

$$\Rightarrow \quad G.M = e^{\ln\left(\frac{\sum \ln(x_i)}{n}\right)}$$

Example: 6

Compute the geometric mean of 100, 200 and 300.

Solution:

- ***(Method 1)***

$$G.M = \sqrt[n]{\prod x_i} = \sqrt[3]{(100)(200)(300)} = \sqrt[3]{6000000} = (6000000)^{1/3} = 181.712$$

- ***(Method 2)***

$$\log(G.M) = \frac{1}{n}\sum \log(x_i) = \frac{1}{3}(\log(100) + \log(200) + \log(300))$$

$$= \frac{1}{3}(2.00000 + 2.30103 + 2.4712) = 2.25938$$

$$\Rightarrow \quad G.M = 10^{\log(G.M)} = 10^{2.25938} = 181.712.$$

- ***(Method 3)***

$$\ln(G.M) = \frac{1}{n}\sum \ln(x_i) = \frac{1}{3}(\ln(100) + \ln(200) + \ln(300))$$

$$= \frac{1}{3}(4.60517 + 5.29832 + 5.70328) = \frac{1}{3}(15.60727) = 5.20242$$

$$\Rightarrow \quad G.M = e^{\ln(G.M)} = e^{5.20242} = 181.712.$$

■ Geometric mean for grouped data

The geometric mean of the set of observations is defined by:

$$G.M = \sqrt{(x_1^{f_1} x_2^{f_2} x_3^{f_3} \ldots\ldots\ldots\ldots x_n)}$$

$$= \left[x_1^{f_1} x_2^{f_2} \ldots\ldots\ldots x_n^{f_n}\right]^{\frac{1}{N}} = \left[\prod_{i=1}^{n} x_i^{f_i}\right]^{\frac{1}{N}}$$

$$= 10^{\left(\frac{1}{N}\sum_{i=1}^{n} f_i \, Log\, x_i\right)}, \text{ where } N = \sum_{i=1}^{n} f_i$$

Example: 7

Find the Arithmetic, Geometric Mean of the following data

Class	20-29	30-39	40-49	50-59	60-69	Sum
Frequency(f)	3	5	20	10	5	43

Solution:

Class	f_i	x_i	$x_i f_i$	$f_i \log x_i$
20-29	3	24.5	73.5	4.17
30-39	5	34.5	172.5	7.69
40-49	20	44.5	890	32.97
50-59	10	54.5	545	17.37
60-69	5	64.5	322.5	9.05
Sum	N=43		2003.5	71.24

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i f_i}{\sum_{i=1}^{5} f_i} = \frac{2003.5}{43} = 46.593,$$

$$\because G = AntiLog\left(\frac{1}{N}\sum_{i=1}^{n} f_i \ Log\ x_i\right)$$

$$\therefore G = 10^{\left(\frac{71.24}{43}\right)} = 10^{(1.6567)} = 45.36$$

■ Harmonic mean for ungrouped data

Harmonic mean is one of several kinds of average,

$$H.M. = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$$

Example: 8

Compute the harmonic mean of 100, 200 and 300.

Solution:

$$H.M = \frac{n}{\sum\left(\frac{1}{x_i}\right)} = \frac{3}{\left(\frac{1}{100}+\frac{1}{200}+\frac{1}{300}\right)} = \frac{3}{(0.01000+0.00500+0.00333)}$$

$$= 163.637$$

Example: 9

If $x = \{2, 5, 3, 4, 7, 8, 8\}$, compute the harmonic mean of x.

Solution:

$$H.M = \frac{n}{\sum\left(\frac{1}{x_i}\right)} = \frac{7}{\left(\frac{1}{2}+\frac{1}{5}+\frac{1}{3}+\frac{1}{4}+\frac{1}{7}+\frac{1}{8}+\frac{1}{8}\right)} = \frac{7}{1.68} = 4.17$$

Example: 10

The harmonic mean of the numbers 2, 4 and 8 is

$$H.M=\frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}}=3.43$$

■ **Harmonic mean for grouped data**

The harmonic mean *H.M* of the set of observations is defined by:

$$H.M=\frac{n}{\sum_{i=1}^{n}\frac{f_i}{x_i}}$$

Example: 11

Find the Harmonic Mean for the data in example 6.

Solution:

Class	f_i	x_i	$\frac{f_i}{x_i}$
20-29	3	24.5	0.1224
30-39	5	34.5	0.1449
40-49	20	44.5	0.4494
50-59	10	54.5	0.1835
60-69	5	64.5	0.0775
Sum	N=43		0.9777

$$H.M=\frac{43}{0.9777}=43.9808$$

■ ***The Relation between the Arithmetic, Geometric and Harmonic Means:***

$$H.M\le G.M\le\bar{X}.$$

■ *Mode for ungrouped data*

Mode is the value of a distribution for which the frequency is maximum. In other words, mode is the value of a variable, which occurs with the highest frequency.

- The mode of the list (4, 2, 2, 3, 3, 3, 5) is 3. The mode is not necessarily well defined.
- The list (4, 2, 2, 3, 3, 5) has the two modes 2 and 3.

Example: 12

Find Mode of the data 3, 12, 4, 6, 1, 4, 2, 5, 8

Solution:

Mode=4

Example: 13

Find Mode of the data

a. 5 5 5 3 1 5 1 4 3 5
b. 1 2 2 2 3 4 5 6 6 6 7 9
c. 1 2 3 6 7 8 9 10

Solution:

a. Mode=5
b. Bimodal=2, 6
c. No Mode

■ Mode for grouped data

$$Mode = a + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)L \text{ Or } Mode = a + \frac{f_m - f_1}{2f_m - f_2 - f_1} \times L,$$

where

$a =$ lower class boundary of the modal class

$\Delta_1 =$ difference of frequency between modal class and class before it

$\Delta_2 =$ difference of frequency between modal class and class after

$L =$ size of the median class interval

$f_m =$ frequency of the modal class.

$f_1 =$frequency of then class proceeding to the modal class.

$f_2 =$ frequency of the class succeeding to the modal class.

Note that: The class which has highest frequency is the modal class

Example: 14

Find the mode the following frequency distribution.

Class	10-14	15-19	20-24	25-29	30-34	Sum
Frequency (*f*)	7	11	14	13	5	50

Solution:

The modal class is the third class with frequency 14. $\Delta_1 = 14 - 11 = 3, \Delta_2 = 14 - 13 = 1$, $a = 19.5, L = 24.5 - 19.5 = 5$. Thus,

$$\text{Mode} = a + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)L = 19.5 + \left(\frac{3}{3+1}\right)5 = 23.25.$$

■ Advantages and disadvantages of the mode

Advantages:

- Its result will not be affected by extreme values and open end classes.
- If data are not grouped, it can be determined easily.

Disadvantages:

- It has to be supplemented by other statistics.
- It is difficult to obtain an accurate estimate of the mode if the values are classified into a frequency distribution.

■ Median for ungrouped data

Median = the middle datum, when *n* is odd.

Median = the mean of the two middle data, when *n* is even.

For the set of data	For the set of data
10, 15, 16, 21, 25 ↑ middle datum median = 16	13, 15, 27, 27 ↑ ↑ middle of two data median = (15 + 27) ÷ 2 = 21

Example: 15

Determine the median from the following sets of data.

a. 2, 6, 9, 4, 3, 4, 5, 6, 7

b. 9, 7, 3, 4, 8, 6, 7

c. 10kg, 12kg, 18kg, 10kg, 16kg, 23 kg

Solution:

a. 2, 6, 9, 4, 3, 4, 5, 6, 7

Re-arrange the numbers in sequence

2, 3, 4, 4, (5), 6, 6, 7, 9

↑

Since $n=9$ is odd, then the order of the median is $\frac{n+1}{2}=\frac{9+1}{2}=5(fiveth)$

Median = 5

b. 9, 7, 3, 4, 8, 6, 7

Re-arrange the numbers in sequence:

3, 4, 6, (7), 7, 8, 9

Median = 7

c.10kg, 12kg, 18kg, 10kg, 16kg, 23 kg

Re-arrange the numbers in sequence:

10, 10, 12, 16, 18, 23

Since $n=6$ is even, then the order of the median is

$$\frac{n}{2},\ \frac{n}{2}+1=3,\ 4 \text{(thired and fourth)}$$

$$\text{Median} = \frac{12+16}{2} = 14 \text{ kg.}$$

Example: 16

Find the median of 2, 4, 8, 7, 4, 6, 10, 8, and 5.

Solution:

Array: 2, 4, 4, 5, 6, 7, 8, 8, 10

Middle value = ((9 + 1) /2) th value = 5 th value= $X_{(5)}$

Median = 6

Median for grouped data

Steps to find Median of group data

1) Compute the less than type cumulative frequencies.
2) Determine $n/2$, one-half of the total number of cases.
3) Locate the median class for which the cumulative frequency is more than $n/2$.
4) Determine the lower limit of the median class. This is a.
5) Sum the frequencies of all classes prior to the median class. This is f_1.
6) Determine the frequency of the median class. This is $f_{median} = f_2 - f_1$.
7) Determine the class width of the median class. This is L.

$$Median = a + \left(\frac{\frac{n}{2} - f_1}{f_{median}} \right) L,$$

where, n is the number of items in the data (total frequency).

Example 17

The following table shows the daily of a random sample of construction workers. Calculate its median.

