Principles of Statistics

Dr. AbuBakr A. AbdulMotaal

Faculty of Commerce
South Valley University

Department of
Quantitative Methods

Principles of Statistics

Dr. AbuBakr A. AbdulMotaal

Faculty of Commerce
South Valley University
Department of Quantitative Methods

2022 / 2023

Preface

In this book “Principles of Statistics”, mathematical background needed is basic arithmetic and basic elements of algebra. The primary purpose of the book is to take the mystery out of the subject matter and to present and explain this field of study in a manner which captures the student’s imagination in utilizing the statistical tools for the purpose of business decision making. The text has been written for facilitating usage by all business and economics majors.

Each topic in each chapter is explained by use of solved examples within the chapter so as to demonstrate the applicability of statistical tools described and learned in the chapter. There are additional unsolved problems at the end of each chapter. Unsolved problems are added at the end of each chapter so that the students acquire a reasonable degree of specialization in statistical thinking, decision making and problem solving. The total numbers of solved examples and unsolved problems in the book are 72 and 76 respectively.

The book covers various aspects of statistics in six chapters. All of these chapters deal with descriptive statistics. In the first four chapters, various statistical terms and concepts are explained and analysis of frequency distribution and associated measures of central tendency and dispersion are discussed and explained in details. Chapters five and six cover the area of simple correlation and regression which deal with the strength of relationships among different but dependent variables.

AbuBakr A. AbdulMotaal

Contents

Chapter (1)
Introduction

1.1 What is Statistics:

The word statistics has two meanings. In the more common usage, statistics refers to numerical facts. The numbers that represent the income of a family, the age of a students, etc. The second meaning of statistics refers to the field or discipline of study. In this sense of the word, statistics is defined as follows

Statistics:
Is a group of methods used to collect, analyze and interpret data in order to make decisions.

1.2 Types of Statistics:

Broadly speaking, applied statistics can be divided into two areas: descriptive statistics and inferential statistics.

Suppose we have information on the test scores of students enrolled in a statistics class. In statistical terminology, the whole set of numbers that represents the scores of students is called a data set, the name of each student is called an element, and the score of each student is called an observation. Many data sets in their original forms are usually very large, especially those collected by state agencies. Consequently, such data sets are not very helpful in drawing conclusions or making decisions. It is easier to draw conclusions from summary tables and diagrams than from the original version of a data set. So, we summarize data by constructing tables, drawing graphs, or calculating summary measures such as averages. The portion of statistics that helps us do this type of statistical analysis is called descriptive statistics.

Descriptive Statistics:
Consists of methods for organizing, displaying (presenting), and describing data using tables, graphs and summary measures.

In statistics, the collection of all elements of interest is called a population. The selection of a portion of the elements from this population is called a sample. A major portion of statistics deals with making decisions, inferences, predictions, and forecasts about populations based on results obtained from samples. The area of statistics that deals with such decision-making procedures is referred to as inferential statistics. This branch of statistics is also called inductive reasoning or inductive statistics.

Inferential Statistics:
Consists of methods that use sample results to make decisions about a population.

Population:
Consists of all elements, individuals, items or objects whose characteristics are being studied.

Sample:
A portion of population or few elements selected from a population.

The purpose of conducting a sample survey is to make decisions about the corresponding population. It is important that the results obtained from a sample survey closely match the results that we would obtain by conducting a census. Otherwise, any decision based on a sample survey will not apply to the corresponding population. Such a sample is called a representative sample. Inferences derived from a representative sample will be more reliable.

Representative Sample:
A sample that represents the characteristics of the population as closely as possible.

Sometimes it is impossible to conduct a census. First, it may not be possible to identify and access each member of the population. For example, if a researcher wants to conduct a survey about homeless people, it is not possible to locate each member of the population and include him or her in the survey. Second, sometimes conducting a survey means destroying the items included in the survey. For example, to estimate the mean life of lightbulbs would necessitate burning out all the bulbs included in the survey. The same is true about finding the average life of batteries. In such cases, only a portion of the population can be selected for the survey.

Depending on how a sample is drawn, it may be a random sample or a nonrandom sample.

Random Sample:
A sample is drawn in such a way that each element has some chance of being selected.

From a given population, we can select a large number of samples of the same size. If each of these samples has the same probability of being selected, then it is called simple random sampling.

Simple Random Sample:
If each element has the same chance of being selected, the sample is called a “simple random sample”.

1.3 Basic Terms:

It is very important to understand the meaning of some basic terms that will be used frequently in this text. This section explains the meaning of an element (or member), a variable, a constant, an observation, and a data set.

Element:
Is a specific subject, or object about which the information is collected.

Variable:
Is a characteristic under study that assumes different values for different elements.

Constant:
The value of the constant is fixed.

Observation:
The value of a variable for an element is called an observation or measurement.

Data Set:
Is a collection of observations on one or more variables.

1.4 Types of Variables:

Some variables (such as the price of a home) can be measured numerically, whereas others (such as hair color) cannot. The price of a home is an example of a quantitative variable while hair color is an example of a qualitative variable.

A- Quantitative Variables:

Quantitative Variable:
A variable that can be measured numerically.

Income, height, gross sales, price of a home, number of cars owned, and number of accidents are examples of quantitative variables because each of them can be expressed numerically. Such quantitative variables may be classified as either discrete variables or continuous variables.

Types of Quantitative Variables:

The values that a certain quantitative variable can assume may be countable or noncountable. For example, we can count the number of cars owned by a family, but we cannot count the height of a family member, as it is measured on a continuous scale. A variable that assumes countable values is called a discrete variable. Note that there are no possible intermediate values between consecutive values of a discrete variable. Some variables assume values that cannot be counted, and they can assume any numerical value between two numbers. Such variables are called continuous variables.

(i) Discrete Variable:

Discrete Variable:
A variable whose values are countable. The discrete variable can assume only certain values with no intermediate values.

For example, the number of cars sold on any given day at a car dealership is a discrete variable because the number of cars sold must be 0, 1, 2, 3, . . . and we can count it. The number of cars sold cannot be between 0 and 1, or between 1 and 2. Other examples of discrete variables are the number of people visiting a bank on any day, the number of cars in a parking lot, the number of cattle owned by a farmer, and the number of students in a class.

(ii) Continuous Variable:

Continuous Variable:
A variable that can assume any numerical value over a certain interval.

The time taken to complete an examination is an example of a continuous variable because it can assume any value, let us say, between 30 and 60 minutes. The time taken may be 42.6 minutes, 42.67 minutes, or 42.674 minutes. (Theoretically, we can measure time as precisely as we want.) Similarly, the height of a person. Neither time nor height can be counted in a discrete fashion.

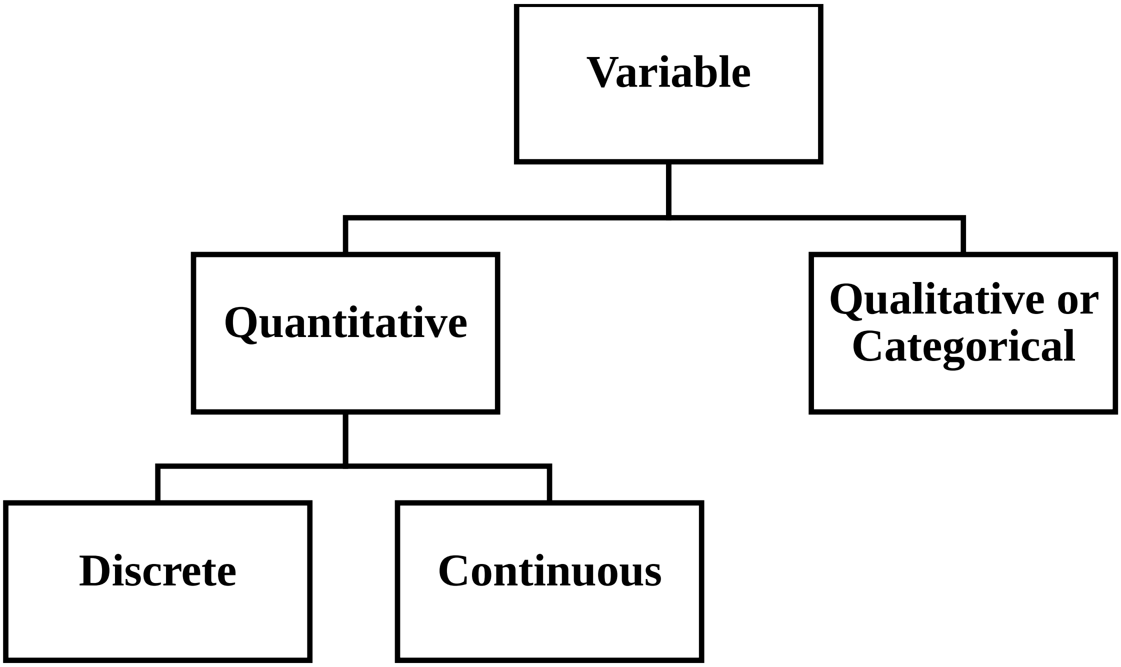
Note that any variable that involves money and can assume a large number of values is typically treated as a continuous variable.

B- Qualitative or Categorical Variables:

Qualitative or Categorical Variable:
A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories.

Qualitative data:
The data collected on a qualitative variable are called qualitative data.

For example, the status of an undergraduate college student is a qualitative variable because a student can fall into any one of four categories: freshman, sophomore, junior, or senior. Other examples of qualitative variables are the gender of a person, the opinions of people, nationality, … etc.


Variable
Quantitative
Qualitative or Categorical
Discrete
Continuous

Problems

1- Briefly describe the two meanings of the word statistics.

2- Briefly explain the types of statistics.

3- Briefly explain the terms: population, sample, representative sample.

4- The following table lists the number of deaths by cause

Cause of Death	Number of Deaths
Heart disease	611,105
Cancer	584,881
Accidents	130,557
Stroke	128,978
Alzheimer's disease	84,767
Diabetes	75,578
Influenza and Pneumonia	56,979
Suicide	41,149

Briefly explain the meaning of an element, a variable, an observation, and a data set with reference to the information in this table.

5- Explain the meaning of the following terms.

a. Quantitative variable
b. Qualitative variable
c. Discrete variable
d. Continuous variable

6- Indicate which of the following variables are quantitative and which are qualitative.

a. The amount of time a student spent studying for an exam.
b. The amount of rain last year in 30 cities.

c. The arrival status of an airline flight (early, on time, late, canceled) at an airport.

d. A person’s blood type.

e. The amount of gasoline put into a car at a gas station.

f. The number of customers in the line waiting for service at a bank at a given time

7- A survey of families living in a certain city was conducted to collect information on the following variables: age of the oldest person in the family, number of family members, number of males in the family, number of females in the family, whether or not they own a house, income of the family, whether or not the family took vacations during the past one year, whether or not they are happy with their financial situation, and the amount of their monthly mortgage or rent.

a. Which of these variables are qualitative variables?

b. Which of these variables are quantitative variables?

c. Which of the quantitative variables of Part b are discrete variables?

d. Which of the quantitative variables of Part b are continuous variables?

Chapter (2)
Organizing and Graphing Data

In addition to hundreds of private organizations and individuals, a large number of government agencies conduct hundreds of surveys every year. The data collected from each of these surveys fill hundreds of thousands of pages. In their original form, these data sets may be so large that they do not make sense to most of us. Descriptive statistics, however, supplies the techniques that help to condense large data sets by using tables, graphs, and summary measures. At a glance, these tabular and graphical displays present information on every aspect of life. Consequently, descriptive statistics is of immense importance because it provides efficient and effective methods for summarizing and analyzing information.

This chapter explains how to organize and display data using tables and graphs. We will learn how to prepare frequency distribution tables for qualitative and quantitative data; how to construct bar graphs, pie charts, and histograms.

2.1 Organizing and Graphing Qualitative Data:

This section discusses how to organize and display qualitative (or categorical) data. Data sets are organized into tables and displayed using graphs. First, we discuss the concept of raw data.

When data are collected, the information obtained from each member of a population or sample is recorded in the sequence in which it becomes available. This sequence of data recording is random and unranked. Such data, before they are grouped or ranked, are called raw data.

2.1.1 Raw Data:

Definition:

Raw Data: Data recorded in the sequence in which they are collected and before they are processed or ranked are called “raw data”.

Suppose we collect information on the ages (in years) of 50 students selected from a university. The data values, in the order they are collected, are recorded in Table 2.1. For instance, the first student’s age is 21, the second student’s age is 19 (second number in the first row), and so forth. The data in Table 2.1 are quantitative raw data.

Table 2.1
Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Suppose we ask the same 50 students about their student status. The responses of the students are recorded in Table 2.2. In this table, F, SO, J, and SE are the abbreviations for freshman, sophomore, junior, and senior, respectively. This is an example of qualitative (or categorical) raw data.

The data presented in Tables 2.1 and 2.2 are also called ungrouped data. An ungrouped data set contains information on each member of a sample or population individually. If we rank the data of Table 2.1 from the lowest to the highest age, they will still be ungrouped data but not raw data.

Table 2.2
Status of 50 students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

2.1.2 Frequency Distributions for Qualitative Data:

A polling agency recently surveyed randomly selected 1015 adults aged 18. These adults were asked, “Please tell me how concerned you are right now about each of the following financial matters, based on your current financial situation—are you very worried, moderately worried, not too worried, or not worried at all.” Among a series of financial situations, one such situation was not having enough money to pay their normal monthly bills. Table 2.3 lists the responses of these adults. The resulting report contained the percent of adults belonging to each category, which we have converted to enough money to pay normal monthly bills. The categories representing this variable are listed in the first column of the table. Note that these categories are mutually exclusive. In other words, each of the 1015 adults belongs to one and only one of these categories. The number of adults who belong to a certain category is called the frequency of that category. A frequency distribution exhibits how the frequencies are distributed over various categories.

Table 2.3 is called a frequency distribution table or simply a frequency table.

Table 2.3
Worries about not having enough money to pay normal monthly bills

Response	Adults
Very worried	162
Moderately worried	203
Not too worried	305
Not worried at all	325
Others	20
Total	1015

Frequency Distribution of a Qualitative Variable:

Definition:

A Frequency Distribution of a Qualitative Variable:
A frequency distribution of a qualitative variable lists all categories and the number of elements that belong to each of the categories.

2.1.3 Relative Frequency and Percentage Distributions:

The relative frequency of a category is obtained by dividing the frequency of that category by the sum of all frequencies. Thus, the relative frequency shows what fractional part or proportion of the total frequency belongs to the corresponding category. A relative frequency distribution lists the relative frequencies for all categories.

Definition:

Calculating Relative Frequency of a Category:

Relative frequency of a category = $\frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$

The percentage for a category is obtained by multiplying the relative frequency of that category by 100. A percentage distribution lists the percentages for all categories.

Calculating Percentage:

Percentage = (Relative frequency) × 100%

The relative frequencies and percentages from Table 2.3 are calculated and listed in Table 2.4.

Table 2.4
Relative frequency and percentage distributions of worries about not having enough money to pay normal monthly bills

Response	Relative Frequency	Percentage
Very worried	162/1015 = 0.16	0.16(100) = 16
Moderately worried	203/1015 = 0.2	0.2(100) = 20
Not too worried	305/1015 = 0.3	0.3(100) = 30
Not worried at all	325/1015 = 0.32	0.32(100) = 32
Others	20/1015 = 0.02	0.02(100) = 2
Total	1	100%

2.1.4 Graphical Presentation of Qualitative Data:

All of us have heard the adage “a picture is worth a thousand words.” A graphic display can reveal at a glance the main characteristics of a data set. The bar graph and the pie chart are two types of graphs that are commonly used to display qualitative data.

Bar Chart:

To construct a bar chart, we mark the various categories on the horizontal axis as in Figure 2.1. Note that all categories are represented by intervals of the same width. We mark the frequencies on the vertical axis. Then we draw one bar for each category such that the height of the bar represents the frequency of the corresponding category. We leave a gap between adjacent bars. Figure 2.1 gives the bar chart for the frequency distribution of Table 2.3.

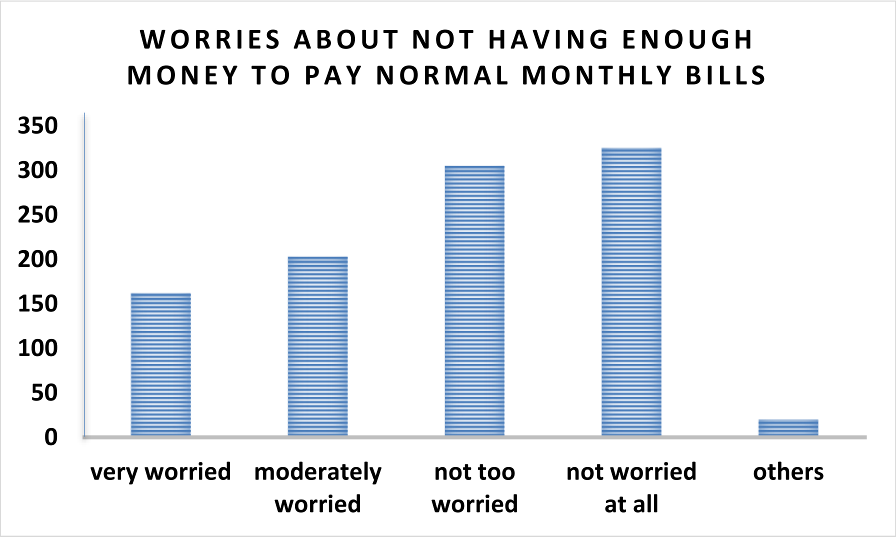


Figure 2.1
Bar Chart for the Frequency Distribution of Table 2.3.

Definition:

Bar Chart: A graph made of bars whose heights represent the frequencies of respective categories is called a bar chart.

The bar charts for relative frequency and percentage distributions can be drawn simply by marking the relative frequencies or percentages, instead of the frequencies, on the vertical axis. Sometimes a bar chart is constructed by marking the categories on the vertical axis and the frequencies on the horizontal axis.

Pie Chart:

A pie chart is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The whole pie (or circle) represents the total sample or population. Then we divide the pie into different portions that represent the different categories.

Definition:

Pie Chart is a circle divided into partitions that represent the relative frequencies or percentages of a population or a sample belonging to different categories is called a pie chart.

Figure 2.2 shows the pie chart for the percentage distribution of Table 2.3.

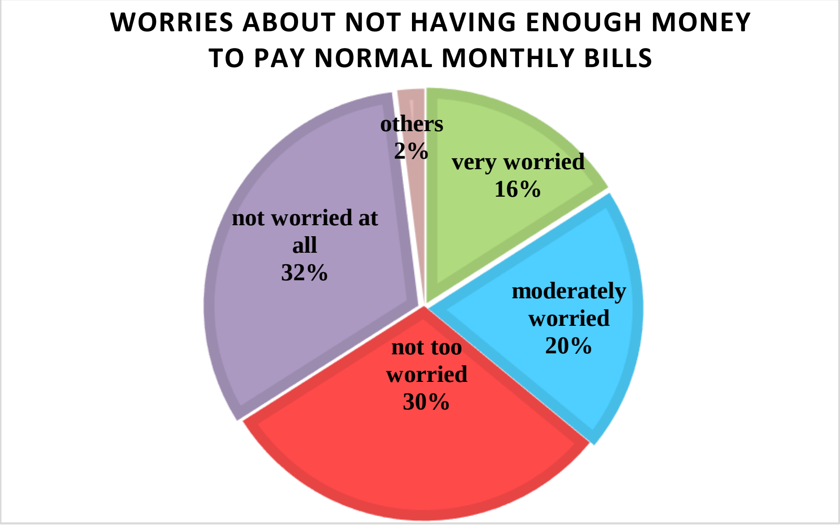


Figure 2.2
Pie Chart for the Percentage Distribution of Table 2.3

2.2 Organizing and Graphing Quantitative Data:

In the previous section we learned how to group and display qualitative data. This section explains how to group and display quantitative data.

2.2.1 Frequency Distributions for Quantitative Data:

The data presented in Table 2.5 represents the family size of 20 families as follows:

Table 2.5
Family size of 20 families

3	1	4	6	3	1	2	5	3	2
4	3	4	3	5	2	4	3	3	5

The frequency table of table 2.5 is given by Table 2.6 which is an illustration of a frequency distribution table for discrete quantitative data.

Table 2.6
Family Size of 20 Families

Family Size	Number of Families
1	2
2	3
3	7
4	4
5	3
6	1
Total	20

While table 2.7 is an illustration of a frequency distribution table for continuous quantitative data. Whereas the data that list individual values are called ungrouped data, the data presented in a frequency distribution table are called grouped data.

Table 2.7 gives the weekly earnings of 100 employees of a large company. The first column lists the classes, which represent the (quantitative) variable weekly earnings. For quantitative data, an interval that includes all the values that fall on or within two numbers (the lower and upper limits) is called a class. Note that

Table 2.7
Weekly Earnings of 100 Employees of a Company

Weekly Earnings (Classes)	Number of Employees (f)
800 –	4
1000 –	11
1200 –	39
1400 –	24
1600 –	16
1800 – 2000	6
Total	100

the classes always represent a variable. As we can observe, the classes are non-overlapping; that is, each value for earnings belongs to one and only one class and there are no gaps between any two successive intervals. The second column in the table lists the number of employees who have earnings within each class. For example, 4 employees of this company earn 800 to less than 1000 per week. The numbers listed in the second column are called the frequencies, which give the number of data values that belong to different classes, where the frequencies are denoted by f_i (the frequency of i^{th} class) and c = 1, 2, 3, …….., c, where c is the number of classes.

For quantitative data, the frequency of a class represents the number of values in the data set that fall in that class. Table 2.7 contains six classes. Each class has a lower limit and an upper limit. The values 800, 1000, 1200, 1400, 1600, and 1800 give the lower limits, and implicitly the upper limits are the lower limits of the next classes.

Frequency Distribution for Quantitative Data:

Definition:

Frequency Distribution for Quantitative Data:
A frequency distribution for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called grouped data.

The difference between the lower limits of two consecutive classes gives the class width. The class width is also called the class size.

Finding Class Width:

To find the width of a class, subtract its lower limit from the lower limit of the next class. Thus:

Width of a class = Lower limit of the next class
− Lower limit of the current class

Thus, in Table 2.5, Width of the first class = 1000 − 800 = 200

The class widths for the frequency distribution of Table 2.5 are listed in the second column of Table 2.6. Each class in Table 2.6 (and Table 2.5) has the same width of 200.

The class midpoint is obtained by dividing the sum of the two limits of a class by 2.

Definition:

Class Midpoint:

$$\textbf{Class Midpoint (or Mark)} = \frac{\textbf{Lower limit} + \textbf{Upper limit}}{\mathbf{2}}$$

Thus, the midpoint of the first class in Table 2.5 or Table 2.6 is calculated as follows:

$$\text{Midpoint of the first class} = \frac{\mathbf{800\ +1000}}{\mathbf{2}} = 900$$

The class midpoints for the frequency distribution of Table 2.7 are listed in the third column of Table 2.8.

Table 2.8
Class Widths and Class Midpoints for Table 2.5

Class Limits	Class Width	Class Midpoint
800 –	200	900
1000 –	200	1100
1200 –	200	1300
1400 –	200	1500
1600 –	200	1700
1800 – 2000	200	1900

2.2.2 Constructing Frequency Distribution Tables:

When constructing a frequency distribution table, we need to make the following three major steps.

1- Number of Classes:

Usually the number of classes for a frequency distribution table varies from 5 to 20, depending mainly on the number of observations in the data set. It is preferable to have more classes as the size of a data set increases. The decision about the number of classes is arbitrarily based on the data organizer. However, one rule helps us to decide the number of classes is Sturges' formula:

$$\mathbf{c = 1 + 3.3 \log_{10}(\boldsymbol{n})}$$

Where c is the number of classes and n is the number of observations in the data set. The value of log(n) can be obtained by using a calculator.

2- Class Width:

Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are the same size, first find the difference between the largest and the smallest values in the data. Then, the approximate width of a class is obtained by dividing this difference by the number of desired classes.

Calculation of Class Width:

$$\boxed{\textbf{Approximate class width} = \frac{\textbf{Largest Value} - \textbf{Smallest Value}}{\textbf{Number of Classes}}}$$

Usually this approximate class width is rounded to a convenient number, which is then used as the class width. Note that rounding this number may slightly change the number of classes initially intended.

3- Lower Limit of the First Class or the Starting Point:

Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

Example (1):

The following data pertain to weights (in kg) of 32 students of a class:

42 74 38 60 82 109 41 61 75 83 59
63 53 110 76 84 50 67 65 78 77 110
56 95 68 69 104 80 79 79 54 73

Prepare a suitable frequency table.

Solution:

1- $c = 1 + 3.3\log_{10}(32)$ = 5.97 ≅ 6

2- class width = $\frac{110-38}{6}$ = 12

3- Lower limit of the first class = 38

Table 2.9
Weights of Students of a Class

Class Interval	Tally Mark	Frequency
38 –	𝍷𝍷𝍷	3
50 –	𝍸𝍷𝍷	7
62 –	𝍸 𝍷	6
74 –	𝍸 𝍸 𝍷	11
86 –	𝍷	1
98 – 110	𝍷𝍷𝍷𝍷	4
Total	-	32

The frequency distribution consists of only two columns. The column of tallies is not a part of the frequency distribution. Therefore, the frequency distribution of these data is as follows:

Table 2.10
Frequency Distribution
of the Weights of 50 students

Class	Frequency
38 –	3
50 –	7
62 –	6
74 –	11
86 –	1
98 – 110	4
Total	32

Frequency Distribution Table for a Given Number of Classes or a Given Class Width:

Another alternative approach is to determine the number of classes without using Sturge’s formula or even a given class width (which implies a given number of classes), in such case we just skip step (1).

Example (2):

The data given below related to the number of years that 50 workers of a small factory has worked for.

Table 2.11
Years of experience of 50 workers

1.4	2.4	0.6	5.1	4.1	4.8	10.9	3.9	11.6	0.9
11.0	8.6	4.4	0.8	5.7	2.3	1.3	7.6	9.3	14.4
5.4	6.9	8.6	3.2	10.6	6.8	7.1	8.4	2.1	11.3
0.4	4.9	8.2	10.8	15.0	9.3	2.3	0.7	3.9	6.2
2.2	5.7	13.8	10.1	0.7	3.2	4.6	9.8	3.9	2.7

Construct a frequency distribution with width = 2 for each class.

Solution:

1- c = $\frac{\mathbf{15.0 - 0.4}}{\mathbf{2}}$ **=** 7.3 ≅ 8

2- Lower limit of the first class: it would be clearer if we start the first class with 0 instead 0.4 (Table 2.12 in the next page)).

2.2.3 Relative Frequency and Percentage Distributions:

Using Table 2.9, we can compute the relative frequency and percentage distributions in the same way as we did for qualitative data in section 2.1.3. The relative frequencies and percentages for a quantitative data set are obtained as given below. Note that relative frequency is the same as proportion.

Therefore, the frequency distribution of years of experience of 50 students is given in Table2.13.

Table 2.12
Years of Experience of 50 Workers

Class	Tally Mark	Frequency
0 -	𝍸 |||	8
2 -	𝍸 𝍸 |	11
4 -	𝍸 ||||	9
6 -	𝍸	5
8 -	𝍸 ||	7
10 -	𝍸 ||	7
12 -	|	1
14-16	||	2
Total	-	50

Table 2.13
Frequency Distribution
Of Years of Experience of 50 Workers

Class	Frequency
0-	8
2-	11
4-	9
6-	5
8-	7
10-	7
12-	1
14-16	2
Total	50

Relative frequency of a category = $\frac{\textbf{Frequency of That Category}}{\textbf{Sum of All Frequencies}} = \frac{\mathbf{f}}{\sum \mathbf{f}}$

Percentage = (Relative frequency) × 100%

Example (3):

Calculate the relative frequencies and percentages for Table 2.10.

Solution:

The relative frequencies and percentages for the data in Table 2.9 are calculated and listed in the second and third columns, respectively, of Table 2.14.

Table 2.14
Relative Frequency and Percentage Distributions of the weights of students of a class

Class	Relative Frequency	Percentage (%)
38 –	3/32 = 0.09375	9.375
50 –	7/32 = 0.21875	21.875
62 –	6/32 = 0.1875	18.75
74 –	11/32 = 0.34375	34.375
86 –	1/32 = 0.03125	3.125
98 – 110	4/32 = 0.125	12.5
Total	1	100 %

2.2.4 Cumulative Frequency Tables:

Sometimes, we may be interested in locating the relative position of a given score in a distribution. For example, we may be interested in finding out how many or what percentage of our sample was younger than 40 or older than 60. Frequency distributions can be presented in a cumulative fashion to answer such questions. There are two types of cumulative frequency tables:

1- Less-Than (Ascending) Cumulative Frequency Table:

It gives the total number of values that fall below the upper limit of each class. Cumulative frequency (CF) for a given class is obtained by adding the frequency of this class to the frequencies of all classes that come before it. Table 2.15 represents a less-than cumulative frequency distribution of table 2.13.

Table 2.15
Less-Than Frequency Distribution for Years of Experience of 50 Workers

Class	Less-Than Cumulative Frequency	More Effective Way
Less than 0	0	0
Less than 2	0 + 8 = 8	0 + 8 = 8
Less than 4	0 + 8 + 11 = 19	8 + 11 = 19
Less than 6	0 + 8 + 11 + 9 = 28	19 + 9 = 28
Less than 8	0 + 8 + 11 + 9 + 5 = 33	28 + 5 = 33
Less than 10	0 + 8 + 11 + 9 + 5 + 7 = 40	33 + 7 = 40
Less than 12	0 + 8 + 11 + 9 + 5 + 7 + 7 = 47	40 + 7 = 47
Less than 14	0 + 8 + 11 + 9 + 5 + 7 + 7 + 1 = 48	47 + 1 = 48
Less than 16	0 + 8 + 11 + 9 + 5 + 7 + 7 + 1 + 2 = 50	48 + 2 = 50

2- More-Than (Descending) Cumulative Frequency Table:

It gives the frequency at or above each class of the variable. obtained by adding the frequency in each class to the frequencies of all the classes above it (or by subtracting the frequency for the succeeding classes from the total frequency). Table 2.14 represents more - than cumulative frequency distribution of table 2.9.

Table 2.16
More-than frequency distribution
for weights of students of a class

Class Interval	More-Than Cumulative Frequency	More Effective Way
More Than 38	3 + 7 + 6 + 11 + 1 + 4 = 32	32
More Than 50	7 + 6 + 11 + 1 + 4 = 29	32 – 3 = 29
More Than 62	6 + 11 + 1 + 4 = 22	29 – 7 = 22
More Than 74	11 + 1 + 4 = 16	22 – 6 = 16
More Than 86	1 + 4 = 5	16 – 11 = 5
More Than 98	4	5 – 1 = 4
More Than 110	0	4 – 4 = 0

The following table summarizes the “less than” and “more than” tables for data given in Table (13)**:**

Cumulative Frequency (CF) Tables For Tables 2.13 and 2.10

For Table 2.13		**For Table 2.10**	
Class	**CF**	**Class**	**CF**
Less than 0	0		
Less than 2	8	More Than 38	32
Less than 4	19	More Than 50	29
Less than 6	28	More Than 62	22
Less than 8	33	More Than 74	16
Less than 10	40	More Than 86	5
Less than 12	47	More Than 98	4
Less than 14	48	More Than 110	0
Less than 16	50		

2.2.5 Graphical Presentation of Quantitative Data:

Quantitative data can be displayed in a histogram or a polygon or frequency curve. This section describes how to construct such graphs.

Bar Chart:

A Bar Chart was explained in graphical presentation of qualitative variable, but it also can be used to draw a frequency distribution for quantitative categorical variable. Figure 2.3 represents a Bar Chart for table 2.6

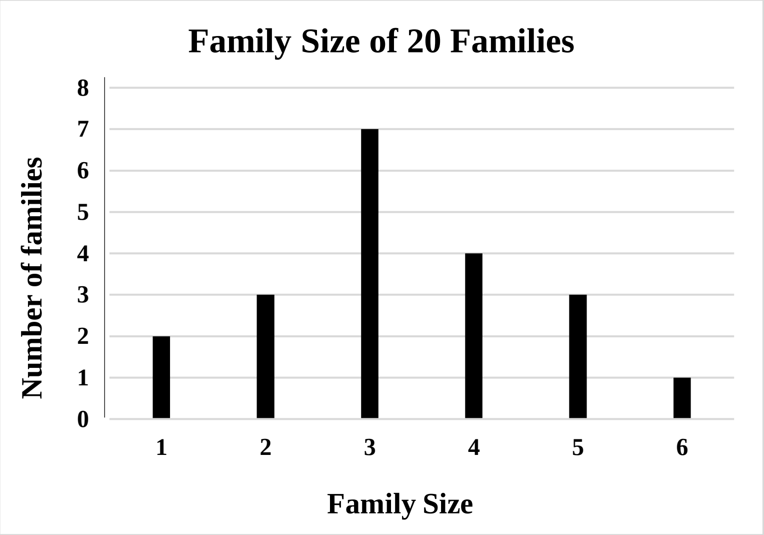


Figure 2.3
Histogram for the Frequency Distribution of Table 2.9.

Histogram:

A histogram can be drawn for a frequency distribution, a relative frequency distribution, or a percentage distribution. To draw a histogram, we first mark classes on the horizontal axis and frequencies (or relative frequencies or percentages) on the vertical axis. Next, we draw a bar for each class so that its height represents the frequency of that class. The bars in a histogram are drawn adjacent to each other with no gap between them.

A histogram is called a frequency histogram, a relative frequency histogram, or a percentage histogram depending on whether frequencies, relative frequencies, or percentages are marked on the vertical axis. Figures 2.4 and 2.5 show the frequency and the percentage histograms, respectively, for the data of Table 2.9.

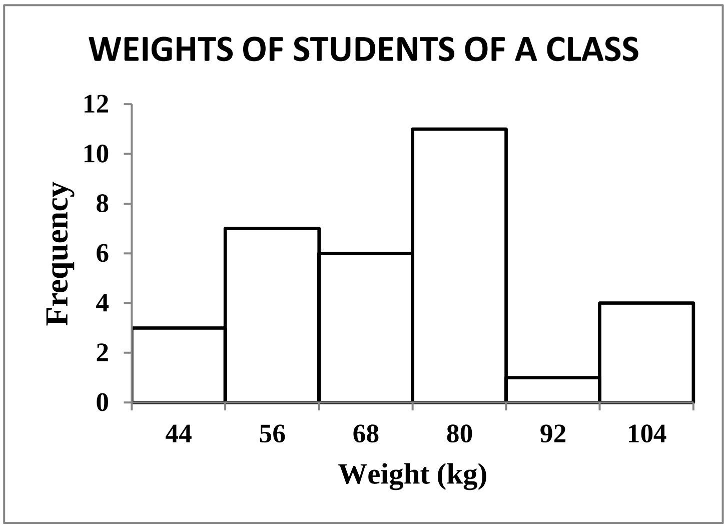


Figure 2.4
Histogram for the frequency distribution of Table 2.9

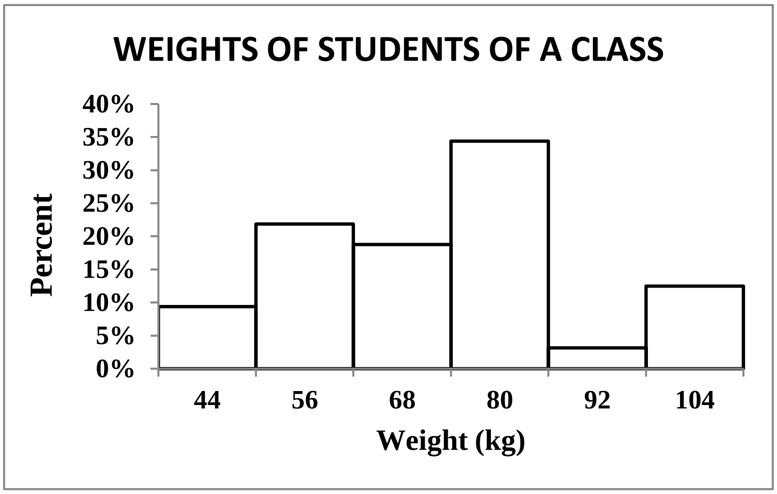


Figure 2.5
Histogram for the percentage distribution of Table 2.14

Frequency Polygon (Line graph):

A frequency polygon (Line Graph) is another device that can be used to represent quantitative data in graphic form. To draw a frequency polygon, we first mark a dot above the midpoint of each class at a height equal to the frequency of that class. This is the same as marking the midpoint at the top of each bar in a histogram. The resulting line graph is called a frequency polygon or simply a polygon.

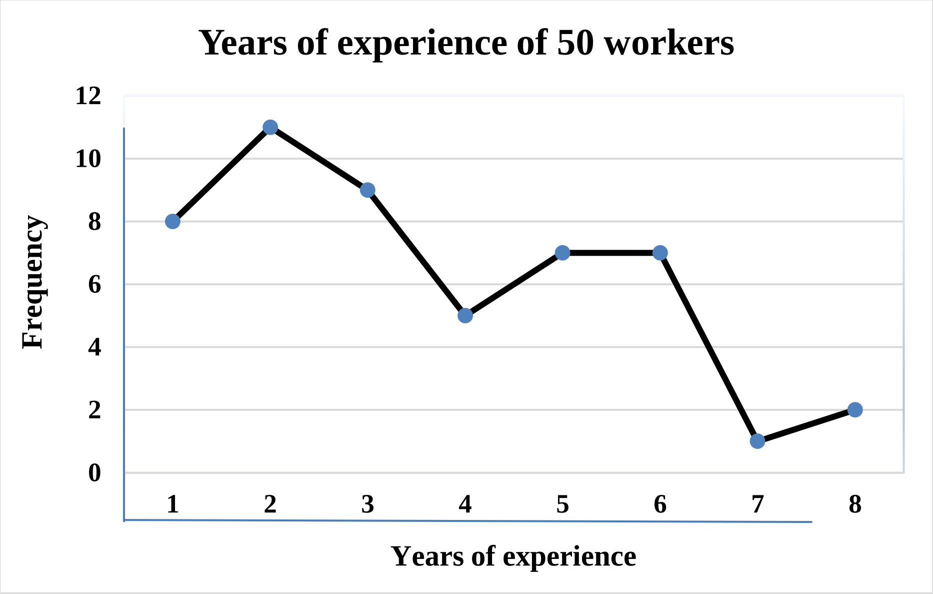


Figure 2.6
A Frequency Polygon (Line Graph)
for the Frequency Distribution of Table 2.13

Cumulative Frequency Curve (Ogive):

Cumulative frequency curves, also known as ogives, are graphs that can be used to determine how many data values lie above (or below) a particular value in a data set. Figures 2.7 and 2.8 show the less-than ogive and the more-than ogive, respectively, for the data of Tables 2.13.

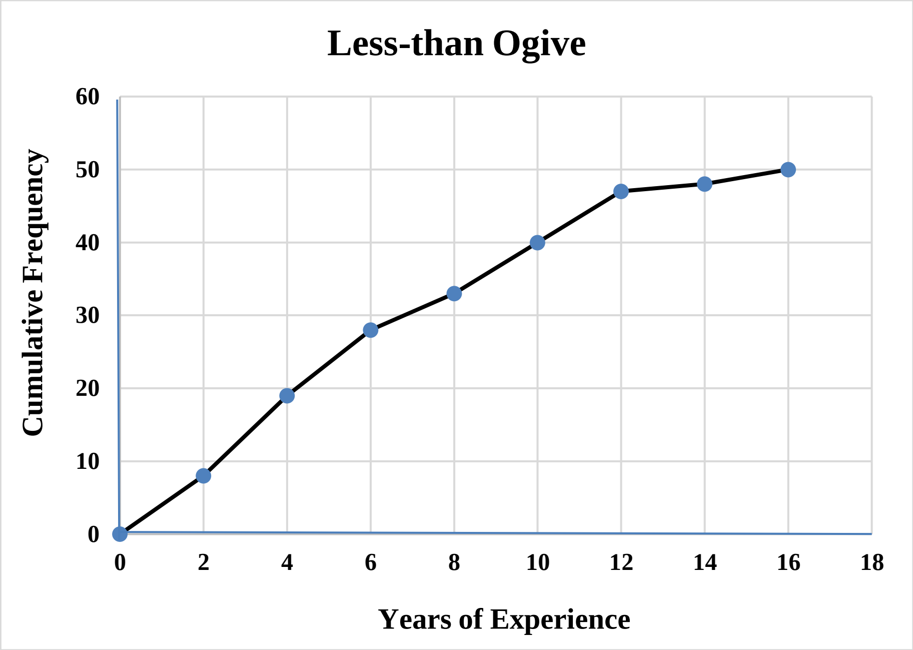


Figure 2.7 Less-than Ogive
for the Cumulative Frequency Distribution of Table 2.13

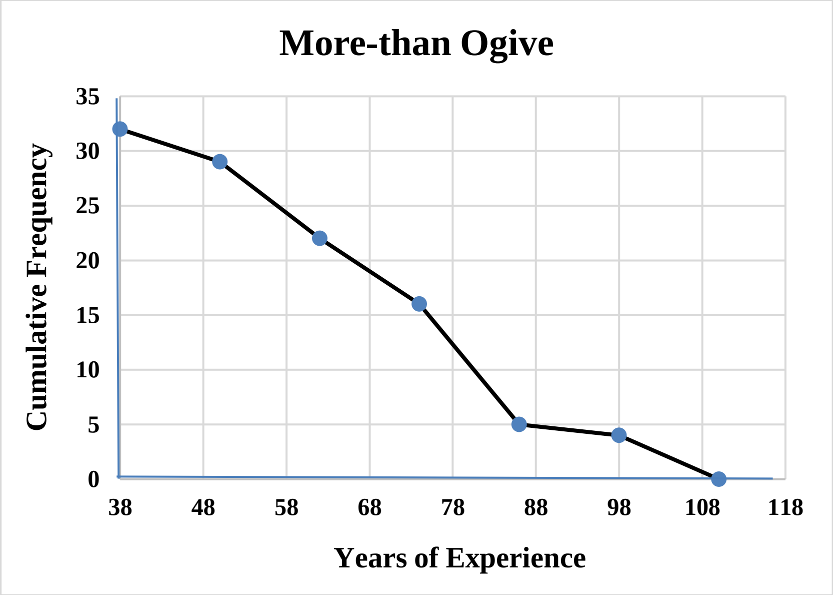


Figure 2.8 More-Than Ogive
for the Cumulative Frequency Distribution of Table 2.14

Exercises

1- The following data give the results of a sample survey. The letters Y, N, and D represent the three categories.

D	N	N	Y	Y	Y	N	Y	D	Y
Y	Y	Y	Y	N	Y	Y	N	N	Y
N	Y	Y	N	D	N	Y	Y	Y	Y
Y	Y	N	N	Y	Y	N	N	D	Y

a. Prepare a frequency distribution table.

b. Calculate the relative frequencies and percentages for all categories.

c. What percentage of the elements in this sample belongs to category Y?

d. What percentage of the elements in this sample belong to category N or D?

e. Draw a pie chart for the percentage distribution.

2- A market research company asked residents of Qena to name their favorite pizza topping. The possible responses included the following choices**:** meats (M); seafood, for example, tuna, or crab (S); vegetables and fruits (V); poultry (PO); and cheese (C). The following data represent the responses of a random sample of 36 people.

V	M	M	M	V	PO	S	M	V	S	V	S
M	S	V	V	V	M	S	S	V	M	C	V
V	V	C	V	S	PO	V	M	S	M	PO	M

a. Prepare a frequency distribution table.

b. Calculate the relative frequencies and percentages for all categories.

c. What percentage of the respondents mentioned vegetables and fruits, poultry, or cheese?

3- The following data show the method of payment by 16 customers in a supermarket checkout line. Here, C refers to cash, CK to check, CC to credit card, D to debit card, and O stands for other.

C	CK	CK	C	CC	D	O	C
CK	CC	D	CC	C	CK	CK	CC

a. Construct a frequency distribution table.

b. Calculate the relative frequencies and percentages for all categories.

c. Draw a pie chart for the percentage distribution.

4- In a 2018 survey of employees conducted by financial magazine, employees were asked about their overall financial stress levels. The following table shows the results of this survey.

Financial Stress Level	**Percentage of Responses (%)**
No financial stress	14
Some financial stress	63
High financial stress	18
Overwhelming financial stress	5

Draw a pie chart for this percentage distribution.

5- A local gas station collected data from the day's receipts, recording the gallons of gasoline each customer purchased. The following table lists the frequency distribution of the gallons of gas purchased by all customers on this one day at this gas station.

Gallons of Gas	**Number of Customers**
0 -	31
4 -	78
8 -	49
12 -	81
16 -	117
20 - 24	13

a. How many customers were served on this day at this gas station?

b. Find the class midpoints. Do all of the classes have the same width? If so, what is this width? If not, what are the different class widths?

c. Prepare the relative frequency and percentage distribution columns.

d. What percentage of the customers purchased 12 gallons or more?

e. Explain why you cannot determine exactly how many customers purchased 10 gallons or less.

f. Prepare the less-than and more-than cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the given table.

6- The following data give the one-way commuting times (in minutes) from home to work for a random sample of 50 workers.

23	17	34	26	18	33	46	42	12	37
44	15	22	19	28	32	18	39	40	48
16	11	9	24	18	26	31	7	30	15
18	22	29	32	30	21	19	14	26	37
25	36	23	39	42	46	29	17	24	31

a. Construct a frequency distribution table using the class width = 10.

b. Calculate the relative frequency and percentage for each class.

c. Construct a histogram for the percentage distribution made in Part b.

d. What percentage of the workers in this sample commutes for 30 minutes or more?

e. Prepare the less-than and more-than cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the table of Part a.

7- In a survey it was found that 64 families bought milk in the following quantities (liters) in a particular month:

19	22	09	22	12	39	19	14	23	06	24	16	18
07	17	20	25	28	18	10	24	20	21	10	07	18
28	24	20	14	24	25	34	22	05	33	23	26	29
13	36	11	26	11	37	30	13	08	15	22	21	32
21	31	17	16	23	12	09	15	27	17	21	16	

a. Using Struges' formula, convert the above data into a frequency distribution.

b. Prepare the less-than and more-than cumulative frequency distribution.

c. Prepare the less-than and more than ogive.

8- The marks obtained by 30 students in Statistics test are given below:

42	48	57	31	40	30	12	52	59	45
65	67	29	22	72	62	60	58	55	40
18	25	44	61	75	77	48	32	26	50

a. Construct a frequency distribution table using 7 classes.

b. Prepare the less-than and more than cumulative frequency distribution.

c. Prepare a histogram and frequency polygon for the frequency distribution in part a.

d. Prepare a less-than and more than ogive.

Chapter (3)
Measures of Central Tendency

The measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classified as summary statistics. The mean (often called the average) is the most likely of the measures of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

3.1 Mean:

This section discusses how to find the arithmetic mean (Mean or Average) of ungrouped and grouped data using 3 different methods which are:

(1) Direct method.
(2) Assumed mean method.
(3) Step-deviation method.

The arithmetic mean is normally abbreviated to just the 'mean' The main advantage of the mean is that it uses all the values in the data set, while the main disadvantage that it is severely affected by outliers.

3.1.1 The Mean of Ungrouped Data:

The arithmetic mean of a set of values is defined as 'the sum of all the values' divided by 'the number of values', that is

$$\text{Mean} = \frac{\text{Sum of All Values}}{\text{Number of Values}}$$

Example 1:

(a) If a firm received orders worth (by 1000 L.E)

30, 10, and 56

for three consecutive months, their mean value of orders per month would be calculated as:

$$\frac{30+10+56}{3} = \frac{96}{3} = 32$$

(b) The mean of the values 12 , 8 , 25 , 26 , 10 is calculated as:

$$\frac{12+8+25+26+10}{5} = \frac{81}{5} = 16.2$$

Note that the mean of a set of values is not necessarily the same as any of the original values in the set, as demonstrated in both (a) and (b) above.

When dealing with a set of items that have known values, the values themselves can be written down and manipulated. However, if it is necessary to consider the values of a set of items in general terms, particularly for use in formula, the notation used is x_1 , x_2 , …, x_n, or x_i where i = 1,2,…, n and n is the number of items in the set. This notation is just a compact way of saying: 'the 1st x-value', '2nd x-value', … and so on.

For example, if a salesman completes 4 , 5 ,12 , 8, and 2 sales in consecutive weeks, this would correspond to: $x_1 = 4$ (i.e. the first x - value is 4), $x_2 = 5$, $x_3 =12$, $x_4 = 8$ and $x_5 = 2$. Here, the variable

X is ‘number of sales’ and x_1 is the 1st sale, x_2 is the 2nd sale, … and so on. In this case, n = 5. Note that any letter instead of ‘X’ can be used.

Adding the values of sets of general items together

$$x_1 + x_2 + x_3 + \ldots + x_n \text{ is written as } \sum x$$

∑ is called **“summation notation”**. It is a Greek symbol for capital ‘S’ (for Sum) and ∑ x can be simply translated as ‘add up all the x-values under consideration’.
For the sales example, we have:

$$\sum x = 4 + 5 + 12 + 8 + 2 = 31$$

Using ∑ notation, the mean of ungrouped data x_1 , x_2 , …, x_n can be calculated using 3 different methods

(1) Direct Method:

Mean of Ungrouped Data Using Direct Method:

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

Example 2:

To calculate the mean for the set:

43 , 75 , 50 , 51 , 51 , 47 , 50 , 47 , 40 , 48

Here, n = 10 and ∑ x = 502.

Therefore: $\mathbf{\bar{x} = \frac{\sum x}{n} = \frac{502}{10} = 50.2}$

(2) Method of Assumed Mean:

The first step is to choose one among the x_i’ s as the assumed mean and denote it by ‘a’. We may take ‘a’ to be that x_i which lies in the center of x_1, x_2 , . . ., x_n.

The next step is to find the difference (d_i) between each of the x_i's and some constant a. that is,

$$\mathbf{d_i = x_i - a}$$

Now, let us find the relation between $\bar{\mathbf{d}}$ and $\bar{\mathbf{x}}$:

Since:

$d_1 + d_2 + \ldots + d_n = (x_1 - a) + (x_2 - a) + \ldots + (x_n - a)$
$\sum d_i = \sum x_i - na$

Therefore

$\sum x_i = na + \sum d_i$

Dividing both sides by n, we get $\bar{\mathbf{x}} = \mathbf{a} + \bar{\mathbf{d}}$

Mean of Ungrouped Data Using Assumed Mean Method:

$$\bar{x} = a + \bar{d}$$

Example 3:

If the hourly wage of 12 workers was as follows:

32 , 26 , 27 , 41 , 36 , 29 , 45 , 40 , 32 , 28 , 33 , 38

Find the mean hourly wage using the assumed mean method.

Solution:

Let, a = the center of the data set = (min + max)/2
= (45 + 26) / 2 = 35.5 ≅ 35

Then, subtracting a from each value in the data set, we get the following deviations (d_i):

Hence: d_i's = -3 , -9 , -8 , 6 , 1 , -6 , 10 , 5 , -3 , -7 , -2 , 3

Then, $\bar{d} = \frac{\sum d_i}{n} = \frac{-13}{12} = -1.08$

Therefore,

$\bar{x} = a + \bar{d} = 35 + (-1.08) = 33.92$

Note: It is not necessarily for the value of the constant a to be the average of the minimum and maximum values. You can choose any convenient constant.

(3) Method of Step - Deviation:

If all the values of d_i are divisible by a common factor b (i.e., without remainder), calculations can be more simplified as follows:

Let $$\mathbf{D_i} = \frac{\mathbf{d_i}}{\mathbf{b}} = \frac{\mathbf{x_i - a}}{\mathbf{b}}$$

where a is the assumed mean and b is the common factor for d_i.

Now, let us find the relation between $\mathbf{\overline{D}}$ and $\mathbf{\bar{x}}$:

Since $\mathbf{D_i} = \frac{\mathbf{x_i - a}}{\mathbf{b}}$, then

$$\sum \mathbf{D_i} = \frac{\sum (\mathbf{x_i - a})}{\mathbf{b}}$$

Therefore, $\sum(\mathbf{x_i} - \mathbf{a}) = \mathbf{b}\sum\mathbf{D_i}$

and $\sum\mathbf{x_i} = \mathbf{na} + \mathbf{b}\sum\mathbf{D_i}$

Dividing both sides by n, then we get $\mathbf{\bar{x}} = \mathbf{a} + \mathbf{b\overline{D}}$

Mean of Ungrouped Data Using Step-Deviation Method:

$$\bar{x} = a + b\overline{D}$$

Example 4:

If the height of 7 students was as follows:

175 , 160 , 145 , 170 , 195 , 180 , 155

Find the mean height using assumed mean method.

Solution:

Let, a = the center of the data set = (Min + Max)/2

= (145 + 195) / 2 = 170

Then, subtracting a from each value in the data set, then we get the following deviations (d_i):

5 , -10 , -25 , 0 , 25 , 10 , -15

It is obvious that the common factor of (d_i) is (b = 5)

and dividing (d_i) by (5), then we get (D_i) values

1 , -2 , -5 , 0 , 5 , 2 , -3

$$\mathbf{\bar{D}} = \frac{\sum \mathbf{D_i}}{\mathbf{n}} = \frac{\mathbf{-2}}{\mathbf{7}}$$

Therefore,

$$\mathbf{\bar{x} = a + b\bar{D} = 170 + 5.\left(\frac{-2}{7}\right)}$$

$$= \mathbf{170 + (-1.43) = 168.57}$$

3.1.2 Mean of Grouped Data:

Large sets of data will normally be arranged into a frequency distribution, and thus previous formulas for the mean are not quite appropriate, since no account is taken of frequencies whether the variable is quantitative discrete or quantitative continuous.

1- In Case of Quantitative Discrete Variables:

Consider the following discrete frequency distribution

x_i	10	12	13	14	16	19
f_i	2	8	17	5	1	1

The total of all the values = the sub total of the 10’s (= 10 × 2= 20)
+ the sub total of the 12’s (= 12 × 8 = 96)
+ the sub total of the 13’s (= 13 × 17 =221)
+ the sub total of the 14’s (= 14 × 5 = 70)
+ the sub total of the 16’s (= 16 × 1 = 16)
+ the sub total of the 19’s (= 19 × 1 = 19) = 442

Notice that in order to get the sub-totals 20, 96, 221, …, etc., x_i is being multiplied by f_i each time. In other words, the total is just $\sum \mathbf{x_i f_i}$. Also, since there are 2 “10’s”, 8 “12’s”, 17 “13’s”, …, etc., the number of values included in the distribution is 2 + 8 + 17 + 5 + 1 + 1 = 34, therefore, $\mathrm{n} = \sum \mathbf{f_i}$ = 31.

Using $\sum \mathbf{x_i f_i}$ as the sum of products of x and f, and $\sum \mathbf{f_i}$ as the number of values, the mean in this case is: $\frac{442}{31}$ = 14.26

The mean of grouped quantitative discrete data can be calculated using 3 different methods.

(a) Direct Method:

Mean of Grouped Quantitative Discrete Data Using Direct Method:

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

where $\mathbf{x_i}$ are quantitative discrete values and
$\mathbf{f_i}$ are the corresponding frequencies

Example 5:

For the following discrete frequency distribution

Number of Vehicles Serviceable (x_i)	0	1	2	3	4	5
Number of Days (f_i)	2	5	11	4	4	1

Find the mean daily number of vehicles serviceable using the direct method.

Solution:

The normal layout for calculations is:

x	f	xf
0	2	0
1	5	5
2	11	22
3	4	12
4	4	16
5	1	5
Total	27	60

Thus: $\bar{\mathbf{x}} = \frac{\sum \mathbf{x_i f_i}}{\sum \mathbf{f_i}} = \frac{\mathbf{60}}{\mathbf{27}} = \mathbf{2.22}$

(b) Method of Assumed Mean:

Sometimes when the numerical values of x_i and f_i are large, finding the product of x_i and f_i becomes tedious and time consuming. So, for such situations, we can use the assumed mean method to reduce these calculations.

We can do nothing with f_i's, but we can change each x_i to a smaller number, so that our calculations become easy. How do we do this? What about subtracting a fixed number from each of these x_i's?

The first step is to choose one among the x_i' s as the assumed mean and denote it by 'a'. We may take 'a' to be that x_i which lies in the center of x_1, x_2 , . . ., x_n.

The next step is to find the difference (d_i) between each of x_i values and the assumed mean (a).

Finally, calculating the products of d_i and f_i , the mean is given by

$$\bar{\mathbf{x}} = \mathbf{a} + \bar{\mathbf{d}} = \mathbf{a} + \frac{\sum \mathbf{d_i f_i}}{\sum \mathbf{f_i}}$$

Mean of Grouped Quantitative Discrete Data Using the Assumed Mean Method:

$$\bar{x} = a + \frac{\sum d_i f_i}{\sum f_i}$$

Example 6:

Consider the following frequency distribution of family size:

Family size (x_i)	1	2	3	4	5	6	Total
Number of families (f_i)	2	10	12	13	9	4	50

Find the mean family size using the assumed mean method.

Solution:

- a = Center of x_i's = (min + max)/2 = (1 + 6) / 2 = 3.5
  For purposes of simplifying calculations, let a = 3
- Subtracting (a = 3) from each (x_i) we get the column:
  ($d_i = x_i – 3$)
- Multiplying each (d_i) by its corresponding (f_i) to get the column (d_i f_i).

x	**f**	**d = x - 3**	**df**
1	2	-2	-4
2	10	-1	-10
3	12	0	0
4	13	1	13
5	9	2	18
6	4	3	12
Total	27	60	29

Therefore, $\bar{\mathbf{x}} = \mathbf{a} + \frac{\sum \mathbf{d_i f_i}}{\sum \mathbf{f_i}} = \mathbf{3} + \frac{\mathbf{29}}{\mathbf{50}} = \mathbf{3.58}$ pearson.

(c) Method of Step – Deviation:

If all the values of d_i are divisible by a common factor b (i.e., without remainder), calculations can be more simplified as follows:

Let $\mathbf{D_i = \frac{d_i}{b} = \frac{x_i - a}{b}}$**,** where a is the assumed mean and b is the common factor of d_i.

Multiplying (D_i) by (f_i), the mean can be written as

$$\bar{\mathbf{x}} = \mathbf{a} + \mathbf{b}\bar{\mathbf{D}} = \mathbf{a} + \mathbf{b}\left(\frac{\sum \mathbf{D_i f_i}}{\sum \mathbf{f_i}}\right)$$

Mean of Grouped Quantitative Discrete Data Using the Step-Deviation Method:

$$\bar{x} = a + b\left(\frac{\sum D_i f_i}{\sum f_i}\right)$$

Example 7:

Consider the following apartment rent frequency distribution

x_i	1100	1150	1200	1250	1300	1350	1400	1450	1500	Total
f_i	6	10	5	12	22	16	13	11	5	100

where x_i is the apartment rent and f_i is the corresponding number of apartments.

Find the mean rent using step-deviation method.