Set	200 - 399	400 - 599	600 - 799	800 - 999	1000 - 1199	1200 - 1399
f_i	5	15	25	30	18	7

Solution:

Set	Cumulative Frequency F_i
Less than 199.5	0
>399.5	5
>599.5	20
>799.5	45 → 50
>999.5	75
>1199.5	93
>1399.5	100

$\frac{n}{2}=\frac{100}{2}=50$, so the median lies in the $4th$ class.

$$Median = a+\left(\frac{\frac{n}{2}-f_1}{f_{median}}\right)L,$$

where $a=799.5$ is the lower class boundary,

$f_{median}=f_2-f_1=75-45=30$ and $L=999.5-799.5=200$ is the class interval.

$$Median = 799.5+\left(\frac{50-45}{75-45}\right)200=832.8.$$

■ *Empirical relation between Mean, Median and Mode:*

The relationship between mean, median and mode depends upon the nature of the distribution. A distribution may be symmetrical or asymmetrical. In asymmetrical distribution the mean, median and mode are equal

Mean= Median= Mode

In a moderately asymmetric distribution the difference between the mean and mode is three times the difference between the mean and median.

Mean-Mode = 3(Mean-Median)

■ *Advantages and disadvantages of the median*

Advantages:

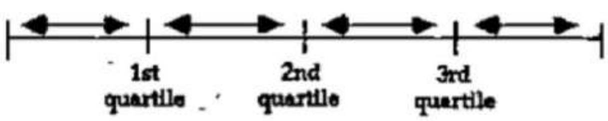
Its result will not be affected by extreme values and open end classes.

■ *Disadvantage:*

It has to be supplemented by other statistics because it does not reflect the distribution in the way that the mean does, that is, including all values.

■ *Quartiles, Deciles and Percentiles*

Three of these divide the data set into four, ten or hundred divisions, respectively.



- Quartiles, Deciles and Percentiles are measures of position useful for comparing scores within one set of data.
- For a set of data you can divide the data into three quartiles (Q_1, Q_2, Q_3), nine deciles $(D_1, D_2, ... D_9)$ and 99 percentiles $(P_1, P_2, ..., P_{99})$.

- Q_1 (Quartile one) covers the first 25% items of the series and it divide the first half of the series into two equal parts. Q_2 (Quartile two) is the median or middle value of the series and Q_3 (quartile three) covers 75% items of the series.

- Deciles are those values which divide the series into ten equal parts. There are nine deciles i.e. $D_1, D_2, ... D_9$ in a series and 5th decile is same as median and 2nd quartile, because those values divide the series in two equal parts.

❖ ***Calculation of Quartiles:***

The calculation of quartiles is done exactly in the same manner as it is in case of the calculation of median.

- **In case of individual and discrete series:**

$$Q_i = Size\ of\ \frac{i(n+1)}{4} th\ item\ of\ the\ series$$

- **In case of continuous series:**

$$Q_i = Size\ of\ \frac{in}{4} th\ item\ of\ the\ series\ ,\ i = 1, 2, 3.$$

- **Interpolation formula for continuous series:**

$$Q_i = a + \left(\frac{\frac{in}{4} - f_1}{f_{Q_i}} \right) L,\ \ i = 1,2,3.$$

The calculation of deciles and percentiles are done exactly in the same manner as it is in case of the calculation of quartiles.

Example 18

Find Q_1 and Q_3 of the following:

(a) 4, 5, 6, 7, 8, 9, 12, 13, 15, 10, 20

(b) 100, 500, 1000, 800, 600, 400, 7000 and 1200

Solution:

(a) Values of the variable are in ascending order:

4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 20

So $n = 11$ (No. of Values)

$$Q_1 = Size\ of\ \frac{(n+1)}{4} th\ item\ of\ the\ series = \text{size of 3rd item} = 6$$

$$Q_3 = Size\ of\ \frac{3(n+1)}{4} th\ item\ of\ the\ series = \text{size of 9th item} = 13$$

Required Q_1 and Q_3 are 6 and 13 respectively,

(b) The values of the variable in ascending order are:

100, 400, 500, 600, 700, 800, 1000, 1200, $n = 8$

$$Q_1 = Size\ of\ \frac{(n+1)}{4} th\ item\ of\ the\ series$$

$$= Size\ of\ \frac{(8+1)}{4} th\ item\ of\ the\ series$$

$$= size\ of\ 2.25th\ item$$

= size of {Second item + 0.25(Third item - Second item)}

= 400 + 0.25 (500 - 400) = 400 + 25 = 425.

$$Q_3 = Size\ of\ \frac{3(n+1)}{4} th\ item\ of\ the\ series$$

$$= Size\ of\ \frac{3(8+1)}{4} th\ item\ of\ the\ series$$

= size of 6.75th item

= size of [6th item + 0.75(7th item - 6th item)]

= 800 + 0.75 (1000 - 800) = 800 + 150 = 950

Required Q_1 and Q_3 are 425 and 950 respectively.

Example 19

The following Table shows the weights of 2,000 students at the Sohag University

Weight	60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	140-149
f_i	138	163	325	541	427	214	110	52	30

Find Q_1, Q_3, D_3 and P_{95}.

Solution:

Set	Cumulative Frequency	
Less than 59.5	0	
>69.5	138	
>79.5	301	$Q_1 \rightarrow$ 500
>89.5	626	$D_3 \rightarrow$ 600
>99.5	1167	$Q_3 \rightarrow$ 1500
>109.5	1594	
>119.5	1808	$P_{95} \rightarrow$ 1900
>129.5	1918	
>139.5	1970	
>149.5	2000	

The order of Q_1 is $\frac{n}{4}=\frac{2000}{4}=500$, so Q_1 lies in the $3th$ class.

Hence,

$$Q_1=a+\left(\frac{\frac{n}{4}-f_1}{f_{Q_1}}\right)L=79.5+\left(\frac{500-301}{626-301}\right)10=85.62kg$$

The order of Q_3 is $\frac{3n}{4}=\frac{6000}{4}=1500$, so Q_3 lies in the $5th$ class.

Hence,

$$Q_3=a+\left(\frac{\frac{3n}{4}-f_1}{f_{Q_3}}\right)L=99.5+\left(\frac{1500-1167}{1594-1167}\right)10=103.78kg$$

$$D_3=P_{30}=a+\left(\frac{\frac{3n}{10}-f_1}{f_{D3}}\right)L=79.5+\left(\frac{600-301}{325}\right)10=88.7kg$$

$$P_{95}=a+\left(\frac{\frac{95n}{100}-f_1}{f_{P95}}\right)L=119.5+\left(\frac{1900-1808}{110}\right)10=127.89Kg.$$

Advantages of Quartiles, Deciles and Percentiles:

(i) These averages can be directly determined in case of open end class intervals without knowing the lower limit of lowest class and upper limit of the largest class.

(ii) These averages can be calculated easily in absence of some data in a series.

(iii) These averages are helpful in the calculation of measures of dispersion.

(iv) These averages are not affected very much by the extreme items.

(v) These averages can be located graphically.

Disadvantages of Quartiles, Deciles and Percentiles:

(i) These averages are not easily understood by a common man. These are not well defined and easy to calculate.

(ii) These averages are not based on all the observations of a series.

(iii) These averages cannot be computed if items are not given in ascending or descending order.

(iv) These averages are affected very much by the fluctuation of sampling.

Exercises:

1. Calculate mean of the following data.

a) 4, 3, 2, 5, 3, 4, 5, 1, 7, 3, 2, 1
b) 30, 70, 10, 75, 500, 8, 42, 250, 40, 36
c) 35, 46, 27, 38, 52, 44, 50, 37, 41, 50

2. Find the mean of first 10 even numbers.
3. Find m, G.M, H.M, median and mode of following

a)

X	5	6	7
f	1	4	3

b)

X	20	21	22	23	24	25	26
f	1	2	4	7	5	3	1

4. Find mean, G.M, H.M, median and mode of following data.

Marks	20	30	40	50	60	70
No. Of Students	8	12	20	10	6	4

5. Find mean, G.M, H.M, median and mode of following

CI	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	6	14	16	27	22	15

6. Find mean, G.M, H.M, median and mode of following data.

CI	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Frequency	10	12	8	20	11	4	5

7. Find A.M, G.M, H.M, median and mode of following data.

CI	10-20	20-40	40-70	70-120	120-200
Frequency	4	10	26	8	2

8. Find the missing frequencies from the data given below if mean is 60.

Marks	50	55	60	65	70	Total
No. Of Students	?	20	25	?	10	100

9. Find the missing frequencies from the data given below if mean is 60.

Marks	60-62	63-65	66-68	69-71	72-74
No. Of Students	15	54	?	81	24

10. In a class of 60 students 10 have failed with an average mark of 15. If the total marks of all the students were 1800, find the average marks of those who have passed?

11. The mean of 10 observations is 10 and the sum of first four observations is 10. Find the 5th observation.

12. The numbers 3,5,7 and 4 have frequencies *x, (x+2), (x-2), (x+1)* respectively. If the arithmetic mean is 4.424. Find the value of *x*.

13.Find the G.M of

a. 3, 6, 24 and 48

b. 2574, 475, 5, 0.8, 0.08,0.005, 0.009

c. 5,10,200,12375,2575

14. Find the G.M following data.

Marks	10	20	30	40	50	60
No. Of Students	12	15	25	10	6	2

15. If mean and G.M of two numbers are 12.5 and 10 respectively. Find those numbers.

16. If mean and G.M of two numbers are 10 and 8 respectively. Find those numbers.

17. Find the H.M of 3,4,12

18. Find H.M of $\frac{1}{5},\frac{1}{6},\frac{1}{7},\frac{1}{8},\frac{1}{9}$

19. Calculate median, quartiles, 5th decile and 45th percentile of the following

a. 391, 591, 407, 384, 1490, 2488, 672, 522, 753, 777

b. 31, 28, 49, 57, 31, 56, 27, 49

c. 16, 14, 11,11, 13, 10, 10, 9, 7, 7, 4, 3, 2, 1

20. Find D_6, P_{65} for the following data

10, 20, 25, 30, 35, 40, 50, 55, 60

21. Find three quartiles, D_7, P_5, P_{90} from the following data.

CI	10-20	20-40	40-70	70-120	120-200
Frequency	4	10	26	8	2

22. Find three quartiles, Q_1, Q_3, D_2, P_5, P_{90} of following data.

Marks	10	20	30	40	50	60
No. Of Students	12	15	25	10	6	2

23. What is median class and modal class?

24. Mean of the 10 observations is 20. If each observation is increased by 5 what is the mean of the resultant series?

25. Mean of the 5 observations is 10. If each observation is doubled then what is the mean of the new series.

26. The GM and HM of two observations are respectively 18 and 10.8. Find the observations.

27. The arithmetic mean of 10 observations is 72.5 and the arithmetic mean of 9 observations is 63.2, find the value of 10^{th} observation.