Solution:

- a = Center of x_i's = (min + max)/2 = (1100 + 1500) / 2 = 1300
- Subtracting (a = 1300) from each (x_i), then we get the column $(d_i = x_i - 1300)$.
- Since all d_i's are divisible by (b = 50), then we can define the column $(D_i = d_i /50)$.
- Multiplying each (D_i) by its corresponding (f_i) to get the column $(D_i f_i)$.

x	f	d = x - 1300	D = d / 50	Df
1100	6	-200	-4	-24
1150	10	-150	-3	-30
1200	5	-100	-2	-10
1250	12	-50	-1	-12
1300	22	0	0	0
1350	16	50	1	16
1400	13	100	2	26
1450	11	150	3	33
1500	5	200	4	20
Total	27	60	29	19

Then, $\bar{\mathbf{x}} = \mathbf{a} + \mathbf{b}\left(\frac{\sum \mathbf{D_i f_i}}{\sum \mathbf{f_i}}\right) = \mathbf{1300} + \mathbf{50}\left(\frac{\mathbf{19}}{\mathbf{100}}\right) = \mathbf{1309.5}$

2- In Case of Quantitative Continuous Variables:

One of the disadvantages of grouping continuous data into the form of a frequency distribution is the fact that individual values of items are lost. This is particularly inconvenient when a mean need to be calculated as it is clearly impossible to calculate the exact value of the mean. However, it is possible to estimate it. This is done by using the class midpoints as representatives for x-values assuming that the observations are evenly distributed in each class. i.e.

X = The Class Midpoint

Also, the mean of grouped quantitative continuous data can be calculated using 3 different methods.

(a) Direct Method:

Mean of Grouped Quantitative Continuous Data Using the Direct Method:

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

where $\mathbf{x_i}$ are class midpoint and,
$\mathbf{f_i}$ are the corresponding frequencies.

Example 8:

For the following hourly wage frequency distribution:

Wage	30-	40-	50-	60-	70-	80-	90-100	Total
Number of Employees	1	4	7	14	11	8	5	50

Find the mean hourly wage using the direct method.

Solution:

- Finding class midpoints x_i = (lower limit + upper limit)/2
- Multiplying each (x_i) by its corresponding (f_i) to get the column ($x_i f_i$).

We can summarize these calculations as shown in the following table:

Wage	**f**	**Class Midpoint (x)**	**xf**
30-	1	35	35
40-	4	45	180
50-	7	55	385
60-	14	65	910
70-	11	75	825
80-	8	85	680
90-100	5	95	475
Total	50	-	3490

$$\bar{\mathbf{x}} = \frac{\sum \mathbf{x_i f_i}}{\sum \mathbf{f_i}} = \frac{3490}{50} = 69.8$$

(c) Method of Assumed Mean:

Choosing the value (a) as an assumed mean and subtracting this value from each class midpoint (i.e. $d_i = x_i - a$), **then:**

Mean of Grouped Quantitative Continuous Data Using the Assumed Mean Method:

$$\bar{x} = a + \frac{\sum d_i f_i}{\sum f_i}$$

Example 9:

For the hourly wage frequency distribution in Example (8), find the mean using the assumed mean method

Solution:

- Let assumed mean = a = midpoints of the class (60 – 70), that is a = (60 + 70)/2 = 65
- Subtracting (a = 65) from each (x_i), then we get the column ($d_i = x_i - 65$)
- Multiplying each (d_i) by its corresponding (f_i) to get the column ($d_i f_i$)

We can summarize these calculations as shown in the following table:

wage	f	Class Midpoint (x)	D = x - 65	df
30-	1	35	-30	-30
40-	4	45	-20	-80
50-	7	55	-10	-70
60-	14	65	0	0
70-	11	75	10	110
80-	8	85	20	160
90-100	5	95	30	150
Total	50	-	-	240

Then: $\bar{x} = a + \frac{\sum d_i f_i}{\sum f_i} = 65 + \frac{240}{50} = 69.8$

Note: You can consider any other value for the assumed mean. However, choosing one of the midpoints may simplify computations.

(c) Method of Step -Deviation:

If all the values of (d_i) can be divided by a common factor (b), then, by defining ($D_i = d_i$ / b), **Then:**

Mean of Grouped Quantitative Continuous Data Using the Step - Deviation Method:

$$\bar{x} = a + b\left(\frac{\sum D_i f_i}{\sum f_i}\right)$$

Example 10:

For the frequency distribution of hourly wage in Example (8) and Example (9), find the mean using the method of step-deviation.

Solution:

- Let a = 65 , and b = 10
- Then, we find the columns of ($D_i = d_i$ /10) and $D_i f_{i.}$

The following table presents calculations required:

Wage Class	f	class midpoint (x)	d = x - 65	D = d /10	Df
30-	1	35	-30	-3	-3
40-	4	45	-20	-2	-8
50-	7	55	-10	-1	-7
60-	14	65	0	0	0
70-	11	75	10	1	11
80-	8	85	20	2	16
90-100	5	95	30	3	15
Total	50	-	-		24

Then: $\bar{x} = a + b\left(\frac{\sum D_i f_i}{\sum f_i}\right) = 65 + 10\left(\frac{24}{50}\right) = 69.8$

The same result obtained for the other two methods (Examples: 8 and 9).

Note: The three methods used for obtaining the value of the mean must lead to the same results.

3.2 Median:

The median is generally considered as an alternative central tendency measure to the mean. This section defines the median and shows how to find its value for ungrouped and grouped data.

Definition:

Median:
The median is the middle value of a set of data after arranging data in an ascending order (or descending order).

3.2.1 The Median of Ungrouped Data:

In order to find the median of ungrouped data, the values have to be sorted in an ascending or descending order, then we choose the item which makes half of the data before this item and the other half after it. We may find one of the following two situations:

(a) Odd Number of Items:

For a set with an odd number (n) of items, the median can be precisely identified as the value of the $\left(\frac{\mathbf{n+1}}{\mathbf{2}}\right)^{\mathbf{th}}$ item. Thus, in a size-ordered set of 15 items, the median would be the $\left(\frac{\mathbf{15+1}}{\mathbf{2}}\right)^{\mathbf{th}}$ which is the 8th value.

Example 11:

Given the hourly wage of 7 workers as follows

75 , 47 , 48 , 50 , 51 , 51 , 43

Find the median hourly wage.

Solution:

- Arranging data in an ascending order, we get:

43 , 47 , 48 , 50 , 51 , 51 , 75

- Since n = 7 is an odd number of items, then the median order is $\left(\frac{\mathbf{7+1}}{\mathbf{2}}\right)^{\text{th}}$ = the 4th item.
- Therefore, the median = 50

(b) Even Number of Values:

When an ordered set of data contains an even number of values, there are two middle values. The order of the first value is $\left(\frac{\mathbf{n}}{\mathbf{2}}\right)$ and the order of the second value is $\left(\frac{\mathbf{n}}{\mathbf{2}}+\mathbf{1}\right)$. The convenient in this situation is to use the mean of these middle two values to give the median.

Example 12:

Given the weight of 10 students as follows

88 , 82 , 91 , 72 , 65 , 85 , 80 , 84 , 73 , 90

Find the median weight

Solution:

- By arranging values in a descending order, we get:

91 , 90 , 88 , 85 , 84 , 82 , 80 , 73 , 72 , 65

- Since n = 10 is an even number of items, order of the first middle item $=\left(\frac{\mathbf{n}}{\mathbf{2}}\right)=\left(\frac{\mathbf{10}}{\mathbf{2}}\right)=\mathbf{5}$ and the order of the second middle item = $\left(\frac{n}{2}+1\right)=\left(\frac{10}{2}+1\right)=6$.
- So, the median = $\frac{84+82}{2}$ = 83

3.2.2 The Median of Grouped Data:

As we mentioned in the previous section, the penalty paid by grouping values is the loss of their individual identities and thus there is no way that a median can be calculated exactly in this situation. However, we can use interpolation for estimating the median. Interpolation in this context is a simple mathematical technique which estimates an unknown value by utilizing immediately surrounding known values.

To be able to find the median, we need to find its order (median point):

The order of the median (Median Point) = $\left(\frac{\sum \mathbf{f_i}}{\mathbf{2}}\right)$

Then, we can use either the less-than cumulative frequency distribution or the more-than cumulative frequency distribution to find the value of the median as follows**:**

(a) Using Less-Than Cumulative Frequency Distribution:

The procedure for estimating the median using a less-than cumulative frequency distribution can be summarized as follows:

- Find the order of the median (median point) which is $\left(\frac{\sum \mathbf{f_i}}{\mathbf{2}}\right)$.
- Prepare a less-than cumulative frequency distribution.
- Identify the location of the median point in the column of F_i , and the location of the unknown value of the median (m) in the column of variable classes.
- Interpolate the value of the median.

Example 13:

For the hourly wage frequency distribution in Example (8), find the median for the hourly wages.

Solution:

- The order of the median (median point) = $\left(\frac{\sum \mathbf{f_i}}{\mathbf{2}}\right) = \frac{\mathbf{50}}{\mathbf{2}} = 25$
- Prepare a less-than cumulative frequency distribution.

Less than wage classes	**CF**
Less than 30	0
Less than 40	1
Less than 50	5
Less than 60	12
Less than 70	26
Less than 80	37
Less than 90	45
Less than 100	50

***m*: value of the median**

Order of The Median

This table can be read as follows:

- No workers with wages of less than L.E. 30
- 1 worker is paid less than L.E. 40
- 5 workers are paid less than L.E. 50
- 12 workers are paid less than L.E. 60
- Half of the workers (25 workers) are paid less than L.E. ‘m’ Which is equivalent to the definition of the median. Therefore, the class 60 - 70 is called **“the median class”**.

- Identify the location of the order of the median (median point) in the cumulative frequency column (F_i) and the location of the unknown value of the median (m) in wage column.

Wage	**F_i**
60	12
m	25
70	26

- Find the value of the median (m) using interpolation as follows:

The main idea of interpolation is that the ratio (25 – 12) to (26 – 12) is proportional to the ratio (m - 60) to (70 – 60), that is

$$\frac{\mathbf{m-60}}{\mathbf{70-60}}=\frac{\mathbf{25-12}}{\mathbf{26-12}} \text{ gives } \frac{\mathbf{m-60}}{\mathbf{10}}=\frac{\mathbf{13}}{\mathbf{14}}$$

$$\mathbf{m = 60 + 10\left(\frac{13}{14}\right) = L.E. 69.29}$$

Note that: the value of the median (m = L.E. 69.29) must lie between the lower and upper limit for the median class), i.e., between 60 and 70.

(b) Using More-Than Cumulative Frequency Distribution:

The same procedure applies here, the more-than frequency distribution is as follows

More than wage classes	Descending F_i
more than 30	50
more than 40	49
more than 50	45
more than 60	38
more than 70	24
more than 80	13
more than 90	5
more than 100	0

***m*: value of the median**

Order of the median = 25

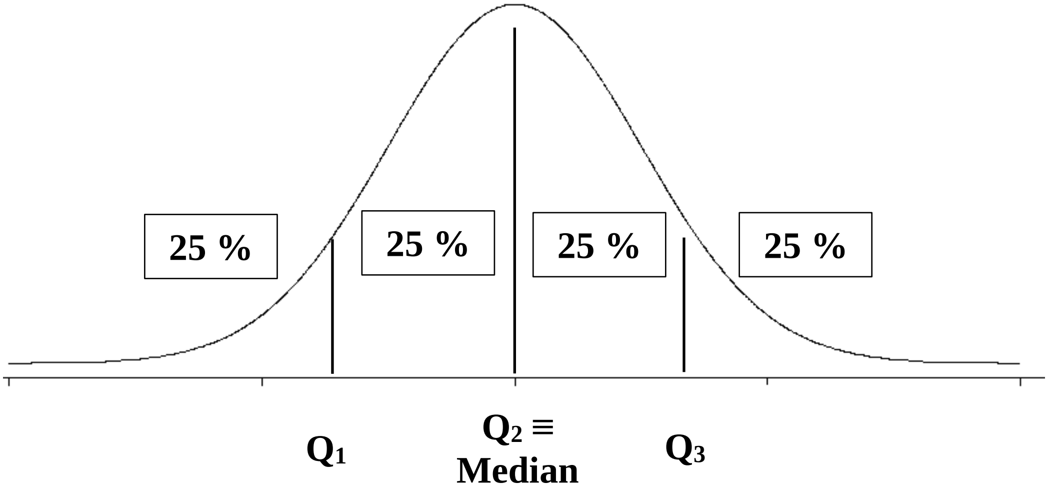
Wage	F_i
60	38
m	25
70	24

$$\frac{m-60}{70-60}=\frac{25-38}{24-38} \text{ gives } \frac{m-60}{10}=\frac{-13}{-14}$$

$$\mathbf{m} = \mathbf{60} + \mathbf{10}\left(\frac{\mathbf{13}}{\mathbf{14}}\right) = \mathbf{L.E. 69.29}$$

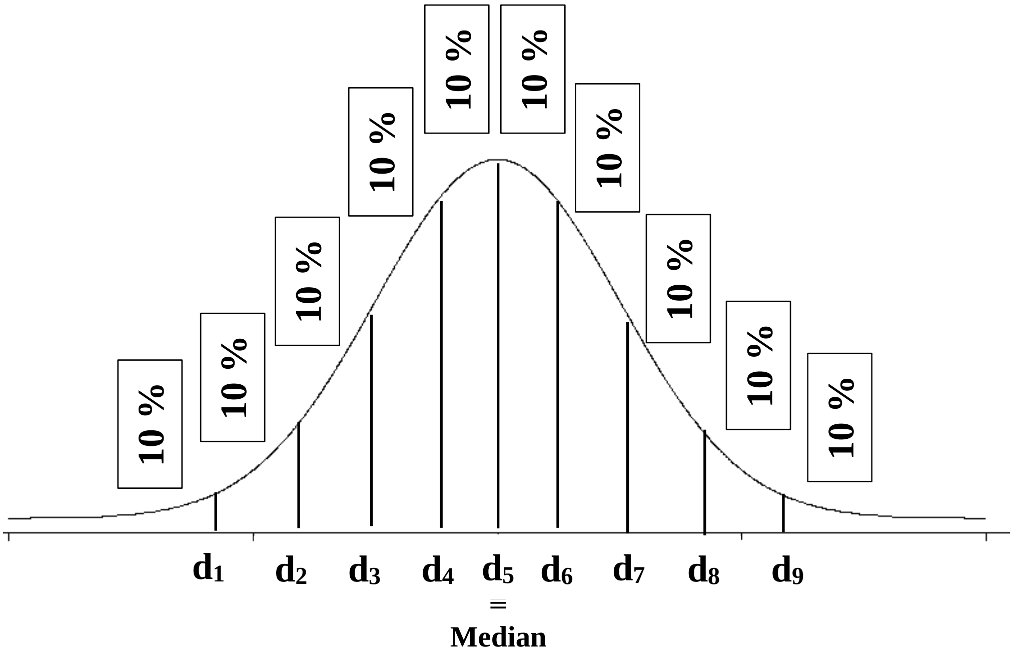
3.3 Quartiles, Deciles and Percentiles:

Quartiles are the summary measures that divide a ranked data set into four equal parts. Three measures will divide any data set into four equal parts. These three measures are the first quartile (denoted by Q_1), the second quartile (denoted by Q_2), the second quartile is the same as the median of a data set and the third quartile (denoted by Q_3).



Each of these partitions contains 25% of the observations of a data set arranged in increasing order. Approximately 25% of the values in a ranked data set are less than Q_1 and about 75% are greater than Q_1. The second quartile, Q_2, divides a ranked data set into two equal parts; hence, the second quartile and the median are the same. Approximately 75% of the data values are less than Q_3 and about 25% are greater than Q_3.

In the same way, deciles are the summary measures that divide a ranked data set into ten equal parts. Nine measures will divide any data set into ten equal parts. These nine measures are the first decile (denoted by d_1), the second decile (denoted by d_2), the third decile (denoted by d_3), and so on. Notice that the fifth decile (denoted by d_5) is the same as the median of a data set.



Each of these partitions contains 10% of the observations of a data set arranged in increasing order. Approximately 10% of the values in a ranked data set are less than (d_1) and about 90% are greater than (d_1). Approximately 20% of the data values are less than (d_2) and about 80% are greater than (d_2), and so on.

Similarly, percentiles are the summary measures that divide a ranked data set into one hundred equal parts. Ninety-nine measures will divide any data set into one hundred equal parts. These 99 measures are the first percentile (denoted by P_1), the second percentile (denoted by P_2), the third percentile (denoted by P_3), so on. Notice that the 25th percentile (P_{25}) is the same as the first quartile (Q_1), the 50th percentile (P_{50}) is the same as the median (m or Q_2) and the 75th (P_{75}) is the same as the third quartile (Q_3).

So, approximately 1% of the values in a ranked data set are less than (P_1) and about 99% are greater than (P_1), …, approximately 19% of the data values are less than (P_{19}) and about 81% are greater than (P_{19}), and so on.

We can use either the less-than cumulative frequency distribution or the more-than cumulative frequency distribution to find the value of either quartiles or deciles or percentiles.

The following steps summarizes how to calculate either quartiles or deciles or percentiles:

(i) Using Less-Than Cumulative Frequency Distribution:

- Form the less - than cumulative frequency distribution.
- Find the order of the measure ($\mathbf{R} \sum \mathbf{f_i}$) where R is the proportion of data which are less than the value of the measure.

For example: since half of the data is less than the median m, then $\mathbf{R} = \frac{\mathbf{1}}{\mathbf{2}}$

Therefore, the order of the median = $\frac{\mathbf{1}}{\mathbf{2}} \sum \mathbf{f_i}$ = $\frac{\sum \mathbf{f_i}}{\mathbf{2}}$

- Identify the location of the order of the median (median point) in the cumulative frequency column (F_i) and the location of the unknown value of the median (m) in the column of classes.
- Determine the value of the measure by interpolation.

Example 14:

The following frequency distribution illustrates the weekly income (L.E) for 300 families:

Income Classe	**750-**	**1000-**	**1250-**	**1500-**	**1750-**	**2000-**	**2250 - 2500**	Total
Number of Families	14	30	45	68	62	49	32	300

Using a less -than cumulative frequency distribution, find**:**

(a) The median. (b) The 1st and the 3rd quartiles.

(c) If 60% of families get less than L.E. x, determine the value of x.

{d) The percentage of families whose income less than L.E.1650

(e) The number of families whose incomes are between 1350 and L.E. 1800.

(f) The number of families whose incomes are at least L.E. 1850.

Solution:

Prepare the less-than cumulative frequency distribution.

Class	**CF**
less than 750	0
less than 1000	14
less than 1250	44
less than 1500	89
less than 1750	157
less than 2000	219
less than 2250	268
less than 2500	300

(a) To find the value of the median:

Since $\frac{1}{2}$ of the data is less than the median, then the order of the median = $\frac{1}{2}\sum \mathbf{f_i}$ = $\frac{1}{2}(\mathbf{300}) = \mathbf{150}$. That is, the median point = 150.

Income	**F_i**
L.T. 1500	89
L.T. m	150
L.T. 1750	157

$$\frac{\mathrm{m}-1500}{1750-1500}=\frac{150-89}{157-89} \quad \text{gives} \quad \frac{\mathrm{m}-1500}{250}=\frac{61}{68}$$

$$\mathrm{m}=1500+250\left(\frac{61}{68}\right)=\mathrm{L.E.}\,1724.26$$

Therefore, the median for the weekly income is L.E. 1724.26.

(b) To find the 1st and the 3rd quartiles:

Since $\frac{1}{4}$ of the data is less than the 1st quartile, then the order of the 1st quartile = $\frac{1}{4}\ .\sum \mathrm{f_i}$ = $\frac{1}{4}(300)=75$, that is, Q_1 point is 75.

Income	F_i
L.T. 1250	44
L.T. Q_1	75
L.T. 1500	89

$$\frac{Q_1-1250}{1500-1250}=\frac{75-44}{89-44} \quad \text{gives} \quad \frac{Q_1-1250}{250}=\frac{31}{45}$$

$$Q_1=1250+250\left(\frac{31}{45}\right)=\mathrm{L.E.}\,1422.22$$

So, the first quartile for the weekly income is L.E. 1422.22

- Similarly, since $\frac{3}{4}$ of the data is less than the 3rd quartile, then the order of the 3rd quartile = $\frac{3}{4}\sum \mathrm{f_i}$ = $\frac{3}{4}(300)=225$

Income	F_A
L.T. 2000	219
L.T. Q_3	225
L.T. 2250	268

$$\frac{Q_3 - 2000}{2250 - 2000} = \frac{225 - 219}{268 - 219}$$

$$\frac{Q_3 - 2000}{250} = \frac{6}{49} \text{ gives } Q_3 = 2000 + 250\left(\frac{6}{49}\right)$$

$$Q_3 = \text{L.E.}\,2030.61$$

Therefore, the third quartile for the weekly income is L.E. 2030.61

(c) Since 60% of families get less than x, then find the value of x. Note that x has the same definition of the 6th decile or the 60th percentile,

Since $\mathbf{\frac{6}{10}}$ (or $\mathbf{\frac{60}{100}}$) of the data is less than the 6th decile (or the 60th percentile), then the order of the 6th decile $= \frac{6}{10}\sum f_i$ $= \mathbf{\frac{6}{10}(300) = 180}$

Income	F_A
L.T. 1750	157
L.T. d_6	180
L.T. 2000	219

$$\frac{d_6 - 1750}{2000 - 1750} = \frac{180 - 157}{219 - 157} \text{ gives } \frac{d_6 - 1750}{250} = \frac{23}{62}$$

$$d_6 = 1750 + 250\left(\frac{23}{62}\right) = \text{L.E.}\,1842.74$$

(d) To find the percentage of families by which their income is less than L.E. 1650, then:

Let k be the number of families whose incomes are less than L.E. 1650. Using the cumulative distribution, we get

Income	F_A
L.T. 1500	89
L.T. 1650	k
L.T. 1750	157

$$\frac{1650-1500}{1750-1500}=\frac{\mathrm{k}-89}{157-89} \text{ gives } \frac{150}{250}=\frac{\mathrm{k}-89}{68}$$

$$\mathrm{k}=89+68\left(\frac{150}{250}\right)=129.8 \cong 130$$

The percentage of families whose weekly income are less than L.E. 1650 is: $\frac{130}{300}\times 100\%$ = 43.3%

(e) To find the number of families by which their weekly income is between 1350 and L.E. 1800.

- Let the number of families whose weekly incomes are less than L.E. 1350 = a
- Let the number of families whose weekly incomes are less than L.E. 1800 = b

Then, the number of families with weekly incomes between 1350 and L.E. 1800 = (b – a) families

Now, we find the value of (a) and (b) as follows

Income	F_A
L.T. 1250	44
L.T. 1350	a
L.T. 1500	89

$$\frac{1350-1250}{1500-1250}=\frac{\mathrm{a}-44}{89-44} \text{ gives } \frac{100}{250}=\frac{\mathrm{a}-44}{45}$$

$\mathrm{a}=44+45\left(\frac{100}{250}\right)=62$, and

Income	F_A
L.T. 1750	157
L.T. 1800	b
L.T. 2000	219

$$\frac{1800-1750}{2000-1750}=\frac{b-157}{219-157} \text{ gives } \frac{50}{250}=\frac{b-157}{62}$$

$$b=157+62\left(\frac{50}{250}\right)=169.4\cong 169$$

Hence, the number of families whose weekly incomes are between L.E. 1350 and L.E. 1800 is equal to:

$$b – a = 169 – 62 = 107$$

(f) The number of families who have weekly income of at least L.E.1850 = 300 – the number of families whose income are less than L.E. 1850.

Let the number of families with weekly incomes of less than L.E. 1850 = k

Then, we have the following calculations

Income	F_A
L.T. 1750	157
L.T. 1850	k
L.T. 2000	219

$$\frac{1850-1750}{2000-1750}=\frac{k-157}{219-157} \text{ gives } \frac{100}{250}=\frac{k-157}{62}$$

$$k=157+62\left(\frac{100}{250}\right)=181.8\cong 182$$

So the number of families whose weekly incomes are more than L.E 1850 is:

$$300 - k = 300 - 182 = 118 \text{ families}$$

(ii) Using More-Than Cumulative Frequency Distribution (Descending Cumulative Frequency Distribution):

- Form a more - than cumulative frequency distribution.
- Find the order of the measure ($\mathbf{R}\sum \mathbf{f_i}$)
  where R is the proportion of data which are more than the value of the measure.
- Identify the location of the order of the measure in the column of cumulative frequencies (F_i), and the location of the unknown value of the measure (say, m) in the column of classes.
- Determine the value of the measure using interpolation.

Example 15:

Solve the previous example using a more-than cumulative frequency distribution

Solution:

Prepare the more-than cumulative frequency distribution

More Than Income Classes	F_D
more than 750	300
more than 1000	286
more than 1250	256
more than 1500	211
more than 1750	143
more than 2000	81
more than 2250	32
more than 2500	0

(a) In order to find the median, we have the following:

Since $\frac{\mathbf{1}}{\mathbf{2}}$ of the data is more than the median, then the order of the median = $\frac{\mathbf{1}}{\mathbf{2}}\sum \mathbf{f_i}$ = $\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{300}) = \mathbf{150}$, i.e, Median Point = 150

Income	F_i
M.T. 1500	211
M.T. m	150
M.T. 1750	143

$$\frac{\mathrm{m} - 1500}{1750 - 1500} = \frac{150 - 211}{143 - 211} \text{ gives } \frac{\mathrm{m} - 1500}{250} = \frac{-61}{-68}$$

$$\mathrm{m} = 1500 + 250\left(\frac{61}{68}\right) = \mathrm{L.E.}\,1724.26$$

Therefore, the median for the weekly income m = L.E. 1724.26 (the same result obtained before).

(b) To find the 1st and the 3rd quartiles:

Since $\frac{\mathbf{3}}{\mathbf{4}}$ of the data is more than the 1st quartile, then the order of the 1st quartile = $\frac{\mathbf{3}}{\mathbf{4}}\sum \mathbf{f_i}$ = $\frac{\mathbf{3}}{\mathbf{4}}(\mathbf{300}) = \mathbf{225}$

Income	F_D
M.T. 1250	256
M.T. Q_1	225
M.T. 1500	211

$$\frac{\mathrm{Q_1} - 1250}{1500 - 1250} = \frac{225-256}{211-256} \text{ gives } \frac{\mathrm{Q_1} - 1250}{250} = \frac{-31}{-45}$$

$$\mathrm{Q_1} = 1250 + 250\left(\frac{31}{45}\right) = 1422.22\ \mathrm{L.E}$$

Therefore, the 1st quartile for the weekly income Q_1 = L.E. 1422.22 (the same result as before).

similarly, since $\frac{\mathbf{1}}{\mathbf{4}}$ of the data is more than the 3rd quartile, then the order of the 3rd quartile = $\frac{\mathbf{1}}{\mathbf{4}}\sum \mathbf{f_i} = \frac{\mathbf{1}}{\mathbf{4}}(\mathbf{300}) = \mathbf{75}$

Income	F_D
M.T. 2000	81
M.T. Q_3	75
M.T. 2250	32

$$\frac{Q_3 - 2000}{2250 - 2000} = \frac{75 - 81}{32 - 81} \text{ gives } \frac{Q_3 - 2000}{250} = \frac{-6}{-49}$$

$$Q_3 = 2000 + 250\left(\frac{6}{49}\right) = \text{L.E.}\,2030.61$$

Therefore, the 3rd quartile for the weekly income Q_3 = L.E. 2030.61 (the same result obtained before).

(c) If 60% of families get less than x, then to find the value of x, note that x has the same definition of the 6th decile or the 60th percentile, so that, we have the following steps:

Since $\frac{\mathbf{4}}{\mathbf{10}}$ (or $\frac{\mathbf{40}}{\mathbf{100}}$) of the data is more than the 6th decile (or the 60th percentile), then the order of the 6th decile = $\frac{4}{10}\sum f_i$ = $\frac{4}{10}(300) = 120$

Income	F_D
M.T. 1750	143
M.T. d_6	120
M.T. 2000	81

$$\frac{\mathrm{d}_6 - 1750}{2000 - 1750} = \frac{120 - 143}{81 - 143} \text{ gives } \frac{\mathrm{d}_6 - 1750}{250} = \frac{-23}{-62}$$

$$\mathrm{d}_6 = 1750 + 250\left(\frac{23}{62}\right) = \mathrm{L.E.}\,1842.74$$

So, 60% of families get less than x = L.E. 1842.74 (the same preceding result).

(d) To find the percentage of families whose weekly incomes are less than L.E. 1650, note that:

The number of families whose incomes are less than L.E. 1650 = 300 – number of families whose incomes are more than L.E. 1650

Let k be the number of families with income more than L.E. 1650, then we have the following calculations:

Income	F_D
M.T. 1500	211
M.T. 1650	k
M.T. 1750	143

$$\frac{1650 - 1500}{1750 - 1500} = \frac{\mathrm{k} - 211}{143 - 211} \text{ gives } \frac{150}{250} = \frac{\mathrm{k} - 211}{-68}$$

$$\mathbf{k} = 211 - 68\left(\frac{150}{250}\right) = 170.2 \approx 170$$

The number of families whose incomes are less than L.E. 1650 is:

300 – k = 300 – 170 = 130

The percentage of families whose incomes are less than L.E. 1650 is $\frac{129.8}{300} \times 100\% = 43.26\%$

(e) To find the number of families with incomes between 1350 and L.E. 1800, we have the following steps:

- Let the number of families whose incomes are more than L.E. 1350 = a
- Let the number of families with incomes more than L.E. 1800 = b

The number of families whose incomes are between 1350 and L.E. 1800 = (a – b)

Therefore, we have to find each of ‘a’ and ‘b’ as follows:

Income	F_D
M.T. 1250	256
M.T. 1350	a
M.T. 1500	211

$$\frac{1350-1250}{1500-1250}=\frac{a-256}{211-256} \text{ leads to } \frac{100}{250}=\frac{a-256}{-45}$$

$$a=256-45\left(\frac{100}{250}\right)=238 \text{ families}$$

Similarly, to find the value of b, we have the following calculations:

Income	F_D
M.T. 1750	143
M.T. 1800	b
M.T. 2000	81

$$\frac{1800-1750}{2000-1750}=\frac{b-143}{81-143} \text{ leads to } \frac{50}{250}=\frac{b-143}{-62}$$

$$b=143-62\left(\frac{50}{250}\right)=130.6\cong 131 \text{ families}$$

Therefore, the number of families whose incomes are between L.E.1350 and L.E. 1800 is

a – b = 238 – 131 = 107 families

(f) To find the number of families with incomes more than L.E. 1850,

Let the number of families with incomes are more than L.E. 850 = k.