CHAPTER (3)

Measures of Variation

Contents.

■ ***Variation***: in a data set is the amount of difference between data values. In a data set with little variation (i.e. 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5), almost all data values would be close to one another. The histogram of such a data set would be narrow and tall. In a data set with a great deal of variation (i.e. 1, 2, 3, 5, 6, 6, 7, 8, 8, 9, 10), the data values would be spread widely. The histogram of this data set would be low and wide. Just as there are many measures of central tendency or location there are many measures of spread or variability.

■ ***Common Measures of Variation:***
A measure of central tendency gives us a typical value that can be used to describe the whole set of data, but using it we lose all the information about how the data was clustered or spread. We will introduce five measures of variation – range, Inter quartile range, variance, standard deviation and coefficient of variation that will give us some indication of data scatter.

■ ***Range For ungrouped data:***
The range is the difference between the largest and smallest data values in a data set.

Range = (highest value – lowest value)

■ ***Range For grouped data:*** the range is the difference between the highest class boundary and the lowest boundary.

Example 1

a) Find the range of the data: 1, 5, 2, 12, 3, 3, 19

b) Find the range of the grouped data

Class	10 - 14	15 - 19	20 - 24	25 - 29
Frequency	12	18	7	3

Solution

a) The range = 19-1=18

b)

Class	Frequency	x_i
10 - 14	12	(9.5+14.5)/2=12
15 - 19	18	(14.5+19.5)/2=17
20 - 24	7	(19.5+24.5)/2=22
25 - 29	3	(24.5+29.5)/2=27

The range = 27-12=15.

■ *Inter quartile range*

$$\text{Inter quartile range} = Q_3 - Q_1,$$

where Q_1, Q_2, Q_3 are called quartiles which divide the data (which have been ranked, i.e. arranged in order) into four equal parts. Moreover,

Q_2 is the median of the whole set of data,

Q_1 is the median of the lower half,

Q_3 is the median of the upper half.

■ *Quartile deviation*

$$\text{Q.D.} = \frac{(Q_3 - Q_1)}{2}$$

Example 2

Find the inter quartile range of

11, 17, 18, 20, 28, 13, 29, 25, 27, 16, 19, 34, 32, 33, 30

Solution

Rearranging the data into ascending order:

11, 13, 16, 17, 18, 19, 20, 25, 27, 28, 29, 30, 32, 33, 34

- The rank of Q_1,

$$i = \frac{1}{4}(15) = 3.75$$

Q_1 is located at the $3 + 1 = 4^{th}$ term

$$\therefore Q_1 = 17$$

- The rank of Q_3,

$$i = \frac{3}{4}(15) = 11.25$$

Q_3 is located at the $11 + 1 = 12^{th}$ term

$$\therefore Q_3 = 30$$

The inter-quartile range$= Q_3 - Q_1 = 13$.

Example 3

Find the inter quartile range of

a) 1 3 5 5 6 7 8 9 15

b) 1 3 5 5 6 7 8 9

Solution

a) $n = 9$ (is odd)

- The rank of Q_1,

$$i = \frac{1}{4}(9) = 2.25$$

Q_1 is located at the $2 + 1 = 4^{th}$ term

$$\therefore Q_1 = 5$$

- The rank of Q_3,

$$i = \frac{3}{4}(9) = 6.75$$

Q_3 is located at the $6 + 1 = 7^{th}$ term

$$\therefore Q_3 = 8.$$

The inter-quartile range $= Q_3 - Q_1 = 3$.

b) $n = 8$ (is even)

$$Q_1 = Size\ of\ \frac{(n+1)}{4} th\ item\ of\ the\ series$$

$$= Size\ of\ \frac{(8+1)}{4} th\ item\ of\ the\ series$$

$$= Size\ of\ 2.25th\ item\ of\ the\ series$$

= size of {Second item + 0.25(Third item - Second item)}

$= 3 + 0.25\ (5 - 3) = 3 + 0.5 = 3.5.$

$$Q_3 = Size\ of\ \frac{3(n+1)}{4} th\ item\ of\ the\ series$$

$$= Size\ of\ \frac{3(8+1)}{4} th\ item\ of\ the\ series$$

$$= Size\ of\ 6.75th\ item\ of\ the\ series$$

= size of {Six item + 0.75(Seven item - Six item)}

$= 7 + 0.75\ (7 - 6) = 7 + 0.75 = 7.75.$

The inter-quartile range $= Q_3 - Q_1 = 7.75 - 3.5 = 4.25$.

The inter-quartile range is good for skewed distributions.

Example 4

The following Table shows the marks of 100 students

mark	60-69	70-	80-	90-	100-	110-	120-	130-	140-149
f_i	12	8	14	16	10	17	11	7	5

Find the inter-quartile range.

Solution:

Set	Cumulative Frequency F_i	
Less than 59.5	0	
>69.5	12	
>79.5	20	Q_1 → 25
>89.5	34	
>99.5	50	
>109.5	60	Q_3 → 75
>119.5	77	
>129.5	88	
>139.5	95	
>149.5	100	

The order of Q_1 is $\frac{n}{4}=\frac{100}{4}=25$, so Q_1 lies in the 3*th* class.

Hence,

$$Q_1 = a + \left(\frac{\frac{n}{4} - f_1}{f_{Q_1}}\right)L = 79.5 + \left(\frac{25-20}{34-20}\right)10 = 83.0714$$

The order of Q_3 is $\frac{3n}{4}=\frac{300}{4}=75$, so Q_3 lies in the 5*th* class.

Hence,

$$Q_3=a+\left(\frac{\frac{3n}{4}-f_1}{f_{Q_3}}\right)L=109.5+\left(\frac{75-60}{77-60}\right)10=118.324$$

The inter-quartile range$=Q_3-Q_1=118.324-83.0714=35.2526$.

■ ***Variance***

The variance is the average of the squared differences between each data value and the mean. Variance is useful for comparing variability in two data sets. The formula for the variance is different depending on whether we are treating the data as a **<u>population</u>** or as a **<u>sample</u>**. Specifically:

- If we treat the data as a population, we use the number of observations, *N*, in the denominator.
- If we treat the data as a sample, we divide by the number of observations minus 1 in the denominator

■ ***Variance for ungrouped data:*** Suppose that a data set of *n* sample measurements $x_1,x_2,\dots,x_n$ with mean $\bar{x}$. Then the sample variance x^2 for an ungrouped data is

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

- If $x_1,x_2,\dots,x_N$ is the whole population with mean μ, then variance σ^2 is

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

■ ***Variance for grouped data:*** Suppose that a data set of *n* sample measurements

$x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ is grouped into k classes in a frequency table, where x_i is a midpoint and f_i is the frequency of the i-th class interval. Then the sample standard deviation s for a grouped data is

$$s^2 = \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}\right)^2 f_i}{n-1}, n = \sum_{i=1}^{k} f_i$$

If $x_1, x_2, \ldots, x_N$ is the whole population with mean μ, then **population variance** σ^2 is

$$\sigma^2 = \frac{\sum_{i=1}^{k} \left(x_i - \mu\right)^2 f_i}{N}, N = \sum_{i=1}^{k} f_i .$$

Example 5

Find the variance of the sample $x = \{50, 60, 70, 80, 90, 100\}$.

Solution

x	$x-\bar{x}$	$(x-\bar{x})^2$
50	-25	625
60	-15	225
70	-5	25
80	5	25
90	15	225
100	25	625
	$\Sigma(x-\bar{x})=0$	$\Sigma(x-\bar{x})^2=1750$

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{1750}{5} = 350$$

■ ***Standard deviation for ungrouped data:*** Suppose that a data set of n sample measurements $x_1, x_2, ..., x_n$ with mean $\bar{x}$. Then the sample standard deviation s for a ungrouped data is

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}.$$

If $x_1, x_2, ..., x_N$ is the whole population with mean μ, then population standard deviation σ is

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

■ ***Standard deviation for grouped data:*** Suppose that a data set of n sample measurements $x_1, x_2, ..., x_n$ with mean $\bar{x}$ is grouped into k classes in a frequency table, where x_i is a midpoint and f_i is the frequency of the i-th class interval. Then the sample standard deviation s for a grouped data is

$$s = \sqrt{\frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 f_i}{n-1}}, n = \sum_{i=1}^{k} f_i$$

If $x_1, x_2, ..., x_N$ is the whole population with mean μ, then population standard deviation σ is

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

Example 6

The following table represents the different marks obtained by 7 students: 33, 42, 51, 59, 67, 75, 83. Find the standard deviation of the marks.

Solution

The mean is $\bar{x} = \frac{33+42+51+59+67+75+83}{7} = \frac{410}{7} \approx 58.57$, and the standard deviation is

$$\sigma = \sqrt{\frac{1}{7}\left((33-58.57)^2 + (42-58.57)^2 + ... + (83-58.57)^2\right)} \approx 13.87.$$

Example 7

The following data is given. Find the variance.