Income	F_i
1750	143
1850	k
2000	81

$$\frac{1850-1750}{2000-1750}=\frac{\mathrm{k}-143}{81-143} \text{ gives } \frac{100}{250}=\frac{\mathrm{k}-143}{-62}$$

$$\mathrm{k}=143-62\left(\frac{100}{250}\right)=118.2\approx 118$$

The number of families whose incomes are more than L.E. 1850 is **118** families.

3.4 Mode:

Sometimes a set of data is obtained where it is appropriate to measure a representative (central tendency) value in terms of ‘popularity’. For example, if a shop sold television sets, the answer to the question ‘what price does the average television set sell at? it is probably best given as the price of the best-selling television. This value is the mode. In this type of questions, the mode would be more representative of the data than, for instance, the mean or median.

3.4.1 Mode of Ungrouped Data:

Definition:

Mode of Ungrouped Data:
The mode of a data set is the value which occurs most often (frequently) or is the most typical value.

Example 16:

If the scores of 10 students were 76, 82, 73, 95, 82, 91, 96, 65, 73, 82. Then the mode would be 82 since this value occurred most often.

In case of ungrouped data, the mode may not exist and sometimes there are more than one mode for the data set. The following examples explain these cases.

Example 17:

Given the temperature of 7 successive days as follows:

35 , 34 , 36 , 33 , 38 , 39 , 37

In such case the mode doesn’t exist since each value happened only once.

Example 18:

Given the working experience of 9 employees:

7 , 12 , 10 , 7 , 16 , 21 , 15, 27, 16

The data set has 2 modes which are 7 and 16, since they have equal frequency and repeated more than other values.

3.4.2 Mode of Grouped Data:

For a grouped frequency distribution, the mode (in line with the mean and median) can’t be determined exactly and so must be estimated. However, we can use the modal class to find out the value of the mode.

Definition:

Modal Class:

The modal class is that class which has the largest frequency.

This section discusses how to find the mode of grouped data using 3 different methods which are:

(a) The method of the midpoint of modal class.

(b) The method of the moments.

(c) The method of differences.

(1) Method of The Midpoint of The Modal Class:

$$\textbf{Mode} = \frac{\mathbf{L_m + U_m}}{\mathbf{2}}$$

Where

$\mathbf{L_m}$: The lower limit of the modal class

$\mathbf{U_m}$: The upper limit of the modal class

The midpoint of a modal class is the simplest method, but it is the least accurate.

Example 19:

The following frequency distribution illustrates the hourly wage of 50 employees:

Wage	30-	40-	50-	60-	70-	80-	90-100	Total
Number of Employees	1	4	7	14	11	8	5	50

Find the mode using the method of the midpoint of the modal class.

Solution:

- The modal class is (60 – 70), therefore $\mathbf{L_m = 60}$ and $\mathbf{U_m = 70}$
- The mode = $\frac{\mathbf{L_m + U_m}}{\mathbf{2}} = \frac{\mathbf{60 + 70}}{\mathbf{2}}$ **=** 65 L.E

(2) Method of The Moments:

$$\textbf{Mode} = \mathbf{L_m} + \frac{\mathbf{f_s}}{\mathbf{f_p + f_s}} \times \mathbf{W_m}$$

Where

L_m: Lower limit for the modal class

f_p: Frequency of the class preceding the modal class

f_s: Frequency for the class succeeding the modal class

W_m: Width of the modal class

Note that, although this method considers the two classes; preceding and succeeding the modal class, it ignores the modal class itself.

Example 20:

Solve the previous example using moments method.

Solution:

- The modal class is (60 – 70), Therefore:

50-	60-	70-
7	14	11

$\mathbf{L_m = 60}$, $\mathbf{f_s = 11}$, $\mathbf{f_p = 7}$, $\mathbf{w_m = 10}$

- Mode = $L_m + \left(\frac{f_s}{f_p + f_s}\right) \times W_m = 60 + \left(\frac{11}{7 + 11}\right) \times 10$

= L.E. 66.61

(3) Method of Differences:

$$\text{Mode} = \mathbf{L_m} + \frac{\mathbf{f_m - f_p}}{(\mathbf{f_m - f_p}) + (\mathbf{f_m - f_s})} \times \mathbf{W_m}$$

Where

L_m: lower limit of the modal class.
f_m: frequency of the modal class.
f_p: frequency of the class preceding the modal class.
f_s: frequency of the class succeeding the modal class.
W_m: width of the modal class.

Note that this method considers the frequency of the modal class in addition to the preceding and succeeding frequencies. **Therefore,** it is considered the most accurate method.

Example 21:

Solve Example 19 using the method of principle of moments.

Solution:

- The modal class is (60 – 70), therefore

50-	60-	70-
7	14	11

$$\mathbf{L_m = 60}\ ,\ \ \mathbf{f_m = 14}\ ,\ \ \mathbf{f_s = 11}\ ,\ \ \mathbf{f_p = 7}\ ,\ \ \mathbf{W_m = 10}$$

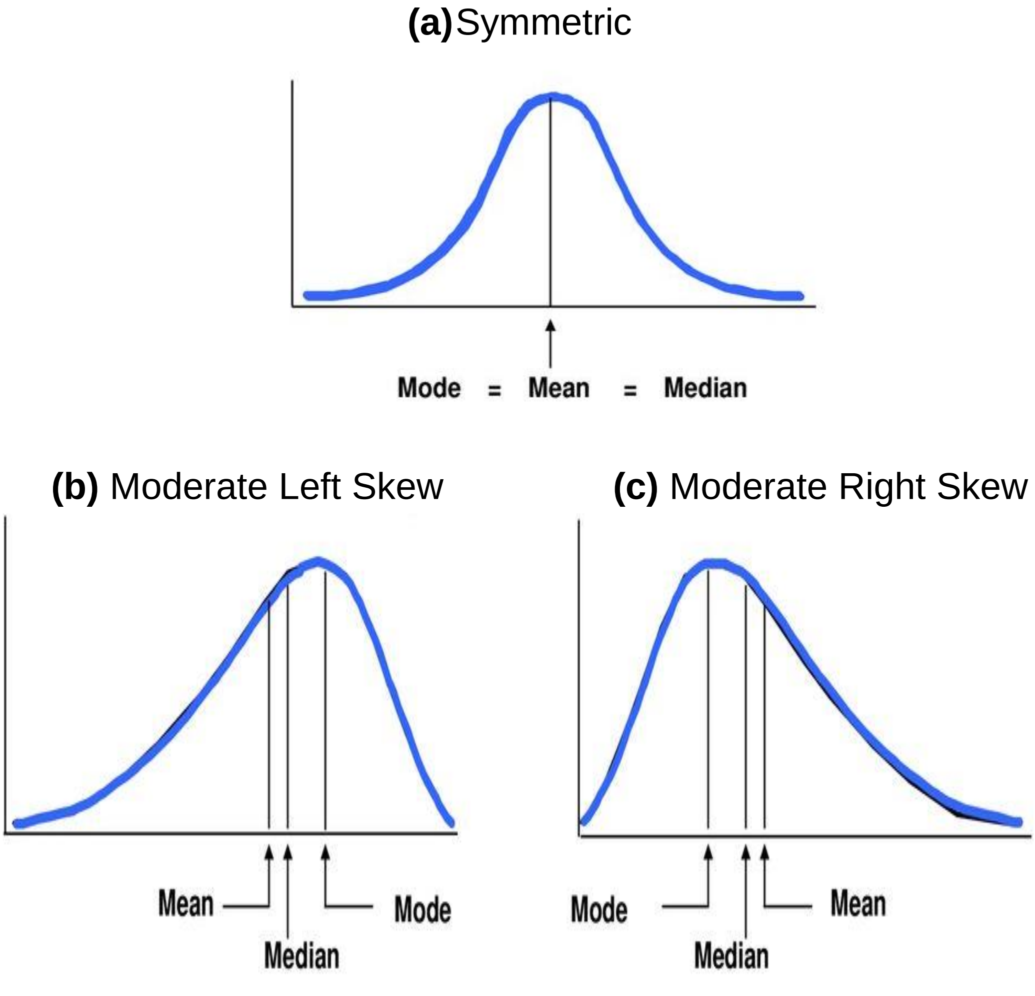
- Mode = $\mathrm{L_m} + \dfrac{\mathrm{f_m - f_p}}{(\mathrm{f_m - f_p}) + (\mathrm{f_m - f_s})} \times \mathrm{w_m}$

$$= 60 + \frac{14-7}{(14-7)+(14-11)} \times 10$$

$$= 60 + 10\left(\frac{7}{7+3}\right) = 67$$

3.5 Relationship Between Mean, Median and Mode:

Frequency curves of distributions may be relatively symmetric, but more often are skewed to some extent. Typical examples of this are distributions of wages, company turnover or times to component failure and it is of some interest to know the approximate relative positions of the three main central tendency measures, the mean, median and mode. Figure 3.1 shows these positions for moderately left skewed, symmetric and moderately right skewed distributions.



(a) Symmetric

(b) Moderate Left Skew

(c) Moderate Right Skew

Figure (3-1)
Graphical Position of Mean, Median and Mode

A useful aid in remembering the positions of the three central tendency measures in Figure 3.1 is to use the following characteristics of the three measures.

- The mode is the item that occurs most frequently and so it must lie under the main ‘hump’.
- The mean is the measure that is most affected by extremes and so it must lie towards the ‘tail’ of the distribution (except, of course, for a symmetric distribution).
- The median is the middle item and it also lies in the middle of the other two central tendency measures (but slightly closer to the mean by a factor of 2 to 1 approximately).

(1) For Symmetric Distributions:

In such case frequencies around the middle class are symmetrical and the following relation holds

Mean = Median = Mode

[see diagram (a) in Figure 3.1 for graphical representation]

Example 22:

Given the frequency distribution of statistics grades for 65 students

Grade	10-	22-	34-	46-	58-	70-	82-94	Total
Number of students	4	8	13	15	13	8	4	65

Find the mean, median and mode. Comment.

Solution:

- **Mean:**

$$\bar{\mathbf{x}} = \frac{\sum \mathbf{x_i f_i}}{\sum \mathbf{f_i}} = \frac{\mathbf{3380}}{\mathbf{65}} = \mathbf{52}$$ See the table given below.

Grade	**f**	**Class Midpoint (x)**	**xf**
10-	4	16	64
22-	8	28	224
34-	13	40	520
46-	15	52	780
58-	13	64	832
70-	8	76	608
82-94	4	88	352
Total	65	-	3380

- **Median:**

Income	**CF**
less than 10	0
less than 22	4
less than 34	12
less than 46	25
less than 58	40
less than 70	53
less than 82	61
less than 94	65

Median Point = $\frac{1}{2}\sum f_i$ = $\frac{1}{2}(65)$ = **32.5**

Grade	F_i
L.T. 46	25
L.T . m	32.5
L.T. 58	40

$\frac{m-46}{58-46} = \frac{32.5-25}{40-25}$ gives $\frac{m-46}{12} = \frac{7.5}{15}$

$$m = 46 + 12\left(\frac{1}{2}\right) = 118.2$$

- **Mode:**

The modal class is (46 – 58)

34-	46-	58-
13	15	13

$$\mathbf{L_m = 46} \quad , \quad \mathbf{f_s = 13} \quad , \quad \mathbf{f_p = 13} \quad , \quad \mathbf{W_m = 12}$$

$$\text{Mode} = \mathrm{L_m} + \frac{\mathrm{f_m} - \mathrm{f_p}}{(\mathrm{f_m} - \mathrm{f_p}) + (\mathrm{f_m} - \mathrm{f_s})} \times \mathrm{W_m}$$

$$= 46 + \frac{15-13}{(15-13)+(15-13)} \times 12$$

$$= 46 + \frac{2}{2+2} \times 12 = 52$$

We can conclude that:

Mean = Median = Mode = 52

This is because the frequency distribution is symmetric.

(2) For Skewed Distributions:

Skewed frequency distributions can be either right or left skewed depending on the side where frequencies are concentrated. For Right skewed distributions, frequencies are more concentrated in the lower classes. In contrast, for left skewed distributions, frequencies are more concentrated in the higher classes.

[See diagrams (b) and (c) in Figure 3.1 for graphical representation].

For moderately skewed distributions, simple relationships between the mean, median and mode can be worked out. Given the fact that the median lies between the mean and mode, but closer to the mean by a factor of 2 to 1, the relationship: mean - mode = 3 (mean - median) should be approximately true. Using this relationship, any one of the three measures of central tendency can be expressed in terms of the other two measures with a little algebraic re-arrangement. Namely:

- Mean = $\frac{\mathbf{3(Median) - Mode}}{\mathbf{2}}$
- Median = $\frac{\mathbf{2(Mean) + Mode}}{\mathbf{3}}$
- Mode = 3(Median) – 2(Mean)

Positively skewed: mode < median < mean
Negatively skewed: mean < median < mode

Example 23:

For the frequency distribution given in Example 8, Example 13 and Example 21 where median = 69.29 and mode = 67.

- Find the mean as a formula of median and mode.
- Is the frequency distribution negatively skewed? Why?

Solution:

- Mean = $\frac{\mathbf{3(Median) - \ Mode}}{\mathbf{2}} = \frac{\mathbf{3(69.29) - 67}}{\mathbf{2}}$ = 70.435
- Since (mean > median > mode), then the distribution is positively skewed.

3.6 Characteristics of the Measures of Central Tendency:

This section illustrates the advantages and disadvantages for central tendency measures.

3.6.1 Characteristics of The Mean:

Advantages:

1- It is considered as the most important and the most applicable measure of central tendency. It is widely used in applications.

2- In contrary to the median and mode, all the values are considered in computing the mean.

3- The data set do not have to be arranged as the case of the median.

4- Amenable to further statistical analysis and algebraic calculations.

Disadvantages:

1- The mean is sensitive to outliers (extreme values). For example, consider the salaries of 5 employees as 3500, 5000, 4200, 4300 and L.E. 60000. The mean is 15400 which gives unrealistic picture about the level of salaries. Note that the salary 60000 is an outlier.

2- The mean can’t be calculated for frequency distributions with open-ended classes. We can handle this problem by:

- Getting rid of open classes and calculating the mean for the rest of the data, but we will lose information especially when these classes have large frequencies.
- Closing the open classes by supposing a value for the lower limit of the lowest class or a value for the upper limit of the highest class which is subjective and requires experience to deal with.
- In symmetrical or nearly symmetrical frequency distributions, we can use the relationship between mean, median and mode to estimate the mean by using Pearson’s relationship which is as follows:

$$\mathbf{Mean} = \frac{\mathbf{3 \times Median - Mode}}{\mathbf{2}}$$

3- It can’t be used for qualitative data.

3.6.2 Characteristics of The Median:

Advantages:

1- It is the middle value of the data after arranging in ascending (or descending) order, so it is not affected by extreme values like the mean.

2- It can be calculated for frequency distributions with open-ended classes.

Disadvantages:

1- Only a part of the data set is used for computing the value of the median. Therefore, it is less accurate than the mean.

2- It isn’t amenable for algebraic calculations.

3.6.3 Characteristics of The Mode:

Advantages:

1- Occasionally, it is considered as an alternative to the mean or median when the situation calls for the ‘most popular’ value (or attribute) for data.

2- It can be used for qualitative data.

3- It is easy to be understood, not difficult to calculate and can be used when a frequency distribution has open- ended classes.

Disadvantages:

1- Only a part of the data set is used for computing the value of the mode. It is the least accurate measure of central tendency.

2- Sometimes it has more than one value, if any.

3- Although the mode usefully ignores isolated extreme values, it is thought to be too much affected by the modal class when a distribution is significantly skewed.

4- Like the median, the mode isn’t used in advanced statistical analysis.

3.7 Effect of Shifting and Scaling:

Sometimes the entire data set can be changed either by shifting (adding or subtracting the same value) or scaling (multiplying or dividing by the same value). In such cases, it will be useful to know how the measures of central tendency are affected by these changes without repeating any calculations.

(1) The Effect of Shifting:

If we add the same value to the entire data set or subtract it to obtain a new data set, the measures of central tendency for the new data set can be found easily using the following relation:

The measures of central tendency for the new data set change by the same value we add to the entire data set or we subtract from it.

Let the value we add or subtract ‘a’, then the measures of central tendency for the new data set can be written as:

- The mean of the new data set

  $= \bar{\mathbf{x}} + \mathbf{a}$ (In case of addition)

  $= \bar{\mathbf{x}} - \mathbf{a}$ (In case of subtraction)

- The median of the new data set

  $= \mathbf{m} + \mathbf{a}$ (In case of addition)

  $= \mathbf{m} - \mathbf{a}$ (In case of subtraction)

- The mode of the new data set

  $= \mathbf{mode} + \mathbf{a}$ (In case of addition)

  $= \mathbf{mode} - \mathbf{a}$ (In case of subtraction)

Example 24:

For the frequency distribution given in Example 8, Example 13 and Example 21 where mean = 69.8, median = 69.29 and mode = 67. Find the measures of central tendency for the new data set after:

(a) Adding L.E. 10 to the wages of all employees.

(b) Subtracting L.E. 5 from the wages of all employees.

Solution:

(a) In case of adding 10 L.E to the wage for all employees, then:

- The **mean** of the new data set = $\mathbf{\bar{x} + a = 69.8 + 10 = 79.8}$
- The **median** of the new data set = $\mathbf{m + a = 69.29 + \ 10 = 79.29}$
- The **mode** of the new data set = $\mathbf{mode + a = 67 + 10 = 77}$

(b) In case of subtracting L.E. 5 to the wages of all employees:

- The **mean** for the new data set is

  $\mathbf{\bar{x} - a = 69.8 - 5 = L.E. 64.8}$
- The **median** for the new data set is

  $\mathbf{m - a = 69.29 - 5 = 64.29}$ L.E
- The **mode** for the new data set is

  $\mathbf{Mode - a = 67 - 5 = L.E. 62}$

(2) The Effect of Scaling:

If we multiply or divide the entire data set by an arbitrary value (constant), then all measures of central tendency for the new data set can be found easily using the following relation:

The measures of central tendency for the new data set is scaled by the same value (constant) by which we multiply or divide.

Let us assume that the value we multiply or divide 'a', then the measures of central tendency for the new data set can be written as:

- The **mean** of the new data set = $\mathbf{\bar{x} \times a}$ (For multiplication)
  = $\mathbf{\bar{x} / a}$ (For division)
- The **median** of the new data set = $\mathbf{m \times a}$ (For multiplication)
  = $\mathbf{m / a}$ (For division)
- The **mode** of the new data set = $\mathbf{mode \times a}$ (For multiplication)
  = $\mathbf{mode / a}$ (For division)

Example 25:

For the frequency distribution given in Example 8, Example 13 and Example 21, where the mean of wages = L.E. 69.8, the median of wages = L.E. 69.29 and the mode of wages = L.E. 67, find the measures of central tendency for the new data set after:

(a) Increasing the wage by 10% for all employees.

(b) Decreasing the wage by 5% for all employees.

Solution:

(a) In case of increasing the wage by 10% for all employees, then:

new wage = old wage + 10% old wage
= old wage + 0.1(old wage) = (1 + 0.1) old wage
= 1.1 × old wage

Therefore, for a = 1.1, then:

- The **mean** for the new data set $= \bar{x} \times a = 69.8 \times 1.1 = \text{L.E.}\, 76.78$

- The **median** for the new data set $= m \times a = 69.29 \times 1.1 = \text{L.E.}\, 76.22$

- The **mode** for the new data set $= \text{mode} \times a = 67 \times 1.1 = \text{L.E.}\, 73.7$

(b) In case of decreasing the wage by 5% for employees, then:

new wage = old wage - 5% old wage
= wage old - 0.05 (old wage) = (1 - 0.05)(old wage)
= 0.95 × old wage

Therefore, For a = 0.95, then:

- The mean for the new data set $= \bar{x} \times a = 69.8 \times 0.95 = \text{L.E.}\, 66.31$

- The median for the new data set = $\mathrm{m} \times \mathrm{a} = 69.29 \times 0.95$
$= \mathrm{L.E.}\,65.83$
- The mode for the new data set = $\mathrm{mode} \times \mathrm{a} = 67 \times 0.95$
$= \mathrm{L.E.}\,63.65$

Exercises

1- The following data give the ages for 6 persons:

65 82 92 86 5 90

a. Find the mean and the median for these data.

b. Using the results of part (a), how can you investigate that this data set contains outliers? Explain?

c. If you dropped the outlier and the values of the mean and median were recalculated, which of the two measures is expected to change by a larger amount? No calculations required.

d. Which of the mean or median is considered a better measure for these data? Explain?

2- The following data give the daily wages (L.E) earned by a sample of 30 workers as shown below.

36	28	22	44	30	26	49	24	33	34
25	31	39	33	28	37	42	27	23	27
32	25	34	29	43	32	26	20	28	35

a. Prepare a frequency distribution for these data.

Hint: Use Sturges' rule formula to determine the number of classes.

b. Construct a cumulative frequency distribution.

c. On the basis of the cumulative distribution obtained in Part (b), find the percentage of workers with daily wages of at least L.E. 27.

3- In a medium sized city, there are 100 houses for sale of a similar size.

The frequency distribution of prices is as follows.

Price ($10,000)	10-	20-	30-	40-	50-60	Total
No. of Houses	26	35	25	9	5	100

a. On the basis of inspection only (without performing any calculations), what might you expect for the value of the mean compared with that of the median? Justify your answer.

b. Find the price value so that 50% of houses have prices less than this value.

c. Find the mean of prices.

d. What connection do you see between your answers in Parts (a), (b), and (c)?

4- The following frequency distribution reports the electricity cost (in dollars) for a sample of 25 two-bedroom apartments in a city.

Electric Cost	30-	40-	50-	60-	70-	80-90	Total
No. of Apartments	2	4	7	8	3	1	25

Find the mean and median of cost.

5- Consider the following frequency distribution:

Class	5-	15-	25-	35-	45-55
Frequency	3	A	6	B	2

Given that: The mean for this distribution = 29.5, the median=30, and the median Class is (25-35), determine the values of A and B.

6- The following frequency table gives the distribution of the monthly bonus payments (in hundreds of L.E.) made to 50 employees in a company.

Monthly Bonus	30-	40-	50-	60-	70-80	Total
Number of Employees	6	12	16	12	4	50

a. On the basis of inspection only (without performing any calculations), do you agree with the claim that the distribution of the monthly bonus is positively skewed? Explain?

b. Find the mean, median and mode of the monthly bonus.

c. Find the bonus value so that 75% of employees have bonuses less than this value.

7- The following data represents the ages for a sample of internet users as follows:

39	15	31	25	24	23	21	22	22	18
19	16	23	27	34	24	19	20	29	17

a. Organize these data into a frequency distribution using 5 as a width for each class.

b. Based upon the frequency distribution obtained in part (a):

(1) Prepare a “less than” cumulative frequency distribution.

(2) Find the percentage of internet users with age of at least 32 years.

8- Before admission to a college, the students have to take Basic Skills Test in fundamentals of mathematics. The scores of 25 students are recorded below out of a total maximum of 40 points.

Hint: Pass Mark = 60% of Full Mark.

15	12	15	22	28	30	19	25	24	28
10	23	16	20	26	22	18	20	27	14
12	19	21	24	32					

a. Prepare a frequency distribution for these data using 5 as width for each class.

b. Using your result in part (a), prepare a ‘Less than’ cumulative frequency distribution. Then, find the proportion of the students who passed the exam.

9- The following table represents the distribution of the hourly wages for 20 workers in a factory.

Hourly Wage	20-25	Less than 30	Less than 35	Less than 40	Less than 45
No. of Workers	2	A	14	18	20

a. If the value of the median is 32.5, find the value of:

(1) A **(2)** The mean **(3)** The mode

b. How would the measures of central tendency be affected if the hourly wages?

(1) Increased by 20%. **(2)** Decreased by 10%.

(3) Increased by 10 L.E. **(4)** Decreased by 5 L.E.

Chapter (4)
Measures of Dispersion and Skewness

In the previous chapter, we have studied measures of central tendency such as mean, mode, median of ungrouped and grouped data. The averages are representatives of a frequency distribution. But they fail to give a complete picture of the distribution because they do not tell anything about how the observations are scattered within the distribution. For example, consider the following data about the wages of two groups of employees as follows:

Group1

145 , 160 , 191 , 172 , 184 , 195 , 179 , 176 , 155 , 163

Group2

155 , 184 , 132 , 176 , 162 , 148 , 115 , 170 , 232 , 146

If we want to compare between both groups based on average, we may say that level of wages in both groups are equal because they have the same mean value = L.E 172. But comparing both groups based on average only is not accurate, because in group1 wages ranged between L.E. 145 and L.E. 195 (i.e., Range = 50), whereas in group2 wages ranged between L.E. 132 and L.E. 232 (i.e., Range = 100), which indicates that values in group1 are closer to the mean and therefore are more homogenous, whereas values in group 2 are less concentrated around the mean.

> Measures of dispersion describe how spread out or scattered a set or distribution of numeric data is?

4.1 Measures of Dispersion: This section is concerned with some important absolute measures of dispersion such as range, semi-interquartile range, variance, and standard deviation.

4.1.1 The Range:

The range is defined as the numerical difference between the lowest and largest values of the items in a data set or distribution.

A- For Ungrouped Data:

Range = Largest value – Lowest value = H – L

Example 1:

The following data represents ages of 10 persons (years):

31 , 18 , 27 , 41 , 53 , 32 , 56 , 43 , 17 , 22

Find the range.

Solution:

- Arrange the data in an ascending order

17 , 18 , 22 , 27 , 31 , 32 , 41 , 43 , 53 , 56

- Range = Highest Value – Lowest Value = 56 – 17 = 39 years

Example 2:

The following data give the hours worked last week by 5 employees of a company:

42 34 40 85 36

Find the range.

Solution:

- Arrange the data in an ascending order

34 36 40 42 85

- Range = Highest – Lowest Value = 85 – 34 = 51 hours

Hint: For determining the lowest and highest values for a data set, there is no need to arrange data in an ascending (or descending order). It should be pointed out that $\mathbf{x_r}$ denotes the $\mathbf{r^{th}}$ value in a raw data set, whereas $\mathbf{x_{(r)}}$ denotes the $\mathbf{r^{th}}$ value in data values after arranging in ascending order. Therefore, $\mathbf{x_{(1)}}$ represents the lowest value, while $\mathbf{x_{(n)}}$ is the highest value. So, the range formula can be written as:

$$\textbf{Range} = \mathbf{x_{(n)} - x_{(1)}}$$

B- For Grouped Data:

In case of grouped data, we can find the range using two different formulas.

(1) Range = Upper limit of the last class – Lower limit of the first class.
(2) Range = Midpoint of the last class – Midpoint of the first class.

Example 3:

Given the following frequency distribution table:

Electricity Cost	30-	40-	50-	60-	70-	80-90	Total
No. of Apartments	3	8	12	16	7	4	50

Find the range.

Solution:

(1) Range = Upper limit of the last class – Lower limit of the first class

= 90 – 30 = 60

(2) Range = Midpoint of the last class – Midpoint of the first class

$$= \frac{80+90}{2} - \frac{30+40}{2} = 85 - 35 = 50$$

Example 4:

The following is a frequency distribution for the ages of a sample of 20 employees at a company.

Age	20-	30-	40-	50-	60-70	Total
No. of employees	4	5	6	3	2	20

Find the range.

Solution:

(1) Range = Upper limit of the last class – Lower limit of the first class

= 70 – 20 = 50

(2) Range = Midpoint of the last class – Midpoint of the first class

$$= \frac{60 + 70}{2} - \frac{20 + 30}{2} = 65 - 25 = 40$$

Characteristics of the Range:

1- The range is a simple concept and easy to calculate.

2- The range is not used in further advanced statistical work.

3- It Can't be calculated for the open frequency distributions because in such case we can't determine the upper limit of the last class neither the lowest limit of the first class.

4- The major disadvantage of the range is the fact that it only takes two values into account (the lowest and largest) and is thus too obviously affected by extreme values. For example, the following data represents the weights of two groups in an ascending order as follows:

Group (A): 10 , 11 , 12 , 13 , 14 , 15 , 16 , 17 , 18 , 84
Group (B): 35 , 41 , 45 , 58 , 63 , 73 , 82 , 90 , 95 , 98

Range of group (A) = 84 – 10 = 74 kg
Range of group (B) = 98 – 35 = 63 kg

According to range values for both groups we may conclude that data in group (A) is more scattered than data in group (B) which is opposite to the fact that data in group (A) is less scattered than data in group (B). The reason why group (A) has range value greater than that for group (B) is the existence of extreme value (84) in group (A), which inflated the value of the range. After deleting the extreme value and recalculating the value of the range for both groups, range of group (A) = 18 – 10 = 8 and range of group (B) = 95 – 35 = 60 which indicates that data in group (A) is much less scattered than data in group (B). The reason why the results has been affected is the deletion of the extreme value from group (A).

4.1.2 Inter-Quartile Range (IQR):

Definition:

The **interquartile range** is a difference between the first and third quartiles.

$$\mathrm{IQR} = Q_3 - Q_1$$

Example 5:

For the frequency distribution given in Example 4, find the inter-quartile range.

Solution:

Class	CF
Less than 20	0
Less than 30	4
Less than 40	9
Less than 50	15
Less than 60	18
Less than 70	20

Q_1 Point = $20(\frac{1}{4})$ = 5

Age	F_i
L.T. 30	4
L.T. Q_1	5
L.T. 40	9

$\frac{Q_1 - 30}{40 - 30} = \frac{5-4}{9-4}$ gives $\frac{Q_1 - 30}{10} = \frac{1}{5}$

$Q_1 = 30 + 10\left(\frac{1}{5}\right) = 32 \text{ years}$

Q_3 Point = $20(\frac{3}{4})$ = 15

Age	F_i
L.T. 50	15

$\mathbf{Q_3 = 50 \text{ years}}$

IQR = $Q_3 - Q_1$ = 50 – 32 = 18

4.1.3 Quartile Deviation (QD):

The quartile deviation (is also known as semi-interquartile range) is the half-distance between the first quartile and the third quartiles.

$$\text{QD} = \frac{Q_3 - Q_1}{2},$$

It is not affected by extreme values (outliers).

Example 6:

For the frequency distribution given in **Example 3**, find the quartile deviation.

Solution:

Class	CF
Less than 30	0
Less than 40	3
Less than 50	11
Less than 60	23
Less than 70	39
Less than 80	45
Less than 90	50

Q_1 Point = $50(\frac{1}{4}) = 12.5$

Cost	F_i
L.T. 50	11
L.T. Q_1	12.5
L.T. 60	23

$$\frac{Q_1 - 50}{60 - 50} = \frac{12.5 - 11}{23 - 11} = \frac{Q_1 - 50}{10} = \frac{1.5}{12}$$

$$Q_1 = 50 + 10\left(\frac{1.5}{12}\right) = 51.25$$

Q_3 Point = $20(\frac{3}{4}) = 15$

Cost	F_i
L.T. 60	23
L.T. Q_1	37.5
L.T. 70	39

$$\frac{Q_3 - 60}{70 - 60} = \frac{37.5 - 23}{39 - 23} \text{ gives } \frac{Q_3 - 60}{10} = \frac{14.5}{16}$$

$$Q_3 = 60 + 10\left(\frac{14.5}{16}\right) = 69.0625$$

$$\mathrm{QD} = \frac{\mathrm{Q}_3 - \mathrm{Q}_1}{2} = \frac{69.0625 - 51.25}{2} = 8.906$$

Example 7:

The time taken for the weekly maintenance of a group of machines in a workshop over the past 25 weeks is shown in the following table.

Maintenance Time	0-	2-	4-	6-	8-10	Total
No. of Weeks	2	6	10	5	2	25

Find the quartile deviation.

Solution:

Class	**CF**
Less than 0	0
Less than 2	2
Less than 4	8
Less than 6	18
Less than 8	23
Less than 10	25

Q_1 Point = $25(\frac{1}{4})$ = 6.25

Time	**$\mathbf{F_i}$**
L.T. 2	2
L.T. Q_1	6.25
L.T. 4	8

$$\frac{\mathrm{Q}_1 - 2}{4 - 2} = \frac{6.25 - 2}{8 - 2} \text{ gives } \frac{\mathrm{Q}_1 - 2}{2} = \frac{4.25}{6}$$

$$\mathrm{Q}_1 = 2 + 2\left(\frac{4.25}{6}\right) = 3.417 \text{ hours}$$

Q_3 Point = $25(\frac{3}{4})$ = 18.75

Time	F_i
L.T. 6	18
L.T. Q_3	18.75
L.T. 8	23

$\frac{Q_3 - 6}{8 - 6} = \frac{18.75 - 18}{23 - 18}$ gives $\frac{Q_3 - 6}{2} = \frac{0.75}{5}$

$$Q_3 = 6 + 2\left(\frac{0.75}{5}\right) = 6.3 \text{ hours}$$

$$\mathbf{QD} = \frac{\mathbf{Q_3 - Q_1}}{\mathbf{2}} = \frac{\mathbf{6.3 - 3.417}}{\mathbf{2}} = \mathbf{1.4415} \textbf{ hours}$$

4.1.4 Mean Deviation (MD):

The mean deviation is a measure of dispersion that gives the average absolute difference (i.e. ignoring ‘minus’ signs) between each item and the mean.

1- Mean Deviation for Ungrouped Data:

Mean Deviation (MD) of a Set of Values:

$$\mathbf{MD} = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$

Example 8:

Calculate the mean deviation of 43 , 75 , 48 , 39 , 51 , 47 , 50 , 47.

Solution:

First determine the mean as: $\frac{\sum_{i=1}^{n} x_i}{n} = \frac{400}{8} =$ 50 and then we find absolute deviations as follows:

$\mathbf{x_i}$	$\mathbf{x_i - \bar{x}}$	$\mathbf{|x_i - \bar{x}|}$
43	-7	7
75	25	25
48	-2	2
39	-11	11
51	1	1
47	-3	3
50	0	0
47	-3	3
400	0	52

$$\mathrm{MD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{52}{8} = 6.5$$

In other words, each value in the set is, on average, 6.5 units away from the common mean.

2- Mean Deviation for Grouped Data:

Mean Deviation of Grouped data:

$$\mathrm{MD} = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$$

Example 9:

For the frequency distribution given in Example 7, find the mean deviation.

Maintenance Time	No. of Weeks (f_i)	Midpoint (x_i)	$x_i f_i$	$|x_i - \bar{x}|$, $\bar{x} = 4.92$	$|x_i - \bar{x}|\ f_i$
0-	2	1	2	3.92	7.84
2-	6	3	18	1.92	11.52
4-	10	5	50	0.08	0.8
6-	5	7	35	2.08	10.4
8-10	2	9	18	4.08	8.16
Total	25	-	123	-	38.72

Note: For a grouped frequency distribution, $\mathbf{x_i}$ represents the class midpoint for the $\mathbf{i^{th}}$ class.

Mean: $\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{123}{25} = 4.92$

$$\mathbf{MD} = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i} = \frac{38.72}{25} = 1.5488$$

4.1.5 Standard Deviation and Variance:

The standard deviation can be defined as 'the root of the mean of the squares of deviations from the common mean' of a set of values.

1- Standard Deviation and Variance of Ungrouped Data:

Standard Deviation of a Set of Values:

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

Computational Formula for the Standard Deviation:

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

Computational Formula for the variance (S^2):

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

Example 10:

For the following data set: 43 , 75 , 48 , 51 , 47 , 50 , 47 , 40 , 48 find the standard deviation and the variance.

Solution:

$\mathbf{x_i}$	$\mathbf{x_i^2}$	$\mathbf{x_i}$	$\mathbf{x_i^2}$
43	1849	47	2209
75	5625	50	2500
48	2304	47	2209
51	2601	40	1600
51	2601	48	2304

From this table: $\sum_{i=1}^{n} x_i = 509$, $\sum_{i=1}^{n} x_i^2 = 25802$

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

$$= \sqrt{\frac{25802}{10} - \left(\frac{500}{10}\right)^2} = 8.96$$

Variance = $s^2 = (8.96)^2 = 80.2816$

Example 11:

For data in Example 2, find the standard deviation and the variance.

Solution:

$\mathbf{x_i}$	$\mathbf{x_i^2}$
42	1764
34	1156
40	1600
85	7225
36	1296
237	13041

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

$$= \sqrt{\frac{13041}{5} - \left(\frac{237}{5}\right)^2} = 19.01$$

Variance = s^2 = $(19.01)^2$ = 361.4

2- Standard Deviation for Grouped Data:

For large sets of data, a frequency distribution is normally compiled, and the computational formula for the standard deviation needs to be duly adapted. The adapted formula is given as follows using three different methods.

(a) Direct Method:

Computational formula for the standard deviation of a frequency distribution using the direct method:

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

Example 12:

The following table presents the distribution of the monthly income (in thousands of Egyptian pounds). Calculate the mean, standard deviation and variance of monthly income using the direct method.

Monthly Income	25-	30-	35-	40-	45-50	Total
Number of Households	2	3	5	22	18	50

Solution:

The standard layout and calculations are shown in the following table:

Monthly Income	**Number of Households (f)**	**Midpoint (x)**	**xf**	**x^2f**
25-	2	27.5	55	1512.5
30-	3	32.5	97.5	3168.75
35-	5	37.5	187.5	7031.25
40-	22	42.5	935	39737.5
45-50	18	47.5	855	40612.5
Total	50	-	2130	92062.5

Mean: $\bar{\mathbf{x}} = \frac{\sum \mathbf{x_i f_i}}{\sum \mathbf{f_i}} = \frac{\mathbf{2130}}{\mathbf{50}} = \mathbf{42.6}$

Standard Deviation: $\boldsymbol{s} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$

$$= \sqrt{\frac{92062.5}{50} - \left(\frac{2130}{50}\right)^2}$$

$$= 5.147$$

Variance = s^2 = $(5.147)^2$ = 26.49

Example 13:

For the frequency distribution table given in Example 7, find the mean, standard deviation and variance.

Solution:

The standard layout and calculations are shown in the following table:

Maintenance Time	No. of Workers (f)	Midpoint (x)	xf	X^2f
0-	2	1	2	2
2-	6	3	18	54
4-	10	5	50	250
6-	5	7	35	245
8-10	2	9	18	162
Total	25	-	123	713

Mean: $\bar{\mathbf{x}} = \frac{\sum x_i f_i}{\sum f_i} = \frac{\mathbf{123}}{\mathbf{25}} = \mathbf{4.92}$

Standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$

$$= \sqrt{\frac{713}{25} - \left(\frac{123}{25}\right)^2} = \mathbf{2.077}$$

Variance = s^2 = $(2.077)^2$ = 4.314

(b) Method of Assumed Mean:

Computational formula for the standard deviation of a frequency distribution using method of step-deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} d_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$$

Where $d_i = x_i - a$

Example 14:

For the monthly income frequency distribution in Example 12, find the mean, standard deviation and variance using the method of step-deviations.

Solution:

Income	Number of Households (f)	Midpoint (x)	d = x – 37.5	df	d^2f
25-	2	27.5	-10	-20	200
30-	3	32.5	-5	-15	75
35-	5	37.5	0	0	0
40-	22	42.5	5	110	550
45-50	18	47.5	10	180	1800
Total	**50**	**-**	**-**	**255**	**2625**

Mean: $\bar{x} = a + \left(\frac{\sum d_i f_i}{\sum f_i}\right) = 37.5 + \left(\frac{255}{50}\right) = 42.6$

Standard Deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} d_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$$

$$= \sqrt{\frac{2625}{50} - \left(\frac{255}{50}\right)^2} = 5.147$$

Variance = $S^2 = (5.147)^2 = 26.49$

(c) Method of Step - Deviation:

Computational formula for the standard deviation of a frequency distribution using the method of step-deviation:

$$S_D = \sqrt{\frac{\sum_{i=1}^{n} D_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} D_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}\ , \quad \text{then} \quad S = b \times S_D$$

Where $D_i = \frac{x_i - a}{b}$

Example 15:

For the monthly income frequency distribution in Example 12, find the mean, standard deviation and variance using the method of step-deviations.

Solution:

Monthly Income	Number of Households (f_i)	Midpoint (x_i)	$d_i = x_i - 37.5$	$D_i = d_i / 5$	$D_i f_i$	$D_i^2 f_i$
25-	2	27.5	-10	-2	-4	8
30-	3	32.5	-5	-1	-3	3
35-	5	37.5	0	0	0	0
40-	22	42.5	5	1	22	22
45-50	18	47.5	10	2	36	72
Total	**50**	**-**	**-**	**-**	**51**	**105**

Mean: $\bar{x} = a + b.\left(\frac{\sum D_i f_i}{\sum f_i}\right) = 37.5 + 5\left(\frac{51}{50}\right) = 42.6$

Since $S_D = \sqrt{\frac{\sum_{i=1}^{n} D_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} D_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$

$$= \sqrt{\frac{105}{50} - \left(\frac{51}{50}\right)^2} \quad = 1.0293$$

standard deviation: $S_x = b \times S_D = 5 \times 1.0293 = 5.298$

variance = $\mathbf{S_x^2}$ = $(5.298)^2$ = 28.069

Example 16:

In a medium sized city there are 50 houses for sale of similar size. The frequency distribution of prices is as follows.

Price ($10,000)	10-	20-	30-	40-	50-60	Total
No. of houses	14	18	12	4	2	50

Find the mean, standard deviation and variance for the prices using the method of step-deviations.

Solution:

Price ($10,000)	No. of Houses (f)	Midpoint (x)	d = x – 35	D = d/10	Df	D^2f
10-	14	15	-20	-2	-28	56
20-	18	25	-10	-1	-18	18
30-	12	35	0	0	0	0
40-	4	45	10	1	4	4
50-60	2	55	20	2	4	8
Total	**50**	**-**	**-**	**-**	**-38**	**86**

Mean: $\bar{x} = a + b\left(\frac{\sum D_i f_i}{\sum f_i}\right) = 35 + 10\left(\frac{-38}{50}\right) = 27.4$

Since $S_D = \sqrt{\frac{\sum_{i=1}^{n} D_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} D_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$

$$S_D = \sqrt{\frac{86}{50} - \left(\frac{-38}{50}\right)^2} = 1.069$$

Standard deviation: $S_x = b(S_D) = 10(1.069) = 10.69$

Variance = $S_x^2 = (10.69)^2 = 114.276$

Note: The value of the standard deviation for data values is unchanged by adding (or subtracting) a constant, say a, to (or from) each value, but is changed by multiplying (or dividing) each value by a constant, say b. That is,

$$S_x = S_{(X \pm a)} \quad \text{and} \quad S_{(bx)} = b(S_X)$$

Characteristics of the Standard Deviation:

1- The standard deviation is the natural partner of the arithmetic mean in the following respects:

- 'By definition'. The standard deviation is defined in terms of the mean.
- In further statistical analysis, there is a need to deal with one of the most commonly occurring natural distributions, called the Normal distribution, which can only be specified in terms of both the mean and standard deviation.

2- It can be regarded as truly representative of the data, since all data values are taken into account in its calculations.

3- For distributions that are not too skewed:

- Virtually, all of the items should lie within three standard deviations of the mean. i.e., range = 6 × standard deviation (approximately).
- 95% of the items should lie within two standard items deviations of the mean.

4- The standard deviation is affected by extreme values (outliers).

4.2 Measures of Relative Dispersion:

It is sometimes necessary to compare two different distributions with regard to variability. For example, if two machines were engaged in the production of identical components, it would be of considerable value to compare the variation of some critical dimensions of their output. However, the standard deviation is used as a measure for comparison only when the units in the distributions are the same and the respective means are roughly comparable (not widely different). Except in these cases, the measures of relative dispersion are more appropriate.

Now, some measures of relative dispersion are to be considered.

4.2.1 Coefficient of Variation (CV):

The coefficient of variation calculates the standard deviation as a percentage of the mean.

$$\mathbf{CV} = \frac{\mathbf{Standard\ Deviation}}{\mathbf{Mean}} \times \mathbf{100}\%$$

Since the standard deviation is being divided by the mean, the actual units of measurement cancel each other out, leaving the measure unit free and thus very useful for comparison.

Example 17:

Over a period of three months the daily number of components produced by two comparable machines was measured, giving the following statistics.

Machine A: Mean = 242.8 , sd = 20.5
sd stands for “standard deviation”

Machine B: Mean = 281.3 , sd = 23.0

Coefficient of variation for Machine A = $\frac{\mathbf{20.5}}{\mathbf{242.8}} \times \mathbf{100}\% = \mathbf{8.4}\%$

Coefficient of variation for Machine B = $\frac{\mathbf{23.0}}{\mathbf{281.3}} \times \mathbf{100}\% = \mathbf{8.2}\%$

Thus, although the standard deviation for Machine B is higher in absolute terms, the dispersion for Machine A is higher in relative terms.

Example 18:

The following table represents the rates of return over the past 6 years for two mutual funds

Fund A (x)	8.3	- 6	18.9	- 5.7	23.6	20
Fund B (y)	12	- 4.8	6.4	10.2	25.3	1.4

Which fund has higher risk. (High variability implies high risk)

Solution:

x_i	x_i^2
8.3	68.89
- 6	36
18.9	357.21
- 5.7	32.49
23.6	556.96
20	400
59.1	1451.55

y_i	y_i^2
12	144
-4.8	23.04
6.4	40.96
10.2	104.04
25.3	640.09
1.4	1.96
50.5	954.09

Fund A:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{59.1}{6} = 9.85$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} = \sqrt{\frac{1451.55}{6} - \left(\frac{59.1}{6}\right)^2}$$

$$= 12.04$$

Fund B:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{50.5}{6} = 8.42$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2} = \sqrt{\frac{954.09}{6} - \left(\frac{50.5}{6}\right)^2}$$

$$= 9.39$$

Coefficient of variation:

For fund A = $\frac{s_x}{\bar{x}} \times 100\% = \frac{12.04}{9.85} \times 100\% = 122.23\%$

For fund B = $\frac{s_y}{\bar{y}} \times 100\% = \frac{9.39}{8.42} \times 100\% = 111.52\%$

Thus, the relative dispersion for fund A is higher than that for fund B.

4.2.2 Quartile Coefficient of Variation (QCV):

Just as it is necessary to be able to compare distributions in respect of variation involving the mean and standard deviation, so there is a need for such a comparison involving the median and quartiles. Such a measure is the quartile coefficient of variation (QCV), which measures the quartile deviation as a percentage of the median (in the same way that the coefficient of variation measures the standard deviation as a percentage of the mean). Thus, the quartile coefficient of variation is a relative measure of variation.

Quartile coefficient of variation (QCV):

$$\mathrm{QCV} = \frac{\mathrm{QD}}{\mathrm{Median}} \times 100\%$$

$$\mathrm{QCV} = \frac{Q_3 - Q_1}{2 \times \mathrm{Median}} \times 100\%$$

Example 19:

For the frequency distribution given in Example 3, find the quartile coefficient of variation.

Solution:

The values of $\mathbf{Q_1}$ and $\mathbf{Q_3}$ have already obtained in Example 6.

$$Q_1 = 51.25 \quad \text{and} \quad Q_3 = 69.0625$$

The value of the median can be found as follows:

Median Point = 50/2 = 25

A "less than" cumulative frequency table for this distribution is as follows:

Cumulative Distribution

Class	CF
Less than 30	0
Less than 40	3
Less than 50	11
Less than 60	23
Less than 70	39
Less than 80	45
Less than 90	50

Median Point = $50(\frac{1}{2})$ = 25

Cost	F_i
L.T. 60	23
L.T. m	25
L.T. 70	39

$$\frac{\mathrm{m}-60}{70-60}=\frac{25-23}{39-23} \text{ gives } \frac{\mathrm{m}-60}{10}=\frac{2}{16}$$

$$\text{Median} = 60 + 10\left(\frac{2}{16}\right) = 61.25$$

$$\text{QCV} = \frac{Q_3 - Q_1}{2\times\text{Median}} \times 100\% = \frac{69.0625-51.25}{2\times 61.25} \times 100\%$$

$$= 14.54\ \%$$

Example 20:

For the houses' prices distribution in Example 16, find the quartile coefficient of variation.

Solution:

Class	CF
Less than 10	0
Less than 20	14
Less than 30	32
Less than 40	44
Less than 50	48
Less than 60	50

Q_1 Point = $50(\frac{1}{4})$ = 12.5

Time	F_i
L.T. 10	0
L.T. Q_1	12.5
L.T. 20	14

$$\frac{Q_1 - 10}{20 - 10} = \frac{12.5 - 0}{14 - 0} \quad \text{gives} \quad \frac{Q_1 - 10}{10} = \frac{12.5}{14}$$

$$Q_1 = 10 + 10\left(\frac{12.5}{14}\right) = 18.93$$

Median Point = $50(\frac{1}{2})$ = 25

Time	F_i
L.T. 20	14
L.T. m	25
L.T. 30	32

$$\frac{\mathrm{m}-20}{30-20}=\frac{25-14}{32-14} \text{ gives } \frac{\mathrm{m}-20}{10}=\frac{11}{18}$$

$$\text{Median} = 20 + 10\left(\frac{11}{18}\right) = 26.11$$

Q3 Point = $50\left(\frac{3}{4}\right)$ = 37.5

Time	**F$_i$**
L.T. 30	32
L.T. Q$_3$	37.5
L.T. 40	44

$$\frac{\mathrm{Q}_3-30}{40-30}=\frac{37.5-32}{44-32} \text{ gives } \frac{\mathrm{Q}_3-30}{10}=\frac{5.5}{12}$$

$$\mathrm{Q}_3 = 30 + 10\left(\frac{5.5}{12}\right) = 34.58$$

$$\mathbf{QCV} = \frac{\mathbf{Q_3 - Q_1}}{\mathbf{2\times Median}} \times \mathbf{100\%} = \frac{\mathbf{34.58 - 18.93}}{\mathbf{2\times 26.11}} \times \mathbf{100\%}$$

$$= \mathbf{29.97}\ \%$$

4.3 Standardized Value:

Standardized value (also called standard score or normal deviate) are the same thing as Z- scores. It tells us how far from the mean we are in terms of standard deviations.

So far, we’ve been working with what is called raw data. From this raw data or raw scores, we have been able to create frequency distributions and frequency curves and to examine measures of central tendency and of dispersion. Often, we want to compare people whom we assess with people who are more representative of others in general. Raw scores won’t allow us to do this. With raw scores we can only compare the people within the same group who took the same test. We can't compare their scores to people who were not in the group that we tested. Fortunately, we can transform raw scores to standard scores. When we standardize scores, we can compare scores for different groups of people, and we can compare scores on different tests.

The foundational standard score in measurement is the Z-score. A Z-score is based on the normal, bell-shaped curve and is formed from deviation scores. A deviation score is the difference between any one score and the mean ($\boldsymbol{x_i - \bar{x}}$). If this deviation score is divided by the standard deviation (SD) for that group of scores, we have transformed the raw score into a Z-score. The formula of the Z-Score is given as follows.

Definition:

Standardized Value (Z-Score):

A Z-score is the deviation of a score from the mean expressed in standard deviation units:

$$\mathrm{Z_i} = \frac{\mathrm{x_i} - \bar{\mathrm{x}}}{\mathrm{s}}$$

Standardized Value (Z-Score):

A Z-score is the deviation of a score from the mean expressed in standard deviation units

$$\mathbf{Z_i} = \frac{\mathbf{x_i} - \bar{\mathbf{x}}}{\mathbf{s}}$$

Example 21:

If student's scores in Statistics and Mathematics exams are 75 and 80 respectively, and the exams' mean and standard deviation are:

Statistics Exam: $\bar{\mathrm{x}}_1 = 67 \text{ and } s_1 = 4$

Mathematics Exam: $\bar{\mathrm{x}}_2 = 71 \text{ and } s_2 = 6$

Find the students' standard scores on Statistics and Mathematics exams and interpret their values.

Solution:

$$Z_{\text{Stat}} = \frac{x_{\text{Stat}} - \bar{x}_1}{s_1} = \frac{75 - 67}{4} = 2$$

$\mathbf{Z_{Stat}}$ indicates that the student's score lies 2 standard deviations above the mean.

$$\mathbf{Z_{Math}} = \frac{x_{\text{Math}} - \bar{x}_2}{s_2}$$

$$= \frac{80 - 71}{6} = 1.5$$

$\mathbf{Z_{Math}}$ indicates that the student's score lies 1.5 standard deviations above the mean.

Although the student's Math score is higher than his Statistics score, the Z-scores indicates that he has better relative performance in Statistics than in Mathematics.

Example 22:

You take an Economics test that has a mean of 80 with a standard deviation of 6. What grade did you earn if your Z-score was 1.5?

Solution:

$$Z_{\text{Eco.}} = \frac{x_{\text{Eco.}} - \bar{x}}{s}$$

$$1.5 = \frac{x_{\text{Eco}} - 80}{6} \quad \text{gives} \quad x_{\text{Eco.}} - 80 = 1.5 \times 6$$

$$x_{\text{Eco.}} = 80 + (1.5 \times 6)$$

$$x_{\text{Eco.}} = 89$$

4.4 Coefficient of Skewness:

Skewness was described in the previous chapter and it is shown that the degree of skewness could be measured by the difference between the mean and median or the difference between mean and mode. However, for most practical purposes, it is usual to require a measure of skewness to be unit-free (i.e., a coefficient) and the following expression, known as Pearson's measure of skewness (P) is of this type.

Pearson's Measures of Skewness

$$\mathrm{P}_1 = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\mathrm{P}_2 = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

Note that:

P < 0 shows there is left or negative skewness.

P = 0 signifies no skewness (symmetric distribution).

P > 0 means there is right or positive skewness.

The higher the absolute value of the coefficient of skewness, the more asymmetric the distribution is. Pearson's coefficients of skewness lie between -3 and +3 $(-\mathbf{3} \leq \boldsymbol{P_1} \leq +\mathbf{3})$ and $(\mathbf{3} \leq \boldsymbol{P_2} \leq +\mathbf{3})$. The distribution is almost asymmetric if $|\boldsymbol{P_1}|$ and $|\boldsymbol{P_2}|$ are less than or equal 0.5.

Example 23:

For the frequency distribution given in Example 12, Find Pearson's coefficients of skewness.

Solution:

The following results were obtained:

Mean = 42.6, and standard deviation = 5.147

Mode: The modal class is (40 – 55), therefore

35-	40-	45-50
5	22	18

$L_m = 40$, $f_m = 22$, $f_p = 5$, $f_s = 18$, $w_m = 5$

$$\text{Mode} = l_m + \frac{f_m - f_p}{(f_m - f_p) + (f_m - f_s)} \times w_m$$

$$= 40 + \frac{22-5}{(22-5)+(22-18)} \times 5$$

$$= 40 + \left(\frac{17}{17+4}\right) \times 5 = 44.05$$

$$\text{Median Point} = 50\left(\frac{1}{2}\right) = 25$$

Class	CF
Less than 25	0
Less than 30	2
Less than 35	5
Less than 40	10
Less than 45	32
Less than 50	50

Time	F_i
L.T. 40	10
L.T. m	25
L.T. 45	32

$$\frac{m-40}{45-40} = \frac{25-10}{32-10} \text{ gives } \frac{m-40}{5} = \frac{15}{22}$$

$$\text{Median} = 40 + 5\left(\frac{15}{22}\right) = 43.41$$

Pearson' s coefficients skewness:

$$P_1 = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(42.6 - 43.41)}{5.147} = -0.472$$

$$P_2 = \frac{\text{mean} - \text{mode}}{\text{standard deviation}} = \frac{42.6 - 44.05}{5.147} = -0.282$$

So, the frequency distribution is left (negatively) skewed. The distribution can be considered as symmetric distribution because the absolute value for each is less than 0.5.

Example 24:

Calculate Pearson' coefficients of skewness from the following frequency distribution

Payment of Commission*	10-	12-	14-	16-	18-	20-	22-24	Total
No. of salesmen	7	15	18	20	25	10	5	100

*** Payments of commission are given in hundreds of dollars.**

Solution:

Commission	f	x	d = x – 17	D = d /2	Df	D^2f
10-	7	11	- 6	-3	-21	63
12-	15	13	- 4	-2	-30	60
14-	18	15	- 2	-1	-18	18
16-	20	17	0	0	0	0
18-	25	19	2	1	25	25
20-	10	21	4	2	20	40
22-24	5	23	6	3	15	45
Total	**100**	**-**	**-**	**-**	**-9**	**251**

Mean: $\bar{\mathbf{x}} = \mathbf{a} + \mathbf{b}\left(\frac{\sum \mathbf{D_i f_i}}{\sum \mathbf{f_i}}\right) = \mathbf{17} + \mathbf{2}\left(\frac{-\mathbf{9}}{\mathbf{100}}\right) = \mathbf{16.82}$

Since S_D = $\sqrt{\frac{\sum_{i=1}^{n} D_i^2 f_i}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} D_i f_i}{\sum_{i=1}^{n} f_i}\right)^2}$

$$= \sqrt{\frac{\mathbf{251}}{\mathbf{100}} - \left(\frac{\mathbf{-9}}{\mathbf{100}}\right)^{\mathbf{2}}} = 1.582$$

Standard deviation (S_x) $= \mathrm{b} \times S_{\mathrm{D}} = 2 \times 1.582 = 3.164$

Mode:

The modal class is (18 – 20), therefore

16-	18-	2000-
20	25	10

L_m = 1800 , f_m = 25 , f_s = 10 , f_p = 20 , w_m = 200

Mode = $l_m + \frac{f_m - f_p}{(f_m - f_p) + (f_m - f_s)} \times w_m$

$= 18 + \frac{25-20}{(25-20)+(25-10)} \times 2$

$= 18 + \left(\frac{5}{5+15}\right) \times 2 = 18.5$

Median:

Class	**CF**
Less than 10	0
Less than 12	7
Less than 14	22
Less than 16	40
Less than 18	60
Less than 20	85
Less than 22	95
Less than 24	100

Median Point = $100(\frac{1}{2})$ = 50

Time	F_i
L.T. 16	40
L.T. m	50
L.T. 18	60

$$\frac{\mathrm{m}-16}{18-16} = \frac{50-40}{60-40} \text{ gives } \frac{\mathrm{m}-16}{2} = \frac{10}{20}$$

$$\text{Median} = 16 + 2\left(\frac{10}{20}\right) = 17$$

Pearson' skewness coefficients:

$$\mathrm{P}_1 = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3(16.82-17)}{3.164} = -0.17$$

The distribution is almost symmetric.

$$\mathrm{P}_2 = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{16.82-18.50}{3.164} = -0.53$$

Weak negative skewness.

For a symmetric distribution, the median lies exactly halfway between the other two quartiles. If a distribution is skewed to the right (positively skewed), the median is pulled closer to $\mathbf{Q_1}$ (or pulled closer to $\mathbf{Q_3}$ for negative skewness) and this relationship enables the following coefficient to be derived for measuring skewness.

Quartile Coefficient of Skewness (Bowley's Coefficient): Let us denote this coefficient by QCS. This measure was introduced by Bowley. Its value varies between -1 and 1. That is,

$$\mathbf{-1 \leq QCS \leq 1}$$

Quartile Coefficient of Skewness (Bowley's Coefficient:

$$\mathrm{QCS} = \frac{(\mathrm{Q}_3 - \mathrm{median}) - (\mathrm{median} - \mathrm{Q}_1)}{\mathrm{Q}_3 - \mathrm{Q}_1}$$

Note that:

- **QCS < 0** shows there is left or negative skewness.
- **QCS = 0** signifies no skewness (symmetric distribution).
- **QCS > 0** means there is right or positive skewness.

Example 25:

For the frequency distribution given in Example 3, find the quartile coefficient of skewness.