Class	0 - 10	10 -20	20 -30	30 -40	40 -50
f	10	20	40	20	10

Solution:

Of course, it is unnecessary to do everything in the table below, but you should know how to do the problem using both computational and definitional formulas.

class	f_i	x_i	$f_i\,x_i$	$f_i\,x_i^{\,2}$	$(x-\bar{x})$	$f_i\,(x_i-\bar{x})$	$f_i\,(x_i-\bar{x})^2$
0-10	10	5	50	250	-20	-200	4000
10-20	20	15	300	4500	-10	-200	2000
20-30	40	25	1000	25000	0	0	0
30-40	20	35	700	24500	10	200	2000
40-50	10	45	450	20250	20	200	4000
Total	100		2500	74500		0	12000

So $\sum f = n = 100$, $\sum f_i\,x_i = 2500$, $\sum f_i\,x_i^{\,2} = 74500$,

$\bar{x} = \frac{\sum f_i\,x_i}{n} = \frac{2500}{100} = 25$ and, using the computational formula,

$$s^2 = \frac{\sum f_i x_i^2 - n\bar{x}^2}{n-1} = \frac{74600-100(25)^2}{100-1} = \frac{12000}{99} = 121.21212$$

Or using the definitional formula $s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{12000}{99} = 121.21212$. So

$s = \sqrt{121.21212} = 11.0096$ and $C = \frac{s}{\bar{x}} = \frac{11.0096}{25} = 0.4404$.

Example 8

Consider the following sample

Class	0.5 – 1.5	1.5 -2.50	2.5 -3.50	3.50 -4.5
f	1	0	1	2

a) Calculate the variance and standard deviation

b) Calculate the inter quartile range

Solution:

class	f_i	x_i	$f_i x_i$	$f_i x_i^2$	Cumulative Frequency F_i	
0.5 - 1.5	1	1	1	1	>0.5	0
1.5 - 2.5	0	2	0	0	>1.5	1
2.5 - 3.5	1	3	3	9	>2.5	1
3.5 - 4.5	2	4	8	32	>3.5	3
Total	4		12	42	>4.5	7

Q_1

Q_3

- *Calculate the variance*

$\sum f = n = 4, \sum f_i x_i = 12, \sum f_i x_i^2 42,$

$$\bar{x} = \frac{\sum fx}{n} = \frac{12}{4} = 3$$

$$s^2 = \frac{\sum fx^2 - n\bar{x}^2}{n-1} = \frac{42-4(3)^2}{3} = \frac{6}{3} = 2$$

$s = \sqrt{\text{variance}} = \sqrt{2} = 1.414$.

- *Calculate the inter quartile range*

For the first quartile $position = \frac{1}{4}(n+1) = 0.25(5) = 1.25$. This location is above 1 and below 2, so use the class 2.5 to 3.5. Then, we find $Q_1 = 2.5 + \left[\frac{0.25(4)-1}{1}\right]1 = 2.5$

For the third quartile $\frac{3}{4}(n+1)=0.75(5)=3.75$. This location is above 2 and below 4,

so use the class 3.5 to 4.5. Then, we find $Q_3=3.5+\left[\frac{0.75(4)-2}{2}\right]1=4.0$.

$$IQR=Q3-Q1=4.0-2.5=1.5.$$

■ ***The coefficient of variation***

- *Measure of Relative Variation*
- *It is sometimes expressed as a percentage*
- *Shows Variation Relative to Mean*
- *Used to Compare 2 or More Groups*
- *It is the* ***ratio*** *of the sample* ***standard deviation*** *to the sample* ***mean.*** *The formula for the coefficient of variation (C.V) is:*

$CV=\left(\frac{s}{\bar{x}}\right)\times 100\%$ for a sample and $CV=\left(\frac{\sigma}{\mu}\right)\times 100\%$ for a population.

Where $\bar{x}$ = the mean of the sample, μ= the mean of the population
s = the standard deviation of the sample,
σ= the standard deviation of the population

Example 9

Calculate the coefficient of variation for the price of 400 g cans of pet food, given that the mean is 81 cents and s = 6.77 cents. Interpret the results.

Solution

$$\begin{aligned} CV &= 100\left(\frac{s}{\bar{x}}\right)\% \\ &= 100\left(\frac{6.77}{81}\right)\% \\ &= 8.36\% \end{aligned}$$

This means that the standard deviation of the price of a 400g can of pet food is 8.36% of the mean price.

Exercises

1. Find the range, IQR, standard deviation, the coefficient of variation, the coefficients of skewness and kurtosis for each set of ungrouped sample data:
   a. 1,2,2,3,3,3,3,4,4,5
   b. 1,1,1,1,2,3,4,5,5,5
2. For the following sample of ten ungrouped measurements: 4,2,3,5,3,1,6,4,2,3
   a. Find the standard deviation.
   b. How many measurements lie within one standard deviation from the mean.
   c. How many measurements lie within two standard deviations from the mean.
3. Find the mean and standard deviation of the following sample data set
   The grade-level reading scores from a test given to randomly sample of 12 students are
   9 11 11 15 10 12 12 13 8 7 13 12
4. Find the range, IQR, standard deviation, the coefficient of variation and the coefficients of skewness and kurtosis for the set of grouped sample data:

Interval	0.5-3.5	3.5-6.5	6.5-9.5	9.5-12.5
Frequency	2	5	7	1

If a distribution has negative skewness, in what order (lowest to highest) will the averages be?

A) mean, mode, median
B) mean, median, mode
C) mode, median, mean
D) median, mode, mean

5. A distribution with positive kurtosis has ________ than a normal distribution.

A) more cases in the centre and fewer in the tails
B) fewer cases in the centre and more in the tails

6. The coefficient of ________ is a measure of the shape of a distribution.

A) skewness
B) kurtosis
C) both skewness and kurtosis
D) neither skewness nor kurtosis

CHAPTER (4)

Correlation and Simple Regression

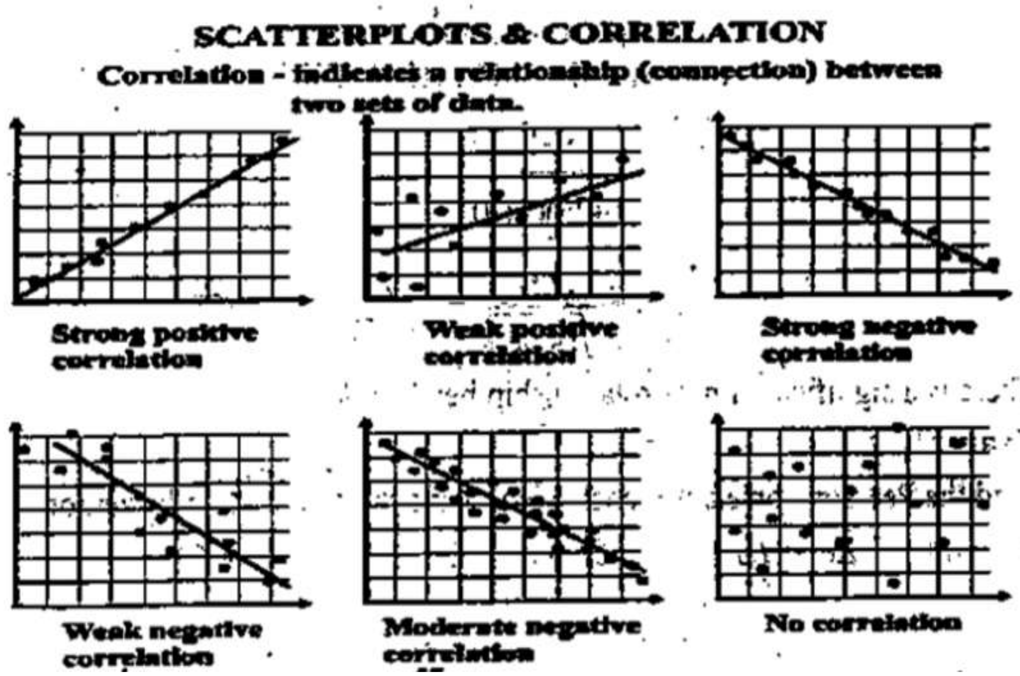
Contents.

What is the relationship between two variables?
The strength of the linear relationship between two variables is called the coefficient of correlation, r.

Correlation= direction and strength of relationship between two variables

■ Properties of Coefficient of Correlation

- r can range from -1.0 to +1.0
- The sign of r tells you whether the relationship between X and Y is a positive (direct) or a negative (inverse) relationship.
- Positive ($+r$) = As X goes up, Y goes up
- *Negative (-r) = As X goes up, Y goes down*



■ **Pearson Correlation Coefficient Formals**

$$r = \frac{\text{covariance of } X \text{ and } Y}{\text{variance of } X \text{ and } Y}$$

Or

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

Or

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

Or, equivalent

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Example 1

If you have $n=8$, $\Sigma X = 38$ $\Sigma Y = 35$ $\Sigma X^2 = 240$ $\Sigma Y^2 = 193$ $\Sigma XY = 209$ compu
Pearson correlation coefficient

Solution

$$r = \frac{n\sum XY - \sum X \cdot \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

$$\therefore r = \frac{8(209) - (38)(35)}{\sqrt{[8(240) - (38)^2][8(193) - (35)^2]}}$$

$$= \frac{1672 - 1330}{\sqrt{(1920 - 1444)(1544 - 1225)}}$$

$$= \frac{342}{\sqrt{(476)(319)}}$$

$$\therefore r = \frac{342}{\sqrt{151844}} = \frac{342}{389.6717} = +0.878$$

There is a significant linear relationship between X and Y.

Example 2

Find the Pearson correlation coefficient using the following information

$\Sigma x = 80$ $\Sigma x^2 = 1,148$ $\Sigma y = 69$ $\Sigma y^2 = 815$ $\Sigma xy = 624$ $n = 7$

Solution

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

$$= \frac{624 - \frac{(80)(69)}{7}}{\sqrt{\left[1{,}148 - \frac{(80)^2}{7}\right]\left[815 - \frac{(69)^2}{7}\right]}}$$

$$= \frac{-164.571}{\sqrt{(233.714)(134.857)}} = \frac{-164.571}{177.533} = -0.927$$

Example 3

If X is the area planted and Y is the quantity of the meat, find the coefficient of correlation between X and Y .

X	305	313	297	289	233	214	240	217
Y	592	603	662	607	635	699	719	747

Solution

$\bar{x} = \frac{\sum x}{n} = \frac{2108}{8} = 263.5$,

$\bar{y} = \frac{\sum y}{n} = \frac{5264}{8} = 658$

x	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})^2$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})$	y
305	-2739	4356	-66	1722.25	41.5	592
313	-2722.5	3025	-55	2450.25	49.5	603
297	134	16	4	1122.25	33.5	662
289	-1300.5	2601	-51	650.25	25.5	607
233	701.5	529	-23	930.25	-30.5	635
214	-2029.5	1681	41	2450.25	-49.5	699
240	-1433.5	3721	61	552.25	-23.5	719
217	-4138.5	7921	89	2162.25	-46.5	747
-13528	23850	0	12040	0	5264	2108

$\sum(x-\bar{x})^2=12040$, $\sum(y-\bar{y})^2=23850$, $\sum(x-\bar{x})(y-\bar{y})=-13528$

Then, the coefficient of correlation is

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$=\frac{-13528}{\sqrt{12040}\sqrt{23850}}$$

$$=\frac{-13528}{(109.727)(154.434)}$$

$$=\frac{-13528}{16945.619}=-0.798$$

There is negative correlation between X and Y

One can use the formula

$$r=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\left[\sum x^2-\frac{(\sum x)^2}{n}\right]\left[\sum y^2-\frac{(\sum y)^2}{n}\right]}}$$

x	y	xy	x^2	y^2
305	592	180560	93025	350464
313	603	188739	97969	363609
297	662	196614	88209	438244
289	607	175423	83521	368449
233	635	147955	54289	403225
214	699	149586	45796	488601
240	719	172560	57600	516961
217	747	162099	47089	558009
2108	5264	1373536	567498	3487562

$\sum x=2108$, $\sum y=5264$, $\sum xy=1373536$ $\sum x^2=567498$ and $\sum y^2=3487562$,

$$r=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\left(\sum x^2-\frac{(\sum x)^2}{n}\right)\left(\sum y^2-\frac{(\sum y)^2}{n}\right)}}$$

$$\therefore r = \frac{1373536 - \frac{(2108)(5264)}{8}}{\sqrt{\left(567498 - \frac{(2108)^2}{8}\right)\left(3487562 - \frac{(5264)^2}{8}\right)}}$$

$$= \frac{-13528}{\sqrt{(12040)(23850)}} = \frac{-13528}{16945.619} = -0.798,$$

which gives the same results.

■ Spearman Correlation Coefficient

Spearman's r is a statistic for measuring the relationship between two variables. It is a nonparametric measure that avoids assumptions that the variables have a straight line relationship and can be used when one or both measures are measured on an ordinal scale. It can have any value between –1 and +1. A value of 0 indicates no relationship and values of +1 or 1 indicate a one to one relationship between the variables or 'perfect correlation'.

1. Rank both sets of data. The ***highest value is ranked first.*** Had there been two or three countries with the same value, they would have been given equal ranking (eg. 1, 2, 3.5, 3.5 [3.5 is the mean of 3 and 4], 5, 7, 7, 7 [7 is the mean of 6, 7, and 8], 9, 10.

2. Calculate the difference, or "d", between the two rankings. Note that it is possible to get negative answers.
3. Calculate "d^2", to eliminate the negative values.
4. Add up ($\sum$) the d^2 values
5. You are now in a position to calculate the correlation coefficient, or "r_S", by using the formula:

$$r_S = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)}$$

r_S = spearman rank,

n = number of samples

$\sum d^2$ = sum of the difference between rank of the values of each matched pair.

Example 4

Calculate spearman's rank correlation coefficient of the following data.

X	125	80	96	65	30	134	54	16	64	72	49
Y	109	76	101	77	27	142	76	12	80	93	82

Solution

X	*Y*	*Rank X*	*Rank Y*	*d*	d^2
125	109	2	2	0	0
80	76	4	8.5	-4.5	20.25
96	101	3	3	0	0
65	77	6	7	-1	1
30	27	10	10	0	0
134	142	1	1	0	0
54	76	8	8.5	-0.5	0.25
16	12	11	11	0	0
64	80	7	6	1	1
72	93	5	4	1	1
49	82	9	5	4	16
					39.5

$$r_S = 1 - \frac{6 \times \sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 3.5}{11(11^2-1)}$$

$$= 1 - \frac{237}{1320} = 0.82$$

The answer of 0.82 shows that there is a strong correlation between the 2 sets of data.

Example 5

The marks of 12 pupils in Statistics and Calculus essays are as follows;

Statistics *(X)*	15	16	19	17	17	15	18	16	18	18	14	10
Calculus *(Y)*	10	12	12	13	11	9	11	13	11	12	8	7

Calculate spearman's rank correlation coefficient.

Solution

First we must rank the data.

Statistics

19 = 1

$18 = \frac{2+3+4}{3} = 3$

17 = ½ (5 + 6) = 5.5

16 = ½ (7 + 8) = 7.5

15 = ½ (9 + 10) = 9.5

14 = 11

10 = 12

Calculus

13 = ½(1+2) = 1.5

$12 = \frac{1}{3}(3+4+5) = 4$

$11 = \frac{1}{3}(6+7+8) = 7$

$10 = 9$

$9 = 10$

$8 = 11$

$7 = 12$

Statistics, x	Calculus, y	Rank, x	Rank, y	$d = x - y$	d^2
15	10	9.5	9	0.5	0.25
16	12	7.5	4	3.5	12.25
19	12	1	4	-3	9
17	13	5.5	1.5	4	16
17	11	5.5	7	-1.5	2.25
15	9	9.5	10	-0.5	0.25
18	11	3	7	-4	16
16	13	7.5	1.5	6	36
18	11	3	7	-4	16
18	12	3	4	-1	1
14	8	11	11	0	0
10	7	12	12	0	0
				$\sum d^2 = 109$	

$$\therefore r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 109}{12(12^2-1)}$$

$$= 1 - \frac{654}{12(144-1)} = 1 - \frac{654}{12 \times 143}$$

$$= 1 - \frac{654}{1716} = \frac{177}{286} = 0.619$$

Some positive correlation between the Statistics and calculus results.

Example 6

These are the marks obtained by 8 pupils in a Maths and Physics. Calculate Spearman's coefficient of rank correlation.

Maths (x)	67	42	85	51	39	97	81	70
Physics (y)	70	59	71	38	55	62	80	76

Solution

Rank x	4	2	7	3	1	8	6	5
Rank y	5	3	6	1	2	4	8	7
d	-1	-1	1	2	-1	4	-2	-2
d^2	1	1	1	4	1	16	4	4

Now

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 32}{8(8^2-1)} = -0.2539$$

Spearman's coefficient of rank correlation is – 0.25.

Example 7

In a study of the relationship between level education (X) and income (Y) the following data was obtained. Find the relationship between them and comment.

X	Preparatory	Primary	University	secondary	secondary	illiterate	University
Y	25	10	8	10	15	50	60

Solution

x	y	*Rank, x*	*Rank, y*	$d = x - y$	d^2
Preparatory	25	5	3	2	4
Primary	10	6	5.5	0.5	0.25
University	8	1.5	7	-5.5	30.25
secondary	10	3.5	5.5	-2	4
secondary	15	3.5	4	-0.5	0.25
illiterate	50	7	2	5	25
University	60	1.5	1	0.5	0.25
					$\sum d^2 = 64$

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 64}{7(7^2-1)} = -0.1.$$

Comment:

There is an indirect weak correlation between level of education and income.

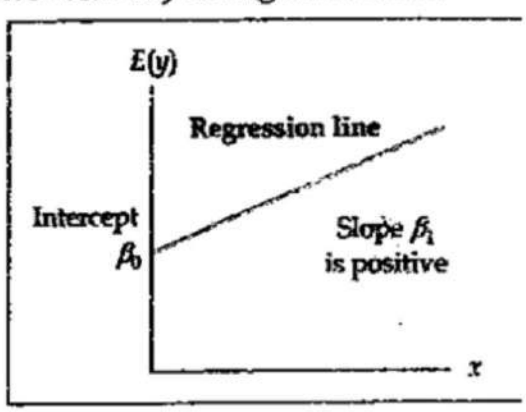
■ **Regression Analysis:** is a statistical procedure used to find relationships among a set of variables. In regression analysis, there is a **dependent variable** (x), which is the one you are trying to explain, and one or more **independent variables** (y) that are related to it. **The equation** that describes how y is related to x and an error term is called the **regression model**. You can express the relationship as a linear equation (**simple linear regression model**), such as:

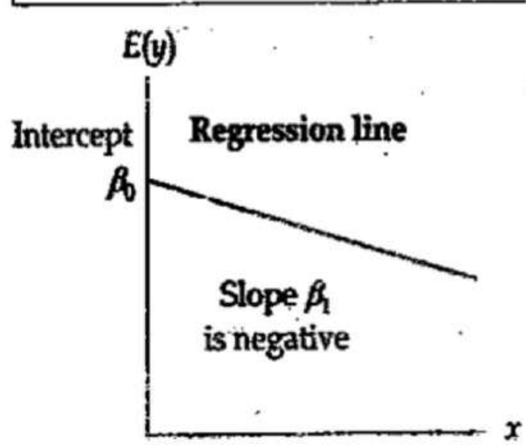
$$y = \beta_0 + \beta_1 x + \varepsilon,$$

- y is the **dependent** variable
- x is the **independent** variable
- β_0 and β_1 **are called** <u>parameters of the model</u> and ε is a random variable called the <u>error term</u>.
- β_0 is a **constant**
- β_1 is the **slope of the** line

The simple **linear regression** equation is: $E(y) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- Graph of the **regression** equation is a straight line.
- $\hat{\beta}_0$ is the **y intercept** of the regression line.
- $\hat{\beta}_1$ is the **slope of the** regression line.
- $E(y) = \hat{y}$ is **the expected** value of y for a given x value.