Solution:

The values of Q_1, median and Q_3 were found to be:

$Q_1 = 51.25$ (Example 6) , Median = 61.25 (Example 19)

$$\mathrm{Q}_3 = 69.0625 \text{ (Example 6)}$$

$$\mathrm{QCS} = \frac{(\mathrm{Q}_3 - \mathrm{median}) - (\mathrm{median} - \mathrm{Q}_1)}{\mathrm{Q}_3 - \mathrm{Q}_1}$$

$$= \frac{(69.0625 - 61.25) - (61.25 - 51.25)}{69.0625 - 51.25} = -0.12$$

So, the frequency distribution is left (negatively) skewed. It can be considered as an almost symmetric distribution.

Example 26:

Calculate the quartile coefficient of skewness from the following frequency distribution:

Monthly Salary$^+$	10-	12-	14-	16-	18-	20-	22-	24-	26-28	Total
No. of Employees	5	14	23	50	52	25	22	7	2	200

+ Monthly salary in hundreds of L.E.

Solution:

To find the values of Q1, Q3, and median, a "less-than" cumulative frequency distribution was found to be as follows.

Class	CF
Less than 10	0
Less than 12	5
Less than 14	19
Less than 16	42
Less than 18	92
Less than 20	144
Less than 22	169
Less than 24	191
Less than 26	198
Less than 28	200

Q_1 Point = $200\left(\frac{1}{2}\right)$ = 100

Salary	F_i
L.T. 16	42
L.T. Q_1	50
L.T. 18	92

$$\frac{Q_1 - 16}{18 - 16} = \frac{50 - 42}{92 - 42} \text{ gives } \frac{Q_1 - 16}{2} = \frac{8}{50}$$

$$Q_1 = 16 + 2\left(\frac{8}{50}\right) = 16.32$$

Median Point = $200\left(\frac{1}{2}\right)$ = 100

Time	F_i
L.T. 18	92
L.T. m	100
L.T. 20	144

$$\frac{\mathrm{m}-18}{2000-18}=\frac{100-92}{144-92} \quad \text{gives} \quad \frac{\mathrm{m}-18}{2}=\frac{8}{52}$$

$$\text{Median} = 18+2\left(\frac{8}{52}\right) = 18.3077$$

Q_3 Point = $200\left(\frac{3}{4}\right) = 150$

Time	F_i
L.T. 2000	144
L.T. Q_3	150
L.T. 2200	169

$$\frac{Q_3-20}{22-20}=\frac{150-144}{169-144} \quad \text{gives} \quad \frac{Q_3-20}{2}=\frac{6}{25}$$

$$Q_3 = 20+2\left(\frac{6}{25}\right) = 20.48$$

Quartile Coefficient of Skewness:

$$\mathrm{QCS}=\frac{(Q_3-\text{median})-(\text{median}-Q_1)}{Q_3-Q_1}$$

$$=\frac{(20.48-18.3077)-(18.3077-16.32)}{20.48-16.32}=0.044$$

Since the quartile coefficient of skewness is too close to zero, the distribution is almost symmetric.

Exercises

1- Given below are 14 statements. Indicate in each case whether the statement is True or False:

(a) The difference between the largest and the smallest observations is called the quartile range.

(b) The dispersion in a series indicates the reliability of the measure of central tendency.

(c) The interquartile range is based on only two values contained in a series.

(d) The square root of the variance gives the standard deviation.

(e) The coefficient of variation is not a relative measure of dispersion.

(f) In measuring dispersion, the standard deviation is more frequently used than the mean deviation.

(g) A major limitation of the range is that it ignores the large number of observations in a series.

(h) Even for an open-ended distribution, it is possible to measure the range.

(i) While calculating variance, every observation in a series is considered.

(j) The coefficient of variation is measured in the same units as the observations in a series.

(k) In a skewed distribution, the mean, median and mode do not have the same value.

(l) In a frequency distribution, if a curve has a longer tail to the right, then it is negatively skewed.

(m) In a positively skewed curve, mean < median < mode.

(n) The median does not always lie between the mean and the mode in a skewed distribution.

2- Multiple Choice Questions:

2.1 Which of the following statements is not correct in respect of the range as a measure of dispersion?

(a) It is difficult to calculate.

(b) Only two points in the data set determine it.

(c) It is affected by extreme values.

(d) There may be considerable change in it from one sample to another.

2.2 If the first and third quartiles are 20.58 and 60.38, respectively, then the quartile deviation is

(a) 39.8 **(b)** 30.3 **(c)** 19.9 **(d)** None of the above

2.3 A series has its mean as 15 and its coefficient of variation as 20, its standard deviation is

(a) 5 **(b)** 10 **(c)** 3 **(d)** 7

2.4 If the first and third quartiles in a series are 15 and 35, then the semi-inter-quartile range is

(a) 30 **(b)** 20 **(c)** 10 **(d)** None of the above

2.5 Which of the following is a measure of relative dispersion?

(a) Variance **(b)** Coefficient of variation

(c) Standard deviation **(d)** All of these

2.6 The square of the variance of a distribution is the

(a) Absolute deviation (**b)** Mean

(c) Standard deviation **(d)** None of these

2.7 Which of the following is not true in respect of mean deviation?

(a) It is simple to understand

(b) It considers each and every item in a series.

(c) It is capable of further algebraic treatment.

(d) The extreme items have less effect on its magnitude.

2.8 Which one is the formula for relative skewness?

(a) Mean = Mode

(b) $(Q_3 - Q_2) - (Q_2 - Q_1)$

(c) $\frac{(Q_3 - \text{median}) - (\text{median} - Q_1)}{Q_3 - Q_1}$

(d) None of the above

2.9 Which of the following relationship is valid in a symmetrical distribution?

(a) $(\text{Median} - Q_1) < (Q_3 - \text{Median})$

(b) $(\text{Median} - Q_1) > (Q_3 - \text{Median})$

(c) $(\text{Median} - Q_1) = (Q_3 - \text{Median})$

(d) None of these

3- The following table gives the heights of students in a class. Find out the quartile deviation.

Height in Inches	50-	53-	56-	59-	62-	65-68
No. of Students	2	7	24	27	13	3

4- Find the mean deviation of each set of numbers:

(a) 12 , 6 , 7 , 3 , 15 , 10 , 18 , 5

(b) 9, 3, 8, 8, 9, 8, 9 and 18.

5- Find the range of weights of 100 students from the data given below:

Weight (Kg)	60-	63-	66-	69-	72-75
No. of Students	5	18	42	27	8

6- The mean of five observations is 4.4 and the variance is 8.24. If three of the five observations are 1, 2 and 6, find the other two.

7- The following is a record of the number of bricks laid each day by two workers A and B:

A: 700 , 675 , 725 , 675 , 800 , 650 , 675 , 625 , 700 , 650
B: 600 , 625 , 675 , 575 , 650 , 625 , 600 , 625 , 550 , 700

Calculate the coefficient of variation in each case and discuss the relative consistency of the two masons. If the figures for A were in every case 20 more and those of B in every case 10 more than the figures given above, how would the answer be affected?

8- The following table gives the marks of 59 students in economics. Calculate the quartile deviation and the quartile coefficient of variation.

Mark	0-	10-	20-	30-	40-	50-	60-70
No. of Students	4	8	11	15	12	6	3

9- For the following distribution, calculate the variance, the standard deviation and the coefficient of variation.

Weight	130-	140-	150-	160-	170-	180-190
No. of Plots	10	20	30	20	10	10

10- In two factories A and B engaged in similar type of industry, the average weekly wages and standard deviations are as given below:

	Factory A	Factory B
Average Weekly Wages	460	490
Standard Deviation of Weekly Wages	50	40

(a) Which, factory A or factory B, pays larger amount as weekly wages?

(b) Which factory shows greater variability in the distribution of weekly wages?

11- Given below are the daily wages paid to workers in two factories X and Y.

Daily Wages	Number of Workers	
	Factory X	Factory Y
20-	15	25
30-	30	40
40-	44	60
50-	60	35
60-	60	20
70-	14	15
80-90	7	5

Using arithmetic mean and standard deviation, answer the following questions:

(a) Which factory pays higher average wages? By how much?

(b) In which factory are wages more variable?

12- Calculate standard deviation for average life of a particular band of T.V. sets.

Life in Years	0-	2-	4-	6-	8-	10-12
No. of Sets	5	16	13	7	5	4

13- Calculate standard deviation for the following distribution of net profits of 100 firms in a certain industry.

Net Profit (%)	0-	5-	10-	15-	20-25
No. of Firms	8	42	36	10	4

14- The distribution of wages of workers in two factories A and B is given below. Determine the factory in which total wages paid to all the workers is more, and the factory in which the wages are more variable.

Daily Wages	Number of Workers	
	Factory X	Factory Y
50-	2	6
100-	9	11
150-	29	18
200-	54	32
250-	11	27
300-350	5	11

15- Calculate Karl Pearson's coefficients of skewness from the following data:

Weekly Sales	10-	12-	14-	16-	18-	20-	22-	24-26
No. of Companies	12	18	35	42	50	45	30	8

Comment on the value obtained.

16- From the following data of age of employees, calculate the quartile coefficient of skewness and comment on the result:

Age (Years)	Number of Employees
Less than 25	8
Less than 30	20
Less than 35	40
Less than 40	65
Less than 45	80
Less than 50	92
Less than 55	100

Chapter (5)
Correlation Analysis

Correlation is a technique used to measure the strength of the relationship between two variables. This chapter describes the general nature and purpose of correlation and gives techniques for measuring correlation. Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other. However, correlation does not imply causation. There may be an unknown factor that influences both variables similarly.

5.1 Scatterplot (Scatter Diagram):

A scatter diagram is a useful technique for visually examining the form of relationship, without calculating any numerical value. A scatterplot shows the relationship between two quantitative variables measured for the same individual. The values of one variable appear on the horizontal axis, and the other variable appear on the vertical axis. Many research projects are correlation studies because they investigate the relationships that may exist between variables. Prior to investigating the relationship between two quantitative variables, it is always helpful to create a graphical representation that includes both of the variables. Such a graphical representation is called a "scatterplot" or a "scatter diagram". For example, the following table and figure represent students' GPA and students' achievement motivation.

Table (5.1)
GPA and Achievement Motivation For A Group of 14 Students

ID	GPA	Motivation
1	2.0	50
2	2.0	48
3	2.0	100
4	2.0	12
5	2.3	34
6	2.6	30
7	2.6	78
8	3.0	87
9	3.1	84
10	3.2	75
11	3.6	83
12	3.8	90
13	3.8	90
14	4.0	98

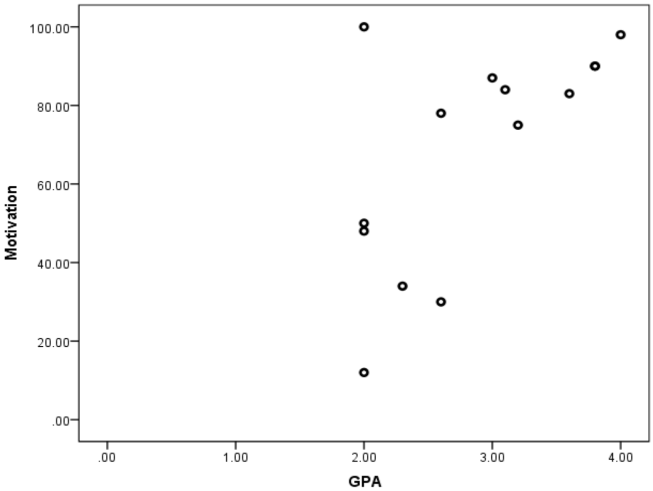


Figure 5.1
Scatterplot For Table 5.1

In this example, the relationship between students' achievement motivation and their GPA is being investigated (GPA stands for "Grade Point Average". It is a number that indicates how well or how high you scored). Table 5.1 includes a small group of individuals for whom GPA and scores on a motivation scale have been recorded. GPAs can range from 0 to 4 and motivation scores in this example range from 0 to 100. Individuals in this table were ordered based on their GPA. Simply looking at the table shows that, in general, as GPA increases, motivation scores also increase. However, with a real set of data, which may have hundreds or even thousands of individuals, a pattern cannot be detected by simply looking at the numbers. Therefore, a very useful strategy is to represent the two variables graphically to illustrate the relationship between them. The image on the right is an example of a scatterplot and displays the data from the table on the left. GPA scores are displayed on the horizontal axis and motivational scores are displayed on the vertical axis. Each dot on the scatterplot represents one individual from the data set. The location of each point on the graph depends on both the GPA and motivation scores. Individuals with higher GPAs are located further to the right and individuals with higher motivation scores are located higher up on the graph. The purpose of a scatterplot is to provide a general illustration of the relationship between the two variables. In this example, in general, as GPA increases so does an individual's motivation score.

Interpreting Scatterplots:

As in any graph of data, look for the overall pattern and for striking departures from that pattern. The overall pattern of a scatterplot can be described by direction, form, and strength of the relationship. An important kind of departure is an outlier, an individual value that falls outside the overall pattern of the relationship.

One important component to a scatterplot is the direction of the relationship between two variables.

- Two variables have a positive association when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.
- Two variables have a negative association when above-average values of one tend to accompany below-average values of the other.

Figure 5.2 compares students' achievement motivation and their GPA. These two variables have a positive association because as GPA increases, so does motivation.

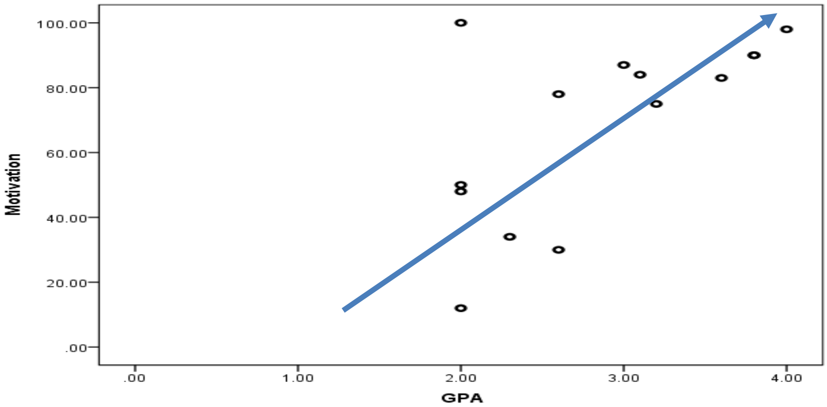


Figure 5.2
Achievement Motivation and GPA for Students

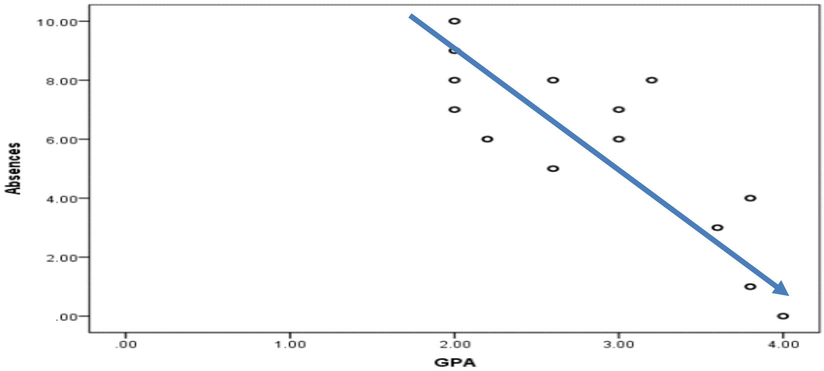


Figure 5.3
GPA and Number of Absences for Students

Figure 5.3 compares students' GPA and their number of absences. These two variables have a negative association because, in general, as a student's number of absences decreases, their GPA increases

Another important component to a scatterplot is the form (type) of the relationship between the two variables.

Linear Relationship:

Figure 5.2 illustrates a linear relationship. This means that the points on the scatterplot closely resemble a straight line.
A relationship is linear if one variable increases by approximately the same rate as the other variables change by one unit.

Curvilinear Relationship:

Figure 5.4 illustrates a relationship that has the form of a curve, rather than a straight line. This is due to the fact that one variable does not increase at a constant rate and may even start decreasing after a certain point.

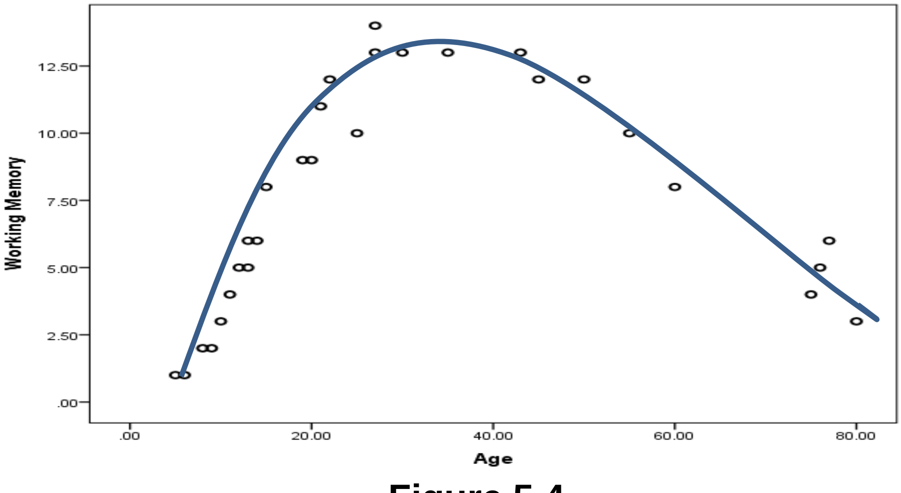


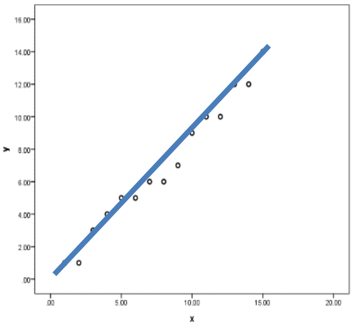
Figure 5.4
The Relationship Between
Age and Working Memory

This example describes a curvilinear relationship between the variable "age" and the variable "working memory". In this example, working memory increases through childhood, remains steady in adulthood, and begins decreasing around age 50.

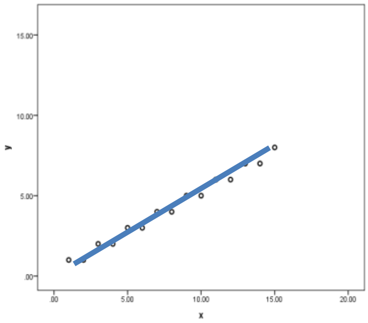
Strength of the Relationship:

Another important component to a scatterplot is the strength of the relationship between the two variables. The slope provides information on the strength of the relationship.

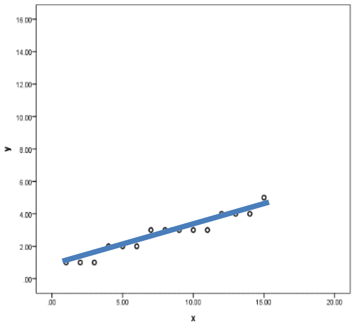
Strong Relationship



Moderate Relationship



Weak Relationship



The strongest relationship occurs when the slope is 1. This means that when one variable increases by one, the other variable increases by the same amount. This line is at a 45-degree angle. The strength of the relationship between two variables is a crucial piece of information. Relying on the interpretation of a scatterplot is too subjective. More precise evidence is needed, and this evidence is obtained by computing a coefficient that measures the strength of the relationship under investigation.

5.2 Pearson’s Correlation Coefficient (r):

A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables. A correlation coefficient measures the strength of that relationship. Calculating a Pearson’s correlation coefficient requires the assumption that the relationship between the two variables is linear (this point will be covered later when dealing with regression).

Pearson's Correlation Coefficient Formula:

$$\mathrm{r} = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

Note that:

- **r** is always a number between -1 and 1.
- **r > 0** indicates a positive (direct) relationship.
- **r < 0** indicates a negative (inverse) relationship.
- The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.
- The extreme values r = -1 and r = 1 occur only in the case of perfect linear relationship.

There is a rule of thumb for interpreting the strength of a relationship based on its r value (use the absolute value of r):

Absolute Value of r	**Strength of Linear Relationship**
0	No relationship
$0 < r < 0.2$	Very weak
$0.2 \le r < 0.4$	Weak
$0.4 \le r < 0.6$	Moderate
$0.6 \le r < 0.8$	Strong
$0.8 \le r < 1$	Very Strong
1	Perfect relationship

Characteristics of Pearson's Correlation Coefficient:

- The order of variables in Pearson's correlation coefficient is not important, which means it does not make difference which one is x and which one is y.

- It provides evidence of association, not causation.
- It is affected by outliers.
- It doesn’t change when the units of the measure of x, y, or both are changed (i.e. not affected by mathematical operations).

Coefficient of Determination $(\mathbf{r^2})$:

Definition:

The coefficient of determination gives the proportion of the variation in the y-values that is explained by the variation in the x-values.

Example 1:

The following table relates the weekly maintenance cost ($) to the age (in months) of ten machines of similar type in a manufacturing company.

Machine	1	2	3	4	5	6	7	8	9	10
Age (x)	5	10	15	20	30	30	30	50	50	60
***Cost* (y)**	*190*	*240*	*250*	*300*	*310*	*335*	*300*	*300*	*350*	*395*

Calculate the Pearson’s correlation coefficient and the coefficient of determination.

Solution:

The table in the next page present the calculations required for determining the values of the correlation coefficient and then the coefficient of determination.

From this table, the following results were obtained:

n = 10 , $\sum_{i=1}^{n} x_i = 300$, $\sum_{i=1}^{n} y_i = 2970$,

$\sum_{i=1}^{n} x_i y_i = 97650$, $\sum_{i=1}^{n} x_i^2 = 12050$, $\sum_{i=1}^{n} y_i^2 = 913050$

Machine	$\mathbf{x}$	$\mathbf{y}$	$\mathbf{xy}$	$\mathbf{x^2}$	$\mathbf{y^2}$
1	5	190	950	25	36100
2	10	240	2400	100	57600
3	15	250	3750	225	62500
4	20	300	6000	400	90000
5	30	310	9300	900	96100
6	30	335	10050	900	112225
7	30	300	9000	900	90000
8	50	300	15000	2500	90000
9	50	350	17500	2500	122500
10	60	395	23700	3600	156025
Total	300	2970	97650	12050	913050

$$r = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

$$= \frac{\left(\frac{97650}{10}\right) - \left(\frac{300}{10}\right)\left(\frac{2970}{10}\right)}{\sqrt{\left[\frac{12050}{10} - \left(\frac{300}{10}\right)^2\right] \times \left[\frac{913050}{10} - \left(\frac{2970}{10}\right)^2\right]}}$$

$$= \frac{855}{\sqrt{305 \times 3096}} = 0.88$$

There is a very strong positive relationship. This means as the age of the machine increases the cost of its maintenance increases, and vice versa.

Coefficient of determination: $\mathrm{r}^2 = (0.88)^2 = 0.7744$

This means that 77.44% of the variation in maintenance cost is explained by the age of machines, while 22.56% is due to other factors. This is according to the linear relationship between the cost of maintenance and the age of the machine.

Example 2:

The following data, obtained from claims drawn on life assurance policies, relates age at official retirement to age at death.

ID	1	2	3	4	5	6	7	8	9
Age of Retirement	57	62	60	57	65	60	58	62	56
Age At Death	71	70	66	70	69	67	69	63	70

Calculate the Pearson's correlation coefficient between age at retirement and age at death.

Solution:

ID	x	y	xy	x^2	y^2
1	57	71	4047	3249	5041
2	62	70	4340	3844	4900
3	60	66	3960	3600	4356
4	57	70	3990	3249	4900
5	65	69	4485	4225	4761
6	60	67	4020	3600	4489
7	58	69	4002	3364	4761
8	62	63	3906	3844	3969
9	56	70	3920	3136	4900
Total	537	615	36670	32111	42077

n = 9 , $\sum_{i=1}^{n} x_i = 537$, $\sum_{i=1}^{n} y_i = 615$, $\sum_{i=1}^{n} x_i y_i = 36670$, $\sum_{i=1}^{n} x_i^2 = 32111$, $\sum_{i=1}^{n} y_i^2 = 42077$

$$r = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

$$= \frac{\left(\frac{36670}{9}\right) - \left(\frac{537}{9}\right)\left(\frac{615}{9}\right)}{\sqrt{\left[\frac{32111}{9} - \left(\frac{537}{9}\right)^2\right] \times \left[\frac{42077}{9} - \left(\frac{615}{9}\right)^2\right]}}$$

$$= \frac{-2.78}{\sqrt{7.78 \times 5.78}} = -0.415$$

There is a moderate negative relationship. This indicates that as the age of retirement increases, the age of death decreases and vice versa.

Example 3:

A sample of eight employees is taken from the production department of a light engineering factory. The data which follow relate to the number of weeks experience in the wiring of components, and the number of components which were rejected as unsatisfactory last week.

Employee	A	B	C	D	E	F	G	H
Weeks of Experience	4	5	7	9	10	11	12	14
Number of Rejects	21	22	15	18	14	14	11	13

Calculate the Pearson’s correlation coefficient and coefficient of determination for these data and interpret their values.

Solution:

Employee	x	y	xy	x^2	y^2
A	4	21	84	16	441
B	5	22	110	25	484
C	7	15	105	49	225
D	9	18	162	81	324
E	10	14	140	100	196
F	11	14	154	121	196
G	12	11	132	144	121
H	14	13	182	196	169
Total	72	128	1069	732	2156

n = 8 , $\sum_{i=1}^{n} x_i = 72$, $\sum_{i=1}^{n} y_i = 128$,

$\sum_{i=1}^{n} x_i y_i = 1069$, $\sum_{i=1}^{n} x_i^2 = 732$, $\sum_{i=1}^{n} y_i^2 = 2156$

$$r = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

$$= \frac{\left(\frac{1069}{8}\right) - \left(\frac{72}{8}\right)\left(\frac{128}{8}\right)}{\sqrt{\left[\frac{732}{8} - \left(\frac{72}{8}\right)^2\right] \times \left[\frac{2156}{8} - \left(\frac{128}{8}\right)^2\right]}}$$

$$= \frac{-10.375}{\sqrt{10.5 \times 13.5}} = -0.87$$

There is a very strong negative relationship. That is, as the duration of experience increases the number of rejects decreases and vice versa.

Coefficient of determination: $r^2 = (-0.87)^2 = 0.7569$

This means that 75.69% of the variation in number of rejects is explained by variation in duration of experience, and 24.31% is due to factors other than duration of experience. This comes as a result for the linear relationship between the two variables.

5.3 Spearman's Rank Correlation Coefficient:

An alternative method of measuring correlation, based on the ranks of the sizes of item values, is available and known as rank correlation.

Rank correlation is used in the following circumstances:

- As an approximation to Pearson's correlation coefficient. This is particularly appropriate if the values of numeric bivariate data are difficult to obtain physically or involve great expense and yet can be ranked in size order.
- If one or both of the variables involved is non-numeric, the Pearson's correlation coefficient cannot be calculated. However, as long as the non-numeric values can be ranked in some natural way, rank correlation can be used.

The procedure for obtaining Spearman's rank correlation coefficient is given as follows:

Step 1: Rank the x values (to give $\mathrm{r_x}$ values).

Step 2: Rank the y values (to give $\mathrm{r_y}$ values).

Step 3: For each pair of ranks, calculate $\mathrm{d}^2 = (\mathrm{r_x} - \mathrm{r_y})^2$.

Step 4: Calculate $\sum \mathrm{d}^2$.

Step 5: The value of the rank correlation coefficient can then be found using the following formula:

Rank Correlation Coefficient ($\mathbf{r_s}$):

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where: d = $r_x - r_y$,

n is the number of bivariate pairs.

Notes On the Rank Correlation Coefficient:

- Ranks are usually allocated in ascending order; rank 1 to the smallest item, rank 2 to the next largest and so on, although it is perfectly feasible to allocate in descending order. However, whichever method is selected must be used on both variables.
- If one or more groups of data items have the same value (known as tied values), the ranks that would have been allocated separately must be averaged and this average rank given to each item with this equal value.

Characteristics of Spearman's Rank Correlation Coefficient:

- It is not affected by outliers. That is, if there is a big difference between r and r_s, it is due to outliers that inflates or shrinks the value of r.
- It doesn't change when the units of the measure of x, y, or both are changed (i.e. not affected by mathematical operations).
- If r = 1, the value of r_s must equal 1. But if r_s = 1, it is not necessary for the value of r to be 1.

Example 4:

The following data relate to the number of vehicles owned (per 100 population) and road deaths (per 100,000 population) for the populations of 12 countries.

Vehicles	30	31	32	30	46	30	30	19	35	40	46	57
Road Deaths	30	14	30	23	32	26	26	20	21	23	30	35

Calculate Spearman's correlation coefficient.

Solution:

x	y	r_x	r_y	d	d^2
30	30	3.5	9	-5.5	30.25
31	14	6	1	5	25
32	30	7	9	-2	4
30	23	3.5	4.5	-1	1
46	32	10.5	11	-0.5	0.25
30	26	3.5	6.5	-3	9
19	20	1	2	-1	1
35	21	8	3	5	25
40	23	9	4.5	4.5	20.25
46	30	10.5	9	1.5	2.25
57	35	12	12	0	0
30	26	3.5	6.5	-3	9
Total					127

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6\times127}{12(12^2-1)} = 0.56$$

There is a moderate positive relationship. That is, as the number of vehicles increases the road deaths increases.

Example 5:

The following data give information of the ages (in years) and the number of breakdowns during the past month for 5 machines in a small company.

Machine No.	1	2	3	4	5
Age (year)	7	2	4	8	9
No. of Breakdowns	5	1	2	5	7

Calculate Spearman' correlation Coefficient.

Solution:

$\mathbf{x_i}$	$\mathbf{y_i}$	$\mathbf{r_x}$	$\mathbf{r_y}$	$\mathbf{d_i}$	$\mathbf{d_i^2}$
7	5	3	3.5	-0.5	0.25
2	1	1	1	0	0
4	2	2	2	0	0
8	5	4	3.5	0.5	0.25
9	7	5	5	0	0
Total					0.5

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 0.5}{5(5^2-1)}$$

$$= 0.975$$

There is a very strong positive relationship. That is, as the age of vehicles increases the number of break downs increases.