- For every increase of 1 in x , y changes by an amount equal to b .
- The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.
- Simple regression fits a straight line to the data.
- The observation is denoted by y and the prediction is denoted by y'.
- ε is the prediction error.
-

$$Slope: \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$Or: \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}},$$

where $SS_{xy} = \Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$

$$SS_{xx} = \Sigma\left(x_i - \bar{x}\right)^2$$

Intercept : $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ *and* n = sample size.

■ ***Interpreting the Estimates of β_0 and β_1 in Simple Liner Regression***

- Intercept: β_0 represents the predicted value of y when $x = 0$.

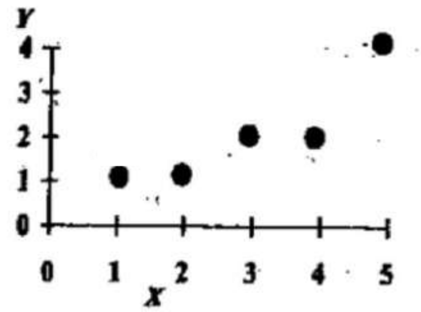
- Slope: β_1 represents the increase (or decrease) in y for every 1-unit increase in x.

Example 7

Calculate the regression line equation of ***y on x*** for the following data.

x	1	2	3	4	5
y	1	1	2	2	4

Solution

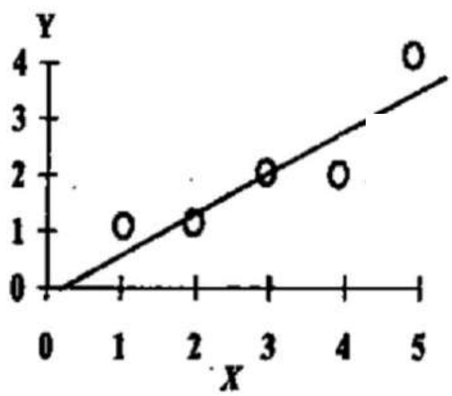


x_i	y_i	$x_i y_i$	x_i^2
1	1	1	1
2	1	2	4
3	2	6	9
4	2	8	16
5	4	20	25
15	9	37	55

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = .70$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - (.70)(3) = -.10$$

$$\hat{y} = -0.1 + 0.7x$$



Example 8

From the following data:

1- Find the coefficient of correlation and plot the scatter plot of the data.

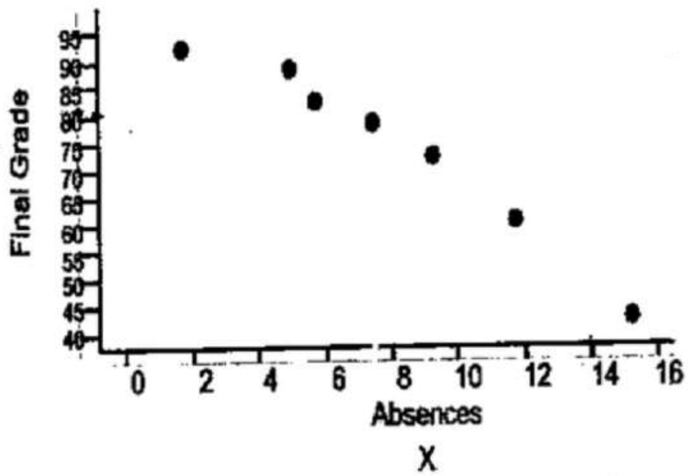
2- Write the equation of the line of regression with x = number of absences and y = final grade.

3- Use this equation to predict the expected grade for a student with

(a) 3 absences (b) 12 absences

Absences x	8	2	5	12	15	9	6
Final Grade y	78	92	90	85	73	74	81

Solution



x_i	y_i	$x_i y_i$	x_i^2	y_i^2
8	78	624	64	6084
2	92	184	4	8464
5	90	450	25	8100
12	58	696	144	3364
15	43	645	225	1849
9	74	666	81	5476
6	81	486	36	6561
57	**516**	**3751**	**579**	**39898**

Pearson Correlation Coefficient: $r = \dfrac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

$$\therefore r = \frac{7(3751)-(57)(516)}{\sqrt{[7(579)-(57)^2][7(39898)-(516)^2]}}$$

$$= \frac{-3155}{\sqrt{804}\sqrt{13030}} = -0.975$$

Now the regression line equation of *y* on x:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{7(3751)-(57)(516)}{7(579)-(57)^2} = -3.924$$

$\bar{y} = \frac{516}{7} = 73.714, \bar{x} = \frac{57}{7} = 8.143,$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 73.714 - (-3.924)(8.143) = 105.667$

The regression equation for number of times absent and final grade is:

$$\hat{y} = 105.667 - 3.924x$$

a) 3 absences

$$\hat{y} = -3.924(3) + 105.667 = 93.895$$

b) 12 absences

$$\hat{y} = -3.924(12) + 105.667 = 58.579$$

Exercises

1. The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Find the equation of the regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report but was ill at the time of the final examination.

2. (a) From the following information draw a scatter diagram and draw the regression line of best fit.

Volume of sales (thousand units)	5	6	7	8	9	10
Total expenses (thousand $)	74	77	82	86	92	95

(b) What will be the total expenses when the volume of sales is 7,500 units?

(c) If the selling price per unit is $11, at what volume of sales will the total income from sales equal the total expenses?

5. Compute and interpret the correlation coefficient for the following grades of 6 students selected at random.

Math. grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

The following table below shows a traffic-flow index and the related site costs in respect of eight service stations of ABC Garages Ltd.

Site No.	Traffic-flow index	Site cost (in 1000)
1	100	100
2	110	115
3	119	120
4	123	140
5	123	135
6	127	175
7	130	210
8	132	200

(a) Calculate the coefficient of correlation for this data.

(b) Calculate the coefficient of rank correlation.

Hypothesis Testing

1. Sampling

One of the major problems in statistics, estimating the properties of a large population from the properties of a **sample** of individuals chosen from that population, is considered in this section. Select at random a sample of n observations $X_1, X_2, \dots X_n$ taken from a population. From these n observations you can calculate the values of a number of statistical quantities, for example the sample mean X. If you choose another random sample of size n from the same population, a different value of the statistic will, in general, result. In fact, if repeated random samples are taken, you can regard the statistic itself as a random variable, and its distribution is called the **sampling distribution** of the statistic.

For example, consider the distribution of heights of all adult men in England, which is known to conform very closely to the normal curve. Take a large number of samples of size four, drawn at random from the population, and calculate the mean height of each sample. How will these mean heights be distributed? We find that they are also normally distributed { about the same mean as the original distribution. However, a random sample of four is likely to include men both above and below average height and so the mean of the sample will deviate from the true mean less than a single observation will. This important general result can be stated as follows:

If random samples of size n are taken from a distribution whose mean is μ_x and whose standard deviation is μ_x , then the sample means form a distribution with mean σ_x and standard deviation $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.

Note that the theorem holds for all distributions of the parent population. However, if the parent distribution is normal, then it can be shown that the sampling distribution of the sample mean is also normal.

The standard deviation of the sample mean, $\sigma_{\bar{x}}$ defined above, is usually called the standard error of the sample mean.

Let us now present three worked examples.

Ex 5.

A random sample is drawn from a population with a known standard deviation of 2. Find the standard error of the sample mean if the sample is of size (i) 9, (ii) 100.
What sample size would give a standard deviation equal to 0.5?

Using the result stated earlier

(i) standard deviation = $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{2}{\sqrt{9}} 0.667$

(ii) standard deviation = $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$

If the standard error equals 0.5, then $\frac{2}{\sqrt{n}} = 0.5$

Squaring then implies that $\frac{4}{n} = 0.25$ or n = 16,

(i.e. the sample size is 16).

Ex 6.

The diameters of shafts made by a certain manufacturing process are known to be normally distributed with mean 2.500cm and standard deviation 0.009 cm. What is the distribution of the sample mean diameter of nine such shafts selected at random? Calculate the percentage of such sample means which can be expected to exceed 2.506 cm.

Solution:
Since the process is normal we know that the sampling distribution of the sample mean will also be normal, with the same mean, 2.500cm, but with a standard error (or standard deviation) $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.009}{\sqrt{9}} = 0.003$ cm.

In order to calculate the probability that the sample mean is bigger than 2.506, i.e. $\bar{X} > 2.506$, we standardise in the usual way by putting $Z = \frac{\bar{X} - 2.5}{0.003}$, and then

$$P\left(\bar{X} > 2.506\right) = P\left(\frac{\bar{X} - 2.5}{0.003} > \frac{2.506 - 2.5}{0.003}\right) = P(Z > 2)$$

$$= 1 - P(Z \leq 2)$$

$$= 1 - 0.9772 = 0.0228$$

Hence, 2.28% of the sample means can be expected to exceed 2.506 cm.

It was stated above that when the parent distribution is normal then the sampling distribution of the sample mean is also normal. When the parent distribution is not normal, then obtain the following theorem (surprising result?):

2- Central limit theorem:

If a random sample of size n; ($n \geq 30$); is taken from ANY distribution with mean μ_x and standard deviation σ_x, then the sampling distribution of X is approximately normal with mean μ_x and standard deviation $\frac{\sigma_x}{\sqrt{n}}$, the approximation improving as n increases.

Ex 8.