Example 6:

The following table represents the daily production (units) and the number of workers assigned for each of 8 days.

Day	1	2	3	4	5	6	7	8
No. of Workers	5	7	6	8	9	7	8	6
Production	12	14	13	17	19	16	16	13

Calculate Spearman's correlation coefficient, Pearson's correlation coefficient. Comment on your results.

Solution:

From the table given in the next page, the following results were found:

n = 8 , $\sum_{i=1}^{n} x_i = 56$, $\sum_{i=1}^{n} y_i = 120$, $\sum_{i=1}^{n} x_i y_i = 861$,
$\sum_{i=1}^{n} x_i^2 = 404$, $\sum_{i=1}^{n} y_i^2 = 1840$, $\sum_{i=1}^{n} d_i^2 = 2.5$

Day	x	y	xy	x^2	y^2	r_x	r_y	d_i	d_i^2
1	5	12	60	25	144	1	1	0	0
2	7	14	98	49	196	4.5	4	0.5	0.25
3	6	13	78	36	169	2.5	2.5	0	0
4	8	17	136	64	289	6.5	7	-0.5	0.25
5	9	19	171	81	361	8	8	0	0
6	7	16	112	49	256	4.5	5.5	-1	1
7	8	16	128	64	256	6.5	5.5	1	1
8	6	13	78	36	169	2.5	2.5	0	0
Total	56	120	861	404	1840	-	-	0	2.5

Pearson's Correlation Coefficient:

$$r = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

$$= \frac{\left(\frac{861}{8}\right) - \left(\frac{56}{8}\right)\left(\frac{120}{8}\right)}{\sqrt{\left[\frac{404}{8} - \left(\frac{56}{8}\right)^2\right] \times \left[\frac{1840}{8} - \left(\frac{120}{8}\right)^2\right]}}$$

$$= \frac{2.625}{\sqrt{1.5 \times 5}} = 0.96$$

Spearman's rank correlation coefficient:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 2.5}{8(8^2-1)}$$

$$= 0.97$$

The value of $\mathrm{r_s}$ is very close to the value of r which indicates that data of x and y do not contain outliers. There is a very strong positive relationship. That is, as the number of workers increases, the daily production increases.

5.4 Yule’s Coefficient of Association:

Measures of associations are used to assess the strength of relationship between qualitative variables. For 2 x 2 tables, one of the best and simplest measures is Yule’s Q. If we simply label the counts in the cells of the 2 X 2 table as:

a	b
c	d

Then:

Yule’s Coefficient of Association:

$$\mathrm{Q} = \frac{\mathrm{ad} - \mathrm{bc}}{\mathrm{ad} + \mathrm{bc}}$$

This coefficient varies from -1 (if either a or d is zero) to +1 (if either b or c is zero). The absolute value is considered because the sign is meaningless with regard to the nature of the relationship between the two attributes.

The categories of the strength of relationship used for r and r_s given before can also be used for Yule's coefficient and Cramer's coefficient which is covered next.

Example 7:

An insurance company is interested in determining whether there is a relationship between automobile accident frequency and cigarette smoking. It randomly sampled 36 policyholders and came up with the following data:

Cigarette Smoking	Number of Accidents		Total
	0	1	
Smokers	8	6	14
Nonsmokers	12	10	22
Total	20	16	36

Does the sample provide sufficient information to conclude that there is a relationship between automobile accident frequency and cigarette smoking?

Solution:

$$\mathrm{Q} = \frac{\mathrm{ad} - \mathrm{bc}}{\mathrm{ad} + \mathrm{bc}} = \frac{(8 \times 10) - (6 \times 12)}{(8 \times 10) + (6 \times 12)} = 0.053$$

There is a very weak relationship between smoking and traffic accidents. The relationship between traffic accidents and whether the driver is a smoker or a nonsmoker is almost non-existent. The percentages of drivers with accidents among smokers and nonsmokers are 42.86% and 45.45%, respectively. The two percentages are very close. This means that the driver with accidents does not depend on whether or not he is a smoker.

Example 8:

The owner of a store is interested to find out if there is any relationship between the time of a purchase is made and the amount of money spent on that purchase. The following table presents the number of purchases for size and time of each purchase.

Size of Purchase	Time of Purchase		Total
	Evening	Night	
$10-20	10	5	15
Over $20	10	25	35
Total	20	30	50

Find the value of Yule’s coefficient of association and interpret your result.

Solution:

$$\mathrm{Q} = \frac{\mathrm{ad} - \mathrm{bc}}{\mathrm{ad} + \mathrm{bc}} = \frac{(10 \times 25) - (5 \times 10)}{(10 \times 25) + (5 \times 10)} = 0.67$$

There a strong relationship between the time of purchase and purchase size. From the table, 50% of those who buy in the evening spent more than $20 in the purchase, while the percentage reaches 83.33% at the time of purchase at night. This means that buyers spend more on their purchases during the night than they spend in the evening.

Example 9:

Two samples, one of 200 students from urban secondary schools and another of 250 students from rural secondary schools, were taken. These students were asked if they have ever smoked. The following table lists the summary of the results.

Type of Car	Car Repair		Total
	Requiring Repair	Not Requiring Repair	
A	11	89	100
B	13	57	70
Total	24	146	170

Find the value of Yule’s coefficient of association and interpret your result.

Solution:

$$\mathrm{Q} = \frac{\mathrm{ad} - \mathrm{bc}}{\mathrm{ad} + \mathrm{bc}} = \frac{(11 \times 57) - (89 \times 13)}{(11 \times 57) + (89 \times 13)} = -0.297$$

There is a weak relationship between the type of car and need for maintenance. The car's need for maintenance depends poorly on whether the car is of type A or B. It is clear from the table that type

B is in need of maintenance at a rate greater that between type A (11% and 18.6%, respectively).

5.5 Contingency Coefficient (Cramer's Coefficient):

Cramer's coefficient of association (denoted by V) can be used to assess the strength of relationship between qualitative variables (attributes) for tables in which either R, C or both > 2, where R is the number of rows and C is the number of columns.

Cramer's Coefficient

$$V = \sqrt{\frac{\chi^2}{N(s-1)}}$$

Where N is the table total.

S either R or C whichever is smaller.

χ^2 is to be defined below (It is read as Chi-Square).

χ^2 Formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Computational Formula for χ^2 :

$$\chi^2 = \sum \frac{O^2}{E} - N$$

Where

E: Expected value

$$E = \frac{\text{Row Marginal} \times \text{Column Marginal}}{N}$$

O: Observed value

N: Table total

Example 10:

A sample of 200 units produced by a machine were classified as good or defective and by the shift on which they were produced. The results are reported in the following table.

Quality	Shift			Total
	First	Second	Third	
Good	76	64	40	180
Defective	4	6	10	20
Total	80	70	50	200

Is there evidence of a relationship between the quality of the units and the shift in which they were produced. If yes, to what extent?

Solution:

Expected values: $E = \frac{\text{Row Marginal} \times \text{Column Marginal}}{\text{N}}$

$$E_{11} = \frac{180 \times 80}{200} = 72 \quad , \quad E_{12} = \frac{180 \times 70}{200} = 63$$

$$E_{13} = \frac{180 \times 50}{200} = 45 \quad , \quad E_{21} = \frac{20 \times 80}{200} = 8$$

$$E_{22} = \frac{20 \times 70}{200} = 7 \quad , \quad E_{23} = \frac{20 \times 50}{200} = 5$$

Chi-Square: $\boldsymbol{\chi^2 = \sum \frac{O^2}{E} - N}$

$= (76)^2 / 72 + (64)^2 / 63 + (40)^2 / 45 + (4)^2 / 8$
$+ (6)^2 / 7 + (10)^2 / 5 – 200 = 7.937$

Cramer’s Coefficient:

$$V = \sqrt{\frac{\chi^2}{N(s-1)}} = \sqrt{\frac{7.937}{200(2-1)}} = 0.199$$

There is a poor relationship between the quality of the product and the shift of production. This means that the quality of the product does not depend, to a large extent, on the timing of the shift in which it was produced.

Note: Sum of Expected Frequencies and Sum of Observed Frequencies are equal for any row or column.

Example 11:

A study regarding the relationship between age and the amount of pressure sales personnel feel in relation to their jobs revealed the following sample information.

Age (years)	Degree of Job Pressure			Total
	Low	Medium	High	
Less than 25	70	20	10	100
25 - 40	70	50	80	200
40 - 60	10	30	160	200
Total	150	100	250	500

Does the sample data provide evidence to conclude that the degree of job pressure depends upon age?

Solution:

Expected values: $E = \dfrac{\text{Row Marginal} \times \text{Column Marginal}}{\text{N}}$

$E_{11} = \dfrac{100 \times 150}{500} = 30$, $E_{12} = \dfrac{100 \times 100}{500} = 20$,

$E_{13} = \dfrac{100 \times 250}{500} = 50$, $E_{21} = \dfrac{200 \times 150}{500} = 60$

$E_{22} = \dfrac{200 \times 100}{500} = 40$. $E_{23} = \dfrac{200 \times 250}{500} = 100$

$E_{31} = \dfrac{200 \times 150}{500} = 60$, $E_{32} = \dfrac{200 \times 100}{500} = 40$

$E_{33} = \dfrac{200 \times 250}{500} = 100$

Chi-square: $\chi^2 = \sum \frac{O^2}{E} - N$

$$\chi^2 = \left(\frac{70^2}{30} + \frac{20^2}{20} + \frac{10^2}{50} + \frac{70^2}{60} + \frac{50^2}{40} + \frac{80^2}{100} + \frac{10^2}{60} + \frac{30^2}{40} + \frac{160^2}{100}\right) - 200 = 173.67$$

Cramer’s Coefficient:

$$\mathrm{V} = \sqrt{\frac{\chi^2}{\mathrm{N}(s-1)}} = \sqrt{\frac{173.67}{500(3-1)}} = 0.417$$

The two attributes have a moderate relationship. The degree of job pressure depends moderately on age. How much job pressure sales personnel are exposed to depends to some extent on his age. From the table we can conclude that the job pressure on the sales personnel increases as his age increases (the proportion of sales personnels who suffer from a high degree of job pressure 1s 0.1, 0.4 and 0.8 in age groups "less than 25", "25 - 40" and "40 - 60", respectively.

Example 12:

Data on the social class and the number of children in a family were obtained as a part of national survey. The results from a sample of 200 families follow.

Number of Children	Social Class			Total
	Lower	Middle	Upper	
1 or 2	12	24	84	120
More than 2	64	12	4	80
Total	76	36	88	200

To what extent you can say that the number of children in a family depends upon the social class of this family?

Solution:

Expected values: $\mathrm{E} = \frac{\text{Row Marginal} \times \text{Column Marginal}}{\mathrm{N}}$

$$\mathrm{E}_{11}=\frac{120\times 76}{200}=45.6 \quad , \quad \mathrm{E}_{12}=\frac{120\times 36}{200}=21.6$$

$$\mathrm{E}_{13}=\frac{120\times 88}{200}=52.8 \quad , \quad \mathrm{E}_{21}=\frac{80\ \times\ 76}{200}=30.4$$

$$\mathrm{E}_{22}=\frac{80\ \times\ 36}{200}=14.4 \quad , \quad \mathrm{E}_{23}=\frac{80\ \times\ 88}{200}=35.2$$

Chi-square: $\chi^2=\sum\frac{\mathrm{O}^2}{\mathrm{E}}-\mathrm{N}$

$$\chi^2=\left(\frac{12^2}{45.6}+\frac{24^2}{21.6}+\frac{84^2}{52.8}+\frac{64^2}{30.4}+\frac{12^2}{14.4}+\frac{4^2}{35.2}\right)-200=108.65$$

Cramer’s Coefficient:

$$\mathrm{V}=\sqrt{\frac{\chi^2}{\mathrm{N}(\mathrm{s}-1)}}=\sqrt{\frac{108.65}{200(2-1)}}=0.737$$

There is a strong relationship between the social class and number of children. The lower the social level of family the more children there are. From the table, we can conclude that 80% of the families who have two or more children have a low social level, while this percentage is 15% and 5% in families of medium and high social level, respectively.

Exercises

Multiple Choice Questions (1- 6)

1- The unit of correlation coefficient between height in feet and weight in kgs is

(i) kg/feet

(ii) percentage

(iii) non-existent

2- The range of simple correlation coefficient is

(i) 0 to infinity

(ii) minus one to plus one

(iii) minus infinity to infinity

3- If r is positive the relation between X and Y is of the type

(i) When Y increases X increases

(ii) When Y decreases X increases

(iii) When Y increases X does not change

4- If r = 0 the variable X and Y are

(i) linearly related

(ii) not linearly related

(iii) independent

5- Of the following three measures which one can measure any type of relationship

(i) Karl Pearson’s coefficient of correlation

(ii) Spearman’s rank correlation

(iii) Scatter diagram

(iv) None of these

6- If precisely measured data are available, the simple correlation coefficient is

(i) more accurate than rank correlation coefficient.

(ii) less accurate than rank correlation coefficient.

(iii) as accurate as the rank correlation coefficient.

7- Can r lie outside the –1 and 1 range depending on the type of data?

8- Does correlation imply causation?

9- When is rank correlation more precise than simple correlation coefficient?

10- Does zero correlation mean independence?

11- Can simple correlation coefficient measure any type of relationship?

12- Interpret the values of r as 1, –1 and 0.

13- Why does rank correlation coefficient differ from Pearson's correlation coefficient?

14- Calculate Pearson's correlation coefficient between the heights of fathers in inches (X) and the heights of their sons (Y).

x	65	66	57	67	68	69	70	72
y	67	56	65	68	72	72	69	71

15- Calculate the correlation coefficient between X and Y and comment on their relationship:

x	-3	-2	-1	1	2	3
y	9	4	1	1	4	9

16- Compute Pearson's correlation coefficient between advertisement cost and sales as given below. Explain your result.

Advertisement	39	65	62	90	82	75	25	98	36	78
Sales	47	53	58	86	62	68	60	91	51	84

17- The table below shows the number of absences, x, in a Statistics course and the final exam grade, y, for 7 students. Find the correlation coefficient and interpret your result.

x	1	0	2	6	4	3	3
y	95	90	90	55	70	80	85

18- Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank, we would like to know that to what extent the knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	47	53	58	86	62	68	60	91

19- Calculate rank correlation coefficient from the following data:

x	78	89	97	69	59	79	68
y	125	137	156	112	107	136	124

20- Calculate Spearman's rank correlation coefficient for the following data:

x	20	38	30	40	50	55
y	17	45	30	35	40	45

21- Calculate rank correlation coefficient for the following data:

x	10	20	30	30	40	45	50
y	15	20	25	30	40	40	50

22- Calculate rank correlation coefficient from the following data:

x	70	70	80	80	90	90	100
y	80	90	90	80	70	60	50

23- Of a group of patients who complained that they did not sleep well, some were given sleeping pills while others were given sugar pills (although they all thought they were getting sleeping pills). They were later asked whether the pills help them or not.

Pills	**Mode of Sleep**		**Total**
	Slept Well	**Did Not Sleep Well**	
Sleeping Pills	44	10	54
Sugar Pills	81	35	116
Total	125	45	170

Does the sample provide sufficient information to conclude that there is a relationship between pills type frequency and mode of sleep?

24- 1,000 students at a college level were graded according to their I.Q. score and the economic conditions of their homes.

Economic Condition	**IQ**		**Total**
	High	**Low**	
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Does the sample provide sufficient evidence to conclude that there is a relationship between economic conditions and IQ level. Explain.

25- Three samples are taken comprising 120 doctors, 150 advocates and 130 university teachers. Each person chosen is asked to select one of the three categories that best represents his feeling toward a certain national policy. The three categories are in favor of policy (F), against the policy (A), and indifferent toward the policy (I).

Occupation	**Reaction**			**Total**
	F	**A**	**I**	
Doctors	80	30	10	120
Advocates	70	40	40	150
Univ. Teachers	50	50	30	130
Total	200	120	80	400

Can you conclude that there is a relationship between occupation and respondents’ reaction toward the policy? Explain

Chapter (6)
Regression Analysis

Regression analysis concerns the study of relationships between variables with the object of identifying, estimating and validating the relationship. The estimated relationship can then be used to predict one variable from the value of the other variable(s).

A regression problem involving a single predictor (also called simple regression) arises when we wish to study the relation between two variables x and y and use it to predict y from x. The variable x acts as an independent variable. The variable y depends on x and is also subjected to unaccountable variations or errors.

x: independent variable, also called predictor variable, causal variable, or input variable.
y: dependent variable or response variable.

Example 1:

Airline Cost

Can the cost of flying commercial airliner be predicted using regression analysis? If so, what variables are related to such cost?

A few of the many variables can potentially contribute are type of plane, distance, number of passengers, amount of luggage, weather conditions, direction of destination, and perhaps even pilot skill. Suppose a study is conducted using only Boeing 737s traveling 500 miles on comparable routes during the same season of the year. Can the number of passengers predict the cost of flying routes? It seems logical that more passengers result in more weight and more baggage, which could, in turn, result in increased fuel consumption and other costs. Suppose the data displayed in

Table 1 are the costs and associated number of passengers for twelve 500-mile commercial airline flights using Boeing 737s during the same season of the year. We will use these data to develop a regression model to predict cost by number of passengers.

Table (6.1)
Airline Cost Data

Number of Passengers (x)	Cost (y)
61	4280
63	4080
67	4420
69	4170
70	4480
74	4300
76	4820
81	4700
86	5110
91	5130
95	5640
97	5560

Usually, the first step in simple regression analysis is to construct a scatter plot. Graphing the data in this way yields preliminary information about the shape and spread of the data. Figure 1 is a scatterplot of the data in Table 1.

The following scatter diagram (figure 6.1) reveals that the relationship is approximately linear in nature; that is, the points seem to cluster around a straight line. Because a linear relation is the simplest relationship to handle mathematically, we present the details of the statistical regression analysis for this case. Other situations can often be reduced to this case by applying a suitable transformation to one or both variables.

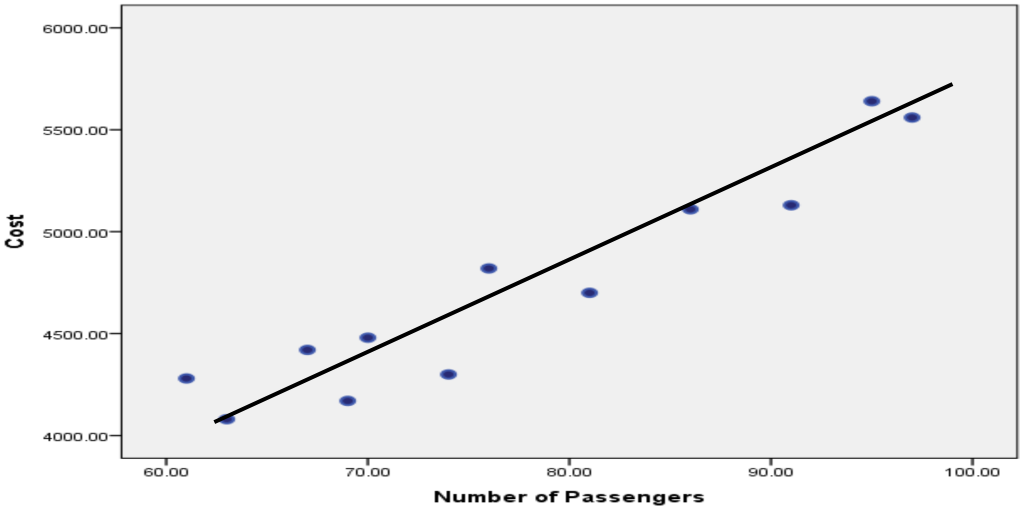


Figure 6.1
Scatterplot of Airline Cost Data

Recall that if the relation between y and x is linear, then the variables are connected by the formula

$$\hat{\mathbf{y}}_{\mathbf{i}} = \mathbf{a} + \mathbf{bx}$$

where “a” indicates the intercept of the line with the y axis and “b” represents the slope of the line, or the change in y per unit change in x (see Figure 2).

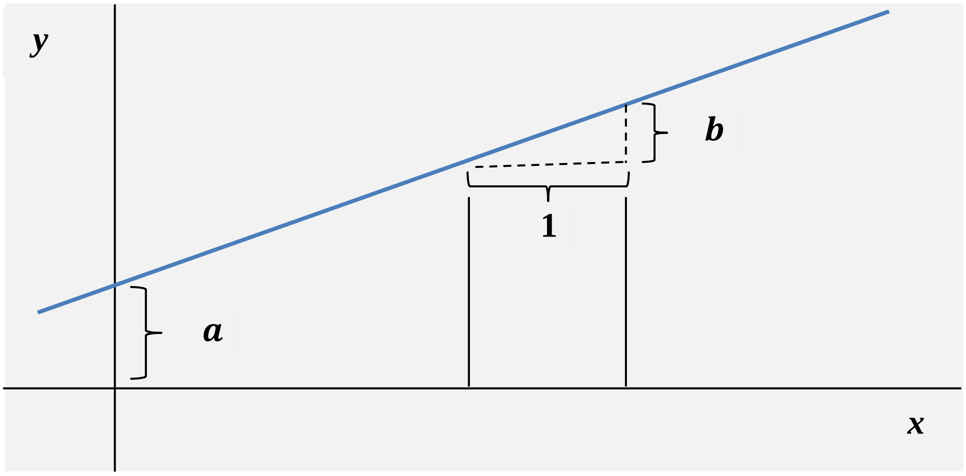


Figure (6.2)
Graph of Straight line $\mathbf{y} = \mathbf{a} + \mathbf{bx}$

General Form of Linear Regression Equation:

$$\hat{y}_i = a + bx_i$$

Where:

- $\hat{y}_i$ read y hat, is the predicted value of the y variable for a selected x value.
- a is the y - intercept. It is the value of y where the regression line crosses x-axis when x = 0.
- b is the slope of the line, or the average change in $\hat{y}$ for each change of one unit (either increase or decrease) in the independent variable x. It is also called "regression coefficient".
- x_i is any value of the independent variable that is selected.

The Formulas of a and b:

Slope of the Regression Line (Regression Coefficient):

$$b = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}$$

or $$b = r \times \left(\frac{S_y}{S_x}\right)$$

Where

r is the correlation coefficient.

$\mathbf{S_y}$ is the standard deviation of y (the dependent variable).

$\mathbf{S_x}$ is the standard deviation of x (the independent variable).

Formula of a: $$a = \bar{y} - b\bar{x}$$

Where

$\bar{\mathbf{y}}$ is the mean of y (the dependent variable).

$\bar{\mathbf{x}}$ is the mean of x (the independent variable).

Note that:

- The regression coefficient (b) and correlation coefficient are of the same sign.
- If **b > 0**, there is a positive relationship between y and x.
- If **b < 0**, there is a negative relationship between y and x.

The Principle of Least Squares:

The difference between each value y_i of the dependent variable and its predicted value $\hat{y}_i$ is called an error. we can choose the estimates a and b to be the values that minimize the distances of the data points to the fitted line. So, we would like to minimize the sum of the squared distances of each observed response to its fitted value. That is, we want to minimize the error sum of squares:

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{e}_{\mathbf{i}}^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y}_{\mathbf{i}} - \hat{\mathbf{y}}_{\mathbf{i}})^{\mathbf{2}}.$$

Example 2:

A researcher is interested in the relationship between the total cost ($1000) against the output (1000 units) of a certain product over a period of 5 weeks, yielding the following data.

Output (x)	4	5	3	6	8
Total cost (y)	40	55	35	70	75

a- Determine the regression equation.

b- Find the predicted cost for producing 7 units.

Solution:

Week	**x**	**y**	**xy**	**x^2**	**y^2**
1	4	40	160	16	1600
2	5	55	275	25	3025
3	3	35	105	9	1225
4	6	70	420	36	4900
5	8	75	600	64	5625
Total	26	275	1560	150	16375

Slope of the regression line (Regression Coefficient):

$$\mathbf{b} = \frac{\left(\frac{\mathbf{1560}}{\mathbf{5}}\right) - \left(\frac{\mathbf{26}}{\mathbf{5}}\right)\left(\frac{\mathbf{275}}{\mathbf{5}}\right)}{\frac{\mathbf{150}}{\mathbf{5}} - \left(\frac{\mathbf{26}}{\mathbf{5}}\right)^{2}} = \frac{\mathbf{26}}{\mathbf{2.96}} = \mathbf{8.784}$$

So, average change in cost per 1 unit change in output is 8.784.

(b) y – intercept:

$$\bar{\mathbf{x}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}}{\mathbf{n}} = \frac{\mathbf{26}}{\mathbf{5}} = \mathbf{5.2} \ \ , \ \ \bar{\mathbf{y}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{y_i}}{\mathbf{n}} = \frac{\mathbf{275}}{\mathbf{5}} = \mathbf{55}$$

$$\mathbf{a} = \bar{\mathbf{y}} - \mathbf{b}\,\bar{\mathbf{x}} = \mathbf{55} - \mathbf{8.784(5.2)} = \mathbf{9.323}$$

So, the cost when there is no output (i.e. fixed costs) is 9.323 ($1000).

Regression equation:

$$\hat{\mathbf{y}}_{\mathbf{i}} = \mathbf{9.323} + \mathbf{8.784x_i}$$

Predicted cost for producing 7 units

$\hat{y}_i = 9.323 + 8.784\ (69) = 70.811$ ($1000)

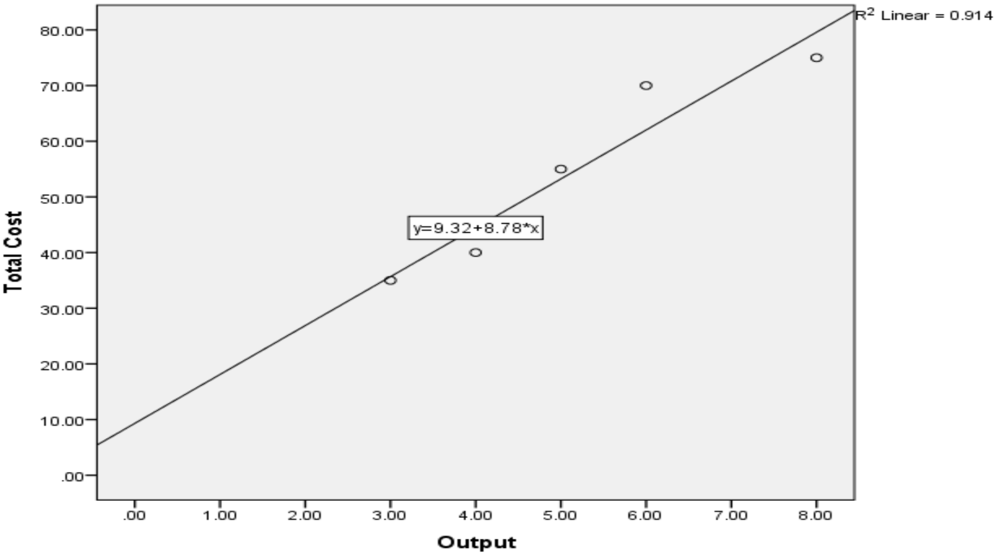


Figure 6.3
Scatterplot of Relationship Between Output and Cost

Example 3:

Using data of **Example 2** in the **previous chapter**:

(a) Determine the regression of age at death on age of retirement.

(b) Predict the age of death for an age of retirement of 65 years.

Solution:

Calculations required were already obtained in **Example 2** in the **previous chapter** as follows:

n = 9 , $\sum_{i=1}^{n} x_i = 537$, $\sum_{i=1}^{n} y_i = 615$, $\sum_{i=1}^{n} x_i y_i = 36670$ $\sum_{i=1}^{n} x_i^2 = 32111$

(a) $$b = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}$$

$$= \frac{\left(\frac{36670}{9}\right) - \left(\frac{537}{9}\right)\left(\frac{615}{9}\right)}{\frac{32111}{9} - \left(\frac{537}{9}\right)^2} = \frac{-2.78}{7.78} = -0.36$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{537}{9} = 59.67 \ , \ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{615}{9} = 68.33$$

$$a = \bar{y} - b\,\bar{x} = 68.33 - (-0.36)(59.67) = 89.81$$

Regression equation:

$$\hat{y}_i = 89.81 - 0.36x_i$$

(b) Predicted the age of death for 56 years age at retirement:

$$\hat{y}_i = 89.81 - 0.36(69) = 64.97 \text{ (\$1000).}$$

Example 4:

For data given in **Example 3** in the **previous chapter:**

(a) Determine the regression of age at death on age of retirement.

(b) Predict the number of rejects for the **Employee E**.

Solution:

Calculations required were already obtained in **Example 3** in the previous chapter as follows:

n = 8 , $\sum_{i=1}^{n} x_i = 72$, $\sum_{i=1}^{n} y_i = 128$,

$\sum_{i=1}^{n} x_i y_i = 1069$, $\sum_{i=1}^{n} x_i^2 = 732$

$$b = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}$$

$$= \frac{\left(\frac{1069}{8}\right) - \left(\frac{72}{8}\right)\left(\frac{128}{8}\right)}{\frac{732}{8} - \left(\frac{72}{8}\right)^2} = \frac{-10.375}{10.5} = -0.99$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{72}{8} = 9 \ , \ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{128}{8} = 16$$

$$a = \bar{y} - b\,\bar{x} = 16 - (-0.99)(9) = 24.91$$

Regression equation:

$$\hat{\mathbf{y}}_{\mathbf{i}} = \mathbf{14.91 - 0.99x_i}$$

(b) Predicted the number of rejects for Employee E that is, for x = 10:

$$\hat{y}_i = 24.91 - 0.99(10) = 15.01 \text{ (\$1000).}$$

The Standard Error of Estimate:

Note in the preceding scatterplot (Figure 6.3) that all of the points do not lie exactly on the regression line. If they all were on the line, there would be no error in estimating the total cost. To put it another way, if all points were on the regression line, total cost could be predicted with 100% accuracy. Thus, there would be no error in predicting the y variable based on x variable.