It is known that a particular make of light bulb has an average life of 800 hrs with a standard deviation of 48 hrs. Find the probability that a random sample of 144 bulbs will have an average life of less than 790 hrs

Solution:

Since the number of bulbs in the sample is large, the sample

mean will be normally distributed with mean = 800 and

standard error $\sigma_{\overline{X}} = \dfrac{48}{\sqrt{144}} = 4$. Put $Z = \dfrac{\overline{X} - \mu_X}{\sigma_X} = \dfrac{(\overline{X} - 800)}{4}$, then

$$P(\overline{X} < 790) = P\left(\frac{\overline{X} - 800}{4} < \frac{790 - 800}{4}\right)$$
$$= P(Z < -2.5) = P(Z > 2.5), \quad \text{by symmetry}$$
$$= 1 - P(Z \le 2.5) = 1 - 0.9938 = \underline{0.0062}.$$

To conclude this section the main results concerning the distribution of the sample mean $\overline{X}$ are summarised. Consider a parent population with mean μ_X and standard deviation σ_X. From this population take a random sample of size n with sample mean $\overline{X}$ and standard error $\sigma_X/\sqrt{n}$. Define $Z = \dfrac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}}$ then

Ex 9.

The percentage of copper in a certain chemical is to be estimated by taking a series of measurements on randomly chosen small quantities of the chemical and using the sample mean to estimate the true percentage. From previous experience individual measurements of this type are known to have a standard deviation of 2 %. How many measurements must be made so that the standard deviation of the estimate is less than 0.3%? If the sample mean w of 45 measurements is found to be 12.91%, give a 95% confidence interval for the true percentage, w.

Solution:

Assume that n measurements are made. The standard error of the sample mean is $(\frac{2}{\sqrt{n}})$%. For the required precision require

$\frac{2}{\sqrt{n}} < 0.3$, i.e. $n > \left(\frac{2}{0.3}\right)^2 = \frac{4}{0.9} = 44.4$.

Since n must be an integer, at least 45 measurements are necessary for required precision.

With a sample of 45 measurements, you can use the central limit theorem and take the sample mean percentage W to be distributed normally with mean ω and standard deviation $\frac{2}{\sqrt{45}}$.

Hence, if ω is the true percentage, it follows that $Z = \frac{W - \omega}{\frac{2}{\sqrt{45}}}$

is distributed as N(0, 1). Since 95% of the area under the standard normal curve lies between $Z = -1.96$ and $Z = 1.96$,

$$P\left(-1.96 \le \frac{W-\omega}{2/\sqrt{45}} \le 1.96\right) = 0.95.$$

Re-arranging, we obtain $P\left(W - 1.96\left(\frac{2}{\sqrt{45}}\right) \le \omega \le W + 1.96\left(\frac{2}{\sqrt{45}}\right)\right) = 0.95.$

Hence, the 95% confidence interval for the true percentage is

$$(12.91 - 1.96(0.298),\ \ 12.91 + 1.96(0.298)) = (12.33,\ \ 13.49).$$

To complete this section we define the sample variance.

Def. Given a sample of n observations
$X_1, X_2, \ldots, X_n$ the **sample variance**, S^2, is given by
$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$, where X denotes the sample mean.
In our discussion of confidence intervals for the mean it was assumed that the population variance σ_x^2 was known. What happens when this it is not known? For samples of size $n > 30$, a good estimate of σ_x^2 is obtained by calculating the sample variance S^2 and using this value. (For small samples, $n < 30$, need to use the t-distribution – not considered in this module).

3. Hypothesis testing

An assumption made about a population is called a statistical hypothesis. From information contained in a random sample we try to decide whether or not the hypothesis is true:
if evidence from the sample is inconsistent with the hypothesis, then the hypothesis is rejected;
if the evidence is consistent with the hypothesis, then the hypothesis is accepted.
The hypothesis being tested is called the null hypothesis (usually denoted by H_0) – it either specifies a particular value of the population parameter or specifies that two or more parameters are equal.

A contrary assumption is called the alternative hypothesis (usually denoted by H_1) – normally specifies a range of values for the parameter.
A common example of the null hypothesis is H_0: $\mu_x = \mu_o$.
Then three alternative hypotheses are

$$(i)\ H_1 : \mu_x > \mu_0 \quad (ii)\ H_1 : \mu_x < \mu_0 \quad (iii)\ H_1 : \mu_x \neq \mu_0$$

Types (i) and (ii) are said to be one-sided (or one-tailed, see figure 6b) – type (iii) is two-sided (or two-tailed, see figure 6a).

The result of a test is a decision to choose H_0 or H_1. This decision is subject to uncertainty, and two types of error are possible:

(i) a type I error occurs when we reject H_0 on the basis of the test although it happens to be true – the probability of this happening is called the level of significance of the test and this is prescribed before testing – most commonly chosen values are 5% or 1%.

(ii) a type II error occurs when you accept the null hypothesis on the basis of the test although it happens to be false.

The above ideas are now applied to determine whether or not the mean, $\bar{X}$, of a sample is consistent with a specified population mean μ_0. The null hypothesis is $H_0 : \mu_x = \mu_0$ and a suitable statistic to use is $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma_x}{\sqrt{n}}}$, where

σ_x^2 is the standard deviation of the population and n is the size of the sample.

Find the range of values of Z for which the null hypothesis would be accepted – known as acceptance region for the test – depends on the pre-determined signi cance level and the choice of H_1.

Corresponding range of values of Z for which H_0 is rejected (i.e. not accepted) is called the rejection region.

Ex 10.

A standard process produces yarn with mean breaking strength 15.8 kg and standard deviation 1.9 kg. A modification is introduced and a sample of 30 lengths of yarn produced by the new process is tested to see if the breaking strength has changed. The sample mean breaking strength is 16.5 kg. Assuming the standard deviation is unchanged, is it correct to say that there is no change in the mean breaking strength?

Solution:

Here $H_0 : \mu_x = \mu_0$, $H_1 : \mu_x \neq \mu_0$,
where $\mu_0 = 15.8$ and μ_x is the mean breaking strength for the new process.

If H_0 is true (i.e. $\mu_X = \mu_0$), then $Z = \dfrac{\overline{X} - \mu_0}{\sigma_X / \sqrt{n}}$ has approximately the $N(0,1)$ distribution, where $\overline{X}$ is the mean breaking strength of the 30 sample values and $n = 30$.

At the 5% significance level there is a rejection region of 2.5% in each tail, as shown in figure 6a (since, under H_0,

$$P(Z < -1.96) = P(Z > 1.96) = 1 - P(Z \leq 1.96) = 1 - \Phi(1.96) = 0.025, \quad \text{i.e. } 2.5\%).$$

This is an example of a two-sided test leading to a two-tailed rejection region.

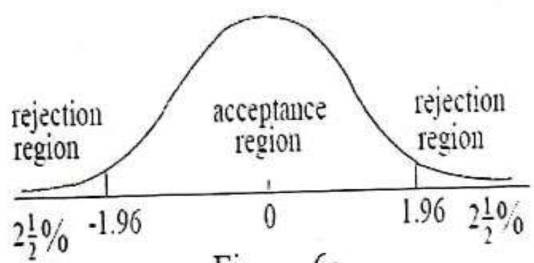


Figure 6a

The test is therefore: accept H_0 if $-1.96 \leq Z \leq 1.96$, otherwise reject.

From the data, $Z = \dfrac{16.5 - 15.8}{1.9/\sqrt{30}} = 2.018$. Hence, H_0 is rejected at the 5% significance level: i.e. the evidence suggests that there IS a change in the mean breaking strength.

Let us now consider a slightly differently worded question.

Suppose the modification was specifically designed so as to increase the strength of the yarn. In this case

$H_0 : \mu_x = \mu_0, \quad H_1 : \mu_x > \mu_0,$

and H_0 is rejected if the value of Z is unreasonably large. In this situation the test is one-sided and acceptance and (one-tailed) rejection regions at the 5% significance level are shown below.

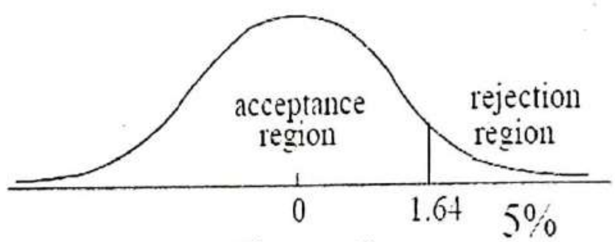


Figure 6b

At 5 % significance level, test is accept H_0 if $Z < 1.64$, otherwise reject:

From earlier work $Z = 2.018$ and again the null hypothesis is rejected.

[Compare the two diagrams above, which illustrate the statement that the rejection region for a test depends on the form of both the alternative hypothesis and the significance level.]

Example 11:

Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15.
A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect, $\alpha = 0.05$.

Solution:

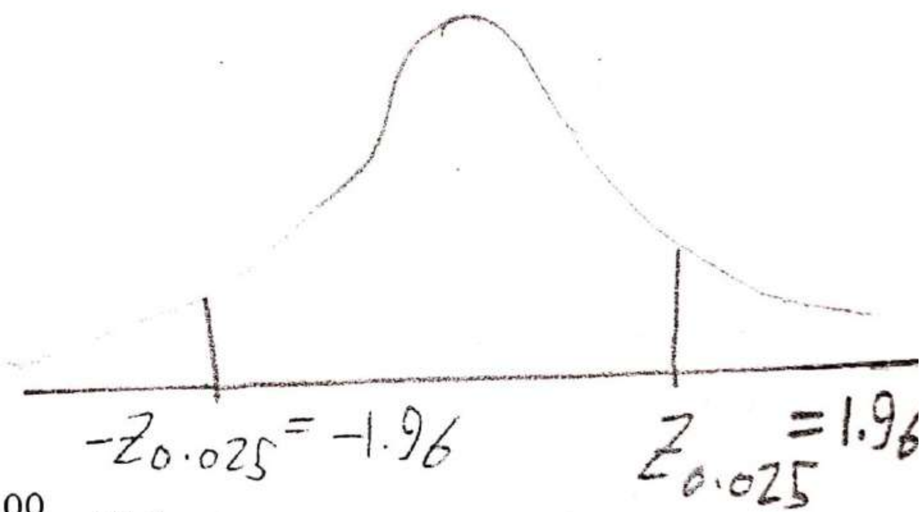
$H_0 : \mu = 100,$

$H_1 : \mu \neq 100$

$n = 30,\ \sigma = 15.$

we use two tailed test

$\frac{\alpha}{2} = 0.025$



$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{140 - 100}{15/\sqrt{30}} = 14.6$$

$Z = 14.6$ *is greater than* $Z_{0.025} = 1.96.$

So we reject the Null hypothesis H_0,
and accept the Alternative hypothesis H_1.

t student's distribution

Introduction

When sample sizes are small, and the standard deviation of the population is unknown it is normal to use the distribution of the t statistic (also known as the t score), whose values are given by:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where $\bar{x}$ is the sample mean, μ_0 is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the t statistic is called the t distribution or student's t distribution.

Example 12:

The maker of a certain model car claimed that his car averaged at least 31 miles per gallon of gasoline. A sample of nine cars was selected and each car was driven with one gallon of regular gasoline. The sample showed a mean of 29.43 miles with a standard deviation of 3 miles. $\alpha = 0.05$ What do you conclude about the manufacturers claim?

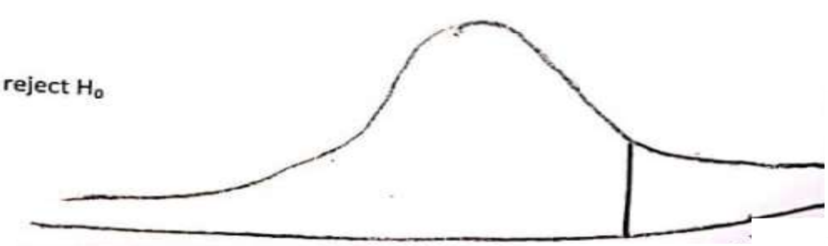
Solution:

$H_0 : \mu = 31,$

$H_1 : \mu \succ 31$

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{29.43 - 31}{\frac{3}{\sqrt{9}}} = -1.57$$

And $t_{0.05,8} = 1.86$ then do not reject H_0



Example 13 :

If we have a factory for car batteries, the owner of the factory thinks that the average age of these batteries is 36 months. In order to test of this claim, a random sample of 10 batteries was selected and measured the age in months, as follows:

27.6	28.7	34.7	29	22.9	29.6	29.4	30.2	36.5	34.7

Do these data show that the average age of these batteries as less than 36 months?

Solution:

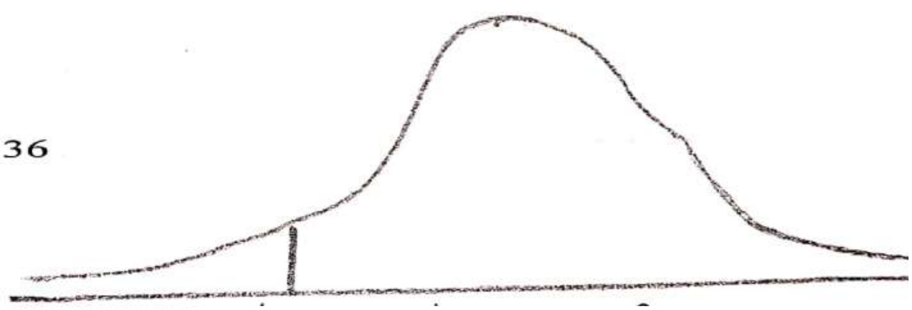
$H_o : \mu = 36,$

$H_1 : \mu < 36$

$$t = \frac{\bar{X} - \mu_o}{\frac{s}{\sqrt{n}}}$$

$n = 10 < 30, \qquad \mu_o = 36$

$$\bar{X} = \frac{\sum X_i}{10} = 30.33$$



$$S = \sqrt{\frac{\sum_{i=1}^{10}(X_i - \bar{X})^2}{n-1}} = 4.011$$

we use

$$t = \frac{\bar{X} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{30.33 - 36}{\frac{4.011}{\sqrt{10}}} = -4.47$$

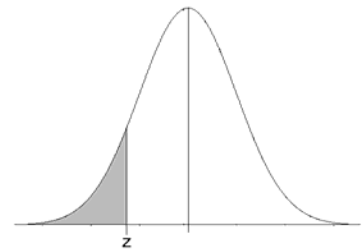
we have :

$t(9,\ 0.01) = 2.821$

$\therefore \quad t = -4.47 < -2.821$

So we reject H_o, and accept H_1

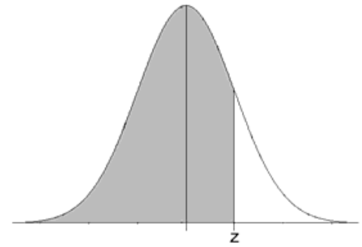
Standard Normal Cumulative Probability Table



Cumulative probabilities for NEGATIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

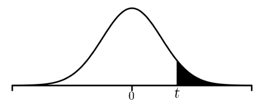
Standard Normal Cumulative Probability Table



Cumulative probabilities for POSITIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Critical Values for Student's t-Distribution.



	Upper Tail Probability: $\Pr(T > t)$									
df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
31	0.853	1.309	1.696	1.810	1.952	2.040	2.144	2.453	2.744	3.633
32	0.853	1.309	1.694	1.808	1.950	2.037	2.141	2.449	2.738	3.622
33	0.853	1.308	1.692	1.806	1.948	2.035	2.138	2.445	2.733	3.611
34	0.852	1.307	1.691	1.805	1.946	2.032	2.136	2.441	2.728	3.601
35	0.852	1.306	1.690	1.803	1.944	2.030	2.133	2.438	2.724	3.591
36	0.852	1.306	1.688	1.802	1.942	2.028	2.131	2.434	2.719	3.582
37	0.851	1.305	1.687	1.800	1.940	2.026	2.129	2.431	2.715	3.574
38	0.851	1.304	1.686	1.799	1.939	2.024	2.127	2.429	2.712	3.566
39	0.851	1.304	1.685	1.798	1.937	2.023	2.125	2.426	2.708	3.558
40	0.851	1.303	1.684	1.796	1.936	2.021	2.123	2.423	2.704	3.551
41	0.850	1.303	1.683	1.795	1.934	2.020	2.121	2.421	2.701	3.544
42	0.850	1.302	1.682	1.794	1.933	2.018	2.120	2.418	2.698	3.538
43	0.850	1.302	1.681	1.793	1.932	2.017	2.118	2.416	2.695	3.532
44	0.850	1.301	1.680	1.792	1.931	2.015	2.116	2.414	2.692	3.526
45	0.850	1.301	1.679	1.791	1.929	2.014	2.115	2.412	2.690	3.520
46	0.850	1.300	1.679	1.790	1.928	2.013	2.114	2.410	2.687	3.515
47	0.849	1.300	1.678	1.789	1.927	2.012	2.112	2.408	2.685	3.510
48	0.849	1.299	1.677	1.789	1.926	2.011	2.111	2.407	2.682	3.505
49	0.849	1.299	1.677	1.788	1.925	2.010	2.110	2.405	2.680	3.500
50	0.849	1.299	1.676	1.787	1.924	2.009	2.109	2.403	2.678	3.496
60	0.848	1.296	1.671	1.781	1.917	2.000	2.099	2.390	2.660	3.460
70	0.847	1.294	1.667	1.776	1.912	1.994	2.093	2.381	2.648	3.435
80	0.846	1.292	1.664	1.773	1.908	1.990	2.088	2.374	2.639	3.416
90	0.846	1.291	1.662	1.771	1.905	1.987	2.084	2.368	2.632	3.402
100	0.845	1.290	1.660	1.769	1.902	1.984	2.081	2.364	2.626	3.390
120	0.845	1.289	1.658	1.766	1.899	1.980	2.076	2.358	2.617	3.373
140	0.844	1.288	1.656	1.763	1.896	1.977	2.073	2.353	2.611	3.361
180	0.844	1.286	1.653	1.761	1.893	1.973	2.069	2.347	2.603	3.345
200	0.843	1.286	1.653	1.760	1.892	1.972	2.067	2.345	2.601	3.340
500	0.842	1.283	1.648	1.754	1.885	1.965	2.059	2.334	2.586	3.310
1000	0.842	1.282	1.646	1.752	1.883	1.962	2.056	2.330	2.581	3.300
∞	0.842	1.282	1.645	1.751	1.881	1.960	2.054	2.326	2.576	3.291
	60%	80%	90%	92%	94%	95%	96%	98%	99%	99.9%
	Confidence Level									

Note: $t(\infty)_{\alpha/2} = Z_{\alpha/2}$ in our notation.

References

1-Introduction to Probability and Statistics , Mendenhall, L. North Scituate, Duxbury(1980).

2- Basic Statistics, Goon, A.M., Gupta, M.K. , Gupta, B.D, Calcutta, (1991).

3- Probability and statistics for Engineers and Scientists, Walpole, R.E. ; Myers, R.H. and Myers, S.L., Prentice Hall, London (1980).

4- Mathematical statistics, John, Freund and Ronald, Walpole, Prentice Hall, Englewood Cliffs, N.J., 1987.

5- An introduction to probability theory and its applications, Feller, W. , Third ed. New York, John Wiley and Sons (1968).

6- Introduction to Mathematical Statistical, Fifth ed. Hoel. P.G. New York, John Wiley and Sons (1984).

7- Introduction to Mathematical Statistical , Hogg, R. and Craig, A. T., Fourth ed. New York, MacMillan Publishing company, (1978).

8- Probability and Statistical inference, Hogg, R. and Elliot, T., Third ed. New York, MacMillan Publishing company, (1988).

9- Introduction to Theory of Statistics, Mood, A.M., Graybill, F.A and Boes, D. C., Third ed McGraw –Hill Inc. (1974).

10- A First Course in probability, Ross, S. Fourth ed. New York, MacMillan Publishing company, (1988).

11- Probability and Statistics with applications, Strait, P. T. New York, Harcaurt Brace Jovanovich Inc. (1983)