Perfect prediction in economics and business is practically impossible. What is needed, then, is a measure that describes how precise the prediction of y is based on x or, conversely, how inaccurate the estimate might be. This measure is called the standard error of estimate. The standard error of estimate measures the dispersion about the regression line.

Definition:

Standard Error of Estimate:
A measure of the dispersion, or scatter, of the observed values around the regression line.

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Computational Formula:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a(\sum_{i=1}^{n} y_i) - b(\sum_{i=1}^{n} x_i y_i)}{n-2}}$$

If S_e is small this means that the data are relatively close to the regression line and the regression equation can be used to predict y with little error. If S_e is large, this means that the data are widely scattered around the regression line and the regression equation will not provide a precise estimate for y.

Coefficient of Determination:

We indicated to the coefficient of determination and what it means in the previous chapter. It is widely used measure of fit for regression model that has a more easily interpreted meaning is the coefficient of determination (r^2). It is computed by squaring Pearson’s correlation coefficient.

Definition:

Coefficient of Determination:
The proportion of the total variation in the dependent variable y that is explained by the variation in the independent variable x.

The coefficient of determination ranges from 0 to 1. An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and there is no regression prediction of y by x. An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is explained by x. Of course, most r^2 values are between extremes. The researcher must interpret whether a particular r^2 is high or low, depending on the use of the model and the context within which the model was developed.

Example 5:

For data in **Example 2**, find the standard error of estimate and the coefficient of determination.

Solution:

Standard error of estimate

$$\boldsymbol{S_e} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a\left(\sum_{i=1}^{n} y_i\right) - b\left(\sum_{i=1}^{n} x_i y_i\right)}{n-2}}$$

$$= \sqrt{\frac{16375 - 9.323(275) - 8.784(1560)}{5-2}} = 6$$

Coefficient of determination (r^2):

$$\mathbf{r} = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

$$= \frac{\left(\frac{1560}{5}\right) - \left(\frac{26}{5}\right)\left(\frac{275}{5}\right)}{\sqrt{\left[\frac{150}{5} - \left(\frac{26}{5}\right)^2\right] \times \left[\frac{16375}{5} - \left(\frac{275}{5}\right)^2\right]}}$$

$$= \frac{26}{\sqrt{2.96 \times 250}} = 0.956$$

$$\mathrm{r}^2 = (0.956)^2 = 0.914$$

So, 91.4% of variation in cost is explained by the variation in the output according to the linear relationship between the two variables, while 8.6% of variation in cost can be explained by some factors other than output.

Perfect Linear Relationship:

A perfect linear relationship should satisfy the following:

- Pearson's correlation coefficient (r) equal to -1 or +1, therefore coefficient of determination (r^2) equal to 1, i.e. 100% of variation in the dependent variable is explained by the independent variable.
- Predicted values equal to observed values $(\hat{y}_i = y_i)$ for each value of i, therefore errors equal to zero $(e_i = 0)$ and also the standard error of estimate equal to zero $(S_e = 0)$.

Example 6:

The following data represents horsepower of motors and monthly cost of electricity.

Horsepower of Motor (x)	4	6	8	10	12
Monthly Cost of Electricity (y)	100	150	200	250	300

a- Determine the regression equation.

b- Find the coefficient of determination.

c- Calculate the standard error of estimate.

Solution:

x	y	xy	x^2	y^2
4	100	400	16	10000
6	150	900	36	22500
8	200	1600	64	40000
10	250	2500	100	62500
12	300	3600	144	90000
40	1000	9000	360	225000

(a) Slope of the Regression Line:

$$\mathrm{b} = \frac{\left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i y_i}}{\mathrm{n}}\right) - \left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}}{\mathrm{n}}\right)\left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y_i}}{\mathrm{n}}\right)}{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i^2}}{\mathrm{n}} - \left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}}{\mathrm{n}}\right)^2}$$

$$= \frac{\left(\frac{9000}{5}\right) - \left(\frac{40}{5}\right)\left(\frac{1000}{5}\right)}{\frac{360}{5} - \left(\frac{40}{5}\right)^2} = \frac{200}{8} = 25$$

So, average change in monthly cost of electricity per 1 unit change in horsepower of motor is 25.

y – intercept:

$$\bar{\mathrm{x}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}}{\mathrm{n}} = \frac{40}{5} = 8$$

$$\bar{\mathrm{y}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y_i}}{\mathrm{n}} = \frac{1000}{5} = 200$$

$$\mathrm{a} = \bar{\mathrm{y}} - \mathrm{b}\,\bar{\mathrm{x}} = 200 - 25(8) = 0$$

Regression Equation:

$$\hat{y}_i = 25x_i$$

(b) Coefficient of Determination (r^2):

$$\mathrm{r} = \frac{\left(\frac{\sum_{i=1}^{n} x_i y_i}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)}{\sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right] \times \left[\frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right]}}$$

$$= \frac{\left(\frac{9000}{5}\right) - \left(\frac{40}{5}\right)\left(\frac{1000}{5}\right)}{\sqrt{\left[\frac{360}{5} - \left(\frac{40}{5}\right)^2\right] \times \left[\frac{225000}{5} - \left(\frac{1000}{5}\right)^2\right]}}$$

$$= \frac{200}{\sqrt{8 \times 5000}} = 1$$

$$\mathrm{r}^2 = (1)^2 = 1$$

So, 100% of variation in monthly cost of electricity is explained by the variation in the horsepower of motor.

(c) Standard Error of Estimate:

$$\boldsymbol{S_e} = \sqrt{\frac{\sum_{i=1}^{n} \mathbf{y}_i^2 - \mathbf{a}(\sum_{i=1}^{n} \mathbf{y}_i) - \mathbf{b}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{y}_i)}{\mathbf{n} - \mathbf{2}}}$$

$$= \sqrt{\frac{225000 - 0(1000) - 25(9000)}{5 - 2}} = 0$$

This means that the relationship between the horsepower of motor and cost of electricity is a perfect relationship.

Note: in this case, you can conclude that $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}}$ for each value of **i** which indicates that there isa perfect positive relationship between the two variables.

Exercises

1- Sketch scatterplots (scatter diagram) from the following data and determine the equation of the regression line.

x	12	21	28	8	20
y	17	15	22	19	24

2- Sketch a scatterplot from the following data and determine the equation of the regression line.

x	140	119	103	91	65	29	24
y	25	29	46	70	88	112	128

3- A corporation owns several companies. The strategic planner for the corporation believes dollars spent on advertising can to some extent be a predictor of total sales dollars. As an aid in long-term planning, she gathers the following sales and advertising information from several companies ($ millions).

Advertising	12.5	3.7	21.6	60	37.6	6.1	16.8	41.2
Sales	148	55	338	994	541	89	126	379

Develop the equation of the simple regression line to predict sales from advertising expenditures using these data.

4- Investment analysts generally believe the interest rate on bonds is inversely related to the prime interest rate for loans; that is, bonds perform well when lending rates are down and perform poorly when interest rates are up. Can the bond rate be predicted by the prime interest rate? Use the following data to construct a regression line to predict bond rates by the prime interest rate.

Bond Rate	5	12	9	15	7
Prime Interest Rate	17	15	22	19	24

5- Is it possible to predict the annual number of business bankruptcies by the number of business starts? The following data are pairs of the number of business bankruptcies (1000s) and the number of business starts (10,000s) for a six-year period. Use the data to develop the equation of the regression model to predict the number of business bankruptcies by the number of business starts. Discuss the meaning of the slope.

Business Bankruptcies (1000)	34.3	35.0	38.5	40.1	35.5	37.9
Business Starts (10,000)	58.1	55.4	57.0	58.5	57.4	58.0

6- It appears that over the past 45 years, the number of farms declined while the average size of farms increased. Use the following data to develop the equation of a regression line to predict the average size of a farm by the number of farms. Discuss the slope and y-intercept of the model.

Year	**Number of Farms (millions)**	**Average Size (acres)**
1960	5.65	213
1965	4.65	258
1970	3.96	297
1975	3.36	340
1980	2.95	374
1985	2.52	420
1990	2.44	426
1995	2.29	441
2000	2.15	460
2005	2.07	469
2010	2.17	434
2015	2.10	444

7- Determine the equation of the regression line for the following data and compute the coefficient of determination.

x	15	8	19	12	5
y	47	36	56	44	21

8- Determine the equation of the regression line for the following data and compute the standard error of estimate.

x	12	21	28	8	20
y	17	15	22	19	24

9- Can the annual new orders for manufacturing be predicted by the raw steel production? Use the following data to develop a regression model to predict annual new orders by raw steel production (100,000s of net tons). Construct a scatterplot and draw the regression line through the points.

Raw Steel Production	**New Orders ($ trillions)**
99.9	2.74
97.9	2.87
98.9	2.93
87.9	2.87
92.9	2.98
97.9	3.09
100.6	3.36
104.9	3.61
105.3	3.75
108.6	3.95

10. For data given in **Example 5** in **Chapter 5**:

(a) Determine the regression line of number of breakdowns on age of machine.

(b) Predict the number of breakdowns for a machine of age 5 years. Comment on your result.

(c) Calculate the standard of estimate. Comment.

Exams

1997 - 2022

South Valley University Faculty of Commerce	Date: January 1997 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 5 Questions, 3 Pages

Question (1):

Briefly explain the following terms:

(1) Representative sample **(2)** Inferential statistics

(3) Quantitative variable **(4)** Population

Question (2):

A group of students has the following scores on a statistics test:

63	79	57	81	42	68	73	55	48	65
51	89	40	59	67	78	84	69	62	74

(1) Construct a frequency distribution table with 5 equal classes.

(2) What can you see from your frequency distribution about the data that was not immediately apparent from the set of raw data.

(3) **(a)** Based on your frequency distribution table constructed in (1), prepare a “less-than” cumulative frequency distribution.

(b) Using the cumulative frequency distribution prepared in (a), find the number of students with scores 65 and more.

(4) Assuming that the lower quartile for the frequency distribution has the value, A. Find, without calculations, the value of the upper quartile in terms of A.

Question (3):

(1) On the basis of the following measures of central tendency. What is the nature of the following distributions:

(a) Mean = 68 Median = 62 Mode = 56

(b) Mean = 62 Median = 62 Mode = 62

(c) Mean = 62 Median = 62 Mode = 30 Mode = 45

(2) The accompanying table is a grouped frequency distribution of the weights of 50 people:

Class	15-	25-	35-	45-	55-	65 -75	Total
Frequency	8	9	12	10	7	4	50

(a) Compute the mean and coefficient of variation of the weights.

(b) From the frequency distribution, can you tell whether the data are positively or negatively skewed? Why?
<u>**Hint**</u>: **No calculations required.**

Question (4):

The following data give the experience (in years) and daily salaries (in L.E.) of ten randomly selected salaries:

Experience (x)	4	7	10	8	6	5	9	12	3	6
Daily Salary (y)	6	9	12	8	7	5	10	11	5	7

(1) What would be the advantage of plotting these data using the scatter diagram? (Do not plot the scatter diagram).

(2) Calculate the coefficient of determination. Explain what it indicates in the context of this problem.

(3) Find the regression line with experience as an independent variable and daily salary as a dependent variable.

(4) Should we use this regression equation for predictive purposes? Why?

(5) **(a)** Predict the daily salary for a secretary with an experience of 4 years.
(b) How can you interpret the difference between the actual and predicted daily salary for this secretary?

(6) Find the standard score for x = 7 and explain what it means.

Question (5):

Patients in hospital were asked questions on smoking habits with the following results:

Cigarettes Smoked

Patients with	**0 and under 5**	**5 and over**	**Total**
Lung cancer	52	48	100
Other diseases	78	22	100
Total	130	70	200

What would you say about the association between smoking and lung cancer?

South Valley University **Faculty of Commerce**	**Date: January 1998** **Time Allowed: 3 hours**

Principles of Statistics

English Teaching Section

<u>Answer the Following Questions</u>: 5 Questions, 4 Pages

<u>Question (1):</u>

(1) The following data give the temperature (in Fahrenheit) observed during eight wintry days in a city (hypothetical data)**:**

20 24 19 26 17 28 19 23

Describe the meaning of variable, element, observation, and data set with reference to these data. Is the variable discrete or continuous?

(2) Briefly explain the following terms**:**
A. Random sample **B.** Population

(3) Write short notes on**:**
A. Types of statistics **B.** Types of variables

<u>Question (2):</u>

The following data set gives the number of years for which 24 workers have been with their current employers:

15	12	9	10	5	12	3	7	16	13	11	14
11	8	7	14	11	8	4	13	2	18	6	19

(1) Construct a frequency distribution table for these data, using **1** as the **lower limit** of the **first class** and **4** as the **width** of each class.

(2) Draw a **histogram** corresponding to the frequency distribution prepared in **Part (1)**.

(3) Calculate the relative frequency and percentage for all classes.

(4) Using the appropriate cumulative distribution, what **percentage** of the employees have been with their current employers for **9** years or more?

Question (3):

(1) Briefly explain the meaning of an outlier. Is the mean or the median a better measure of central tendency for a data set that contains an outlier? Illustrate with the help of an example.

(2) Explain the relationship between the mean, median and mode for symmetric and skewed distributions. Illustrate these relationships with graphs.

(3) The following data give the ages of 10 persons**:**

22 19 25 30 29 27 32 98 18 26

A. Find the **mean** and **median** for these data.

B. Using the results obtained in **Part (1)**, how can you know that these data contain an outlier?

C. Drop the outlier and recalculate the mean and median. Which of the two measures changed by a larger amount when you drop the outlier?

D. Is the mean or the median a better measure for these data? Why?

(4) The following table gives information on the amount (in L.E.) of electric bills for November 1997 for a sample of 50 families**:**

Amount of Bill	0-	5-	10-	15-	20-	25-30	Total
No. of Families	7	8	10	12	8	5	50

A. Calculate the **mean** and **variance**.

B. What proportion of families whose bills are between 12 and 25?

C. Calculate the coefficient of skewness. Comment on the skewness of these data.

Question (4):

(1) Consider the following two data sets**:**

Data Set (X):	5	9	4	11	6
Data Set (Y):	3	7	2	9	4

Note that each value of the second data set (Y) is obtained by subtracting 2 from the corresponding value of the first data set (X).

Without calculations, comment on the relationship between the two data sets in regard to**:**

- The mean	**-** The standard deviation
- The correlation coefficient	**-** The regression line of Y on X

(2) A statistics teacher believes that a knowledge of mathematics is essential for a student to do well in a statistics course. At the start of the semester, the teacher administers a standardized test of general mathematics. Later, he compares these scores to the scores of a statistical test. The data are shown in the accompanying table:

Student	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Math Score	90	85	80	75	70	69	72	60	56	53
Stat Score	94	92	81	78	74	73	75	67	54	52

A. Draw a scatter diagram for these data and describe the relationship between the two tests.

B. Calculate the correlation coefficient between the two sets of scores. Does the correlation coefficient support your description of the scatter diagram?

C. Using the coefficient of determination, describe the relationship between the two tests.

D. Would it be fair to state that knowledge of a student's performance on the mathematics test allows you to predict the student's performance in the statistics course?

E. A student obtained a score of **87** on the mathematics test. What is his/her predicted score in the statistics test?

F. Predict the statistics score for the student B. How can you interpret the difference between the actual and predicted statistics score for this student?

G. Find the standardized score for the mathematics score of the student E, and explain what it means.

Question (5):

The personal department of a large corporation recorded the week days during which individuals in a sample of 360 absentees were away over the past several months. The personnel department categorized absentees according to the shift on which they worked, as shown in the accompanying table.

Shift	Sat.	Sun.	Mon.	Tue.	Wed.	Thurs.
Day	65	40	15	19	16	30
Evening	25	30	20	16	14	70

(1) Is there evidence of a relationship between the days on which employees were absent and the shift on which they worked?

(2) Determine the nature and strength of this relationship (if any).

South Valley University **Faculty of Commerce**	**Date: January 1999** **Time Allowed: 3 hours**	
Principles of Statistics		
English Teaching Section		

Answer the Following Questions: 4 Questions, 3 Pages

Question (1): 14 Points

(1) Briefly **explain** the **meaning** of each of the following:

A. Inferential statistics. **B.** Simple random sample

C. Element **D.** Discrete Variable

Question (2):

(1) The following are the scores of **25** students on a statistics examination.

38 10 23 33 46 59 49 35 28 12 30 43 55

19 36 50 26 44 27 52 29 32 48 32 21

A. Calculate the **mean** score.

B. Display these data in a frequency distribution table using **10** as a width of each class, then find the mean.

C. Explain why the mean of the groped data in **Part (2)** would be different from the mean obtained for the ungrouped data in **Part (1)**.

D. Prepare a "less than" cumulative frequency distribution and then find

i. The median

ii. What **proportion** of students would **fail** if the **pass mark** is **30**?

E. Under what conditions is the median a better measure of central tendency than the mean?

F. Based on your results of **Parts (2) and (4 - a)**, do you think that the distribution of scores is symmetric or skewed? Explain.

(2) The following table gives the weekly wages of **20** workers in a certain factory.

Weekly Wage	35-	45-	55-	65 and more	Total
No. of workers	2	6	8	4	20

Given that the **mean** wage is **58,** find the **highest possible wage.**

(3) For a symmetric distribution, if**:**

Mean = 50 and First quartile (Q_1) = 30

Find the semi - interquartile range for this distribution.

Question (3):

(1) For the two variables X and Y, Given that:

- $\sigma_x = 0.5\sigma_y$
- $y_i = \hat{y}_i$ for all values of i.
- X and y are positively correlated
- For x = 1, $\hat{y} = 2$.

A. Show that the two variables X and Y have the same coefficient of variation.

B. What is the regression coefficient of Y on X?

(2) The advertising expenditure (in thousands of L.E.) and product sales (in millions of L.E.) of a company for 8 consecutive months are as follows:

Month	1	2	3	4	5	6	7	8
Expenditure	4	6	2	7	9	4	3	5
Sales	16	22	10	25	26	15	12	18

A. Calculate the standard deviation for both variables; the advertising expenditure and product sales.

B. For what reasons may standard deviation be inappropriate for comparing the variation of the two variables? Suggest a better alternative measure, and find it for both variables.

C. Explain why we construct a scatter diagram.

D. Draw a scatter diagram for these data.

E. By looking at the **scatter diagram** constructed in **Part (4)**, do you expect the correlation coefficient between these two variables to be closer to 0, -1 or +1?

F. Compute the correlation coefficient between the two variables.

G. Find the coefficient of determination. Explain its meaning.

H. Is the value of the correlation coefficient, calculated in **Part (1)** consistent with what you expected in Part (5)?

I. With the advertising expenditure as an independent variable and the product sales as a dependent variable, what do you expect for the sign of the regression coefficient (b) to be? Explain.

J. **Find the predict**ed sales for **month 4**. How can you interpret the difference between the actual and predicted sales for this month?

K. Find the predicted sales for a month with advertising expenditure of L.E.10,000. Comment on the accuracy of this prediction.

L. What would be the effect on sales if advertising expenditure reduced by L.E.1000 per month?

Question (4):

Suppose that a random sample of men and women indicated their view on a proposal of public importance as follows:

Gender	**View**			**Total**
	Opposed	**In Favor**	**Undecided**	
Men	60	30	10	100
Women	90	40	20	150
Total	150	70	30	250

Can we say that the opinion on this issue is the same for men and women? Explain your conclusion.

South Valley University
Faculty of Commerce

Date: January 2000
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 5 Questions, 3 Pages

Question (1):

Briefly explain the following**:**

(1) Population **(2)** Random sample

(3) Types of variables (give examples)

Question (2):

The following data represent **25** ages for a group of persons**:**

21	8	17	22	19	18	19	14	17	11	6	21	25
19	9	12	16	16	10	29	24	5	21	20	25	

(1) Construct a frequency distribution table. Take 5 as the width of each class.

(2) Calculate the relative frequencies for all classes.

(3) For what percentage of persons in this group was the age less than 15 years?

Question (3):

The following table gives the frequency distribution of heights (in inches) for 100 persons**:**

Height	50-	55-	60-	65-	70- 75	Total
No. of Persons	20	25	30	15	10	100

(1) Calculate the mean and standard deviation of the heights.

(2) Find the median of the heights.

(3) From your answers to **Parts (1) and (2)**, what might you conclude about the type and strength of the skewness of the heights?

(4) What proportion of persons whose heights are less than 65 inches?

(5) What height of a person with a standard score (z) of 1.5? Explain your results.

Question (4):

The following table gives the experience (in years) and the number of items which were rejected as unsatisfactory last week for the employees at a small factory.

Employee	A	B	C	D	E	F	G	H
Years of experience (X)	4	5	7	9	10	11	12	14
No. of rejects (Y)	21	22	15	18	14	14	11	13

Compare between X and Y with regard to variation.

(1) Calculate the standard score of X and Y for the employee G. **Comment** on your results.

(2) Find the correlation coefficient between X and Y.

(3) What proportion of the variability in the number of rejects is explained by the variability in the years of experience?

(4) Determine the regression line of Y on X. Comment on the value you have found for the regression coefficient.

(5) Find the predicted number of rejects for the employee **E**. How can you interpret the difference between the actual and predicted number of rejects for this employee?

(6) Compute the standard deviation of errors. Explain what it means.

Question (5):

The following tables shows the number of good and defective parts on each of the three work shifts at a manufacturing plant during randomly sampled periods.

Shift	Day	Evening	Night	Total
Defectives	10	20	20	50
Non-defectives	50	70	80	200
Total	60	90	100	250

Can we say that the shift is independent of whether or not the part produced is defective or non-defective? Explain.

South Valley University
Faculty of Commerce

Date: 10/1/2001
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 5 Questions, 3 Pages

Question (1):

Briefly explain the following:

(a) Simple random sample **(b)** Types of statistics

(c) Continuous variable

Question (2):

The following are the scores made on an intelligence test by a group of children who participated in an experiment:

94	129	100	136	80	98	114	113	124	96
154	109	120	122	139	112	108	127	132	110

(1) Display these data in a frequency distribution with 5 classes of equal width.

(2) Based on your results obtained in **Part (a)**, Construct the cumulative frequency distribution.

(3) Using:

(a) Raw data (Ungrouped data).

(b) Cumulative frequency distribution obtained in Part (2) What proportion of scores are

(i) less than 110? **(ii)** greater than 125?

Comment on your results.

Question (3):

(1) The following table gives the distribution of bonus payments (L.E.) paid to 50 employees in a company.

Monthly Bonus	0-	10-	20-	30-	40-50	Total
No. of Employees	8	12	15	9	6	50

A. Find the mean, **median**, and **standard deviation.**

B. Based on your results obtained in **Part (1)**, determine the type and degree of skewness.

C. Find the semi- interquartile range (quartile deviation).

(2) For the two variables X and Y, given that: Y = 2X

A. What is the **rank correlation coefficient** between X and Y?

B. Compare between **X** and **Y** in regard to:
(i) Mean **(ii)** Variance **(iii)** Coefficient of variation

Question (4):

Ten randomly selected life-insurance salesmen were surveyed in a company to determine the number of weekly sales calls they made and the number of policy sales they concluded. The data shown in the accompanied table were collected:

Salesman No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Weekly Calls (x)	7	4	3	5	6	8	4	2	6	5
Weekly Sales (y)	12	9	7	11	14	13	6	5	13	10

(1) Calculate the correlation coefficient and coefficient of determination. What do these values tell you about the relationship between the two variables?

(2) Find the regression line of y on x; $\hat{\mathbf{y}} = \mathbf{a} + \mathbf{b}\mathbf{x}$.

(3) Give a brief interpretation of the values of a and b found in **Part (2).**

(4) Predict the number of sales concluded by a salesman who makes **9** calls.

(5) What is the predicted value of weekly sales for the salesman number 6? Find the error of estimation and comment on your results.

Question (5):

Indiscipline and violence have become major problem in schools in Egypt. A random sample of **100** adults were selected and they were asked if they favor giving more freedom to school teachers to punish students for indiscipline and violence. The two-way classification of responses of these adults is presented in the following table.

Opinion Gender	In favor	Against	No Opinion	Total
Men	16	11	3	30
Women	12	6	2	20
Total	28	17	5	50

To what extent the opinion on this issue depends upon gender? **Explain** your conclusion.

South Valley University Faculty of Commerce	Date: 23/1/2002 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 4 Questions, 3 Pages

Question (1): Briefly explain each of the following**:**

(1) Population and sample.

(2) Inferential statistics.

(3) Types of quantitative variables, give an example for each type.

Question (2):

(1) The following figures represent the time (in minutes) lost per day through mechanical failure of machinery collected over a period of **40** consecutive working days.

37	14	9	27	54	30	15	23	42	29
24	32	26	18	11	33	25	10	34	7
50	5	26	17	7	32	28	12	38	29
33	18	22	16	31	28	24	19	23	17

A. Prepare a frequency distribution table for these data using five classes of equal widths.

B. Construct a cumulative frequency distribution for the time lost.

C. What proportion of the days had a lost time of:
(i) 35 minutes at least **(ii)** less than **20** minutes

D. Using the frequency distribution of Part **(1),** find the mean, median, standard deviation, and coefficient of skewness for these data.

(2) The following data give the hours worked last week by **5** employees of a company**:**

42 34 40 85 36

A. Find the mean, median, and mode for these data.

B. Does this data set contain any outlier**?** If yes, drop this value and recalculate the mean and median. Which of the two measures changes by a larger amount when you drop this value**?**

C. Is the mean or the median a better measure for these data**?** Explain.

Question (3):

(1) For the two variables ***x*** and **y**, Given**:**

- The two variables have the same coefficient of variation.
- $\hat{y}_i$ **= y**$_i$ for all values of i.
- **x** and **y** are positively correlated.
- **$\sigma_x = 0.2\sigma_y$.**

Find the **predicted** value of **y** for ***x* = 4**. Comment on this prediction.

(2) The following data give information on the ages (in years) and the number of breakdowns during the past month for a sample of **8** machines in a large company.

Machine No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age	12	7	2	8	13	9	4	9
Number of Breakdowns	9	6	1	5	11	7	2	7

A. Find the correlation coefficient between age and number of breakdowns. Comment on your results.

B. Without doing any further calculations show the effect, if any, on the correlation coefficient if each age is increased by **10%**.

C. What proportion of the variability in the number of breakdowns is not explained by the variability in age?

D. Determine the regression line with number of breakdowns as a dependent variable and age of machine as an independent variable.

E. **(5)** Why is the sign of the regression coefficient always the same as that of the correlation coefficient**?** Explain the meaning of the regression coefficient.

F. Should we use the regression line obtained in Pare **(4)** for purposes of prediction**?** Explain.

G. What will be the predicted number of breakdowns for a machine of:
(i) 8 years age **(ii) 10** years age

H. Comment on your predictions in Parts **(a)** and **(b**).

I. Find the standardized value for the number of breakdowns of the **5th** machine. What does this value mean?

J. Compare between age and number of breakdowns with regard to relative variation.

Question (4):

(1) Consider the following two variables**:**

x:	2	3	1	6	5
y:	7	9	5	15	13

Note: Each value of the variable y is obtained by multiplying 2 by the corresponding value of the first variable ***x*** and then adding **3.**
That is, **y = 2 *x* + 3**

Without calculations, comment on the relationship between the two variables in regard to the
(a) mean **(b)** variance
(c) rank correlation coefficient **(d)** regression line of **y** on ***x***

(2) Two samples, one of **200** students from urban secondary schools and another of **250** students from rural secondary schools, were taken. These students were asked if they have ever smoked. The following table lists the summary of the results.

Smoking	Urban	Rural	Total
Have Never Smoked	120	160	280
Have Smoked	80	90	170
Total	200	250	450

What would you say about the association between the two attributes**?**

Discuss your results in the context of these data.

South Valley University
Faculty of Commerce

Date: 16/1/2003
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 4 Questions, 3 Pages

Question (1):

(1) Briefly define each of the following**:**
A. Inferential statistics **B.** Primary data

(2) Distinguish between a discrete variable and a continuous variable, and give examples.

(3) Describe the application of statistics in the following fields**:**
A. Production **B.** Marketing

Question (2):

(1) Before admission into a college, the students have to take Basic Skills Test in fundamentals of mathematics. In one such exam, **40** students appeared in the test. Their scores are recorded below out of a total maximum of **30** points.

15	12	15	22	28	30	19	25	24	28
10	15	16	20	26	22	18	20	27	14
12	19	21	18	19	30	13	10	21	24
15	20	22	18	20	12	23	29	22	24

A. Construct a frequency distribution for the above data with **5** classes and a suitable width of each class.

B. Based on your frequency distribution constructed in **Part (1):**

i. What proportion of students would **fail** if the pass mark is **15 ?**

ii. Find the **percentage** of students whose scores are **at least 22**.

(2) The following data represent the distribution of the annual incomes (in **thousands** of dollars) of **50** households.

Annual Income	**Number of Households**
5 and under 15	15
15 ,, ,, 25	20
25 ,, ,, 35	8
35 ,, ,, 45	5
45 ,, ,, 55	2

A. Calculate the mean and variance for annual household income.

B. Find the income value so that **50%** of households earn less than this value.

C. Is the dustribution skewed**?** If so, find the coefficient of skewness. Explain.

Question (3):

In economics, the demand function for a product is often estimated by the price charged for such a product. The quantity (in **thousands** of units) of new crying baby dolls sold and the corresponding price charged at **10** stores of a large toy store chain for a one week period is shown in the following table.

Store :	1	2	3	4	5	6	7	8	9	10
Quantity:	2	3	5	8	6	10	9	11	14	12
Price (L.E.):	25	22	20	19	17	16	15	13	10	13

(1) Find the correlation coefficient between quantity and price. What do this value tell you about the relationship between the two variables**?**

(2) What is the proportion of the total variation in quantity that is **not** explained by variation in price**?** Interpret your results in regard to this problem.

(3) What do you conclude from your results in **Part (2)** as far as the reliability of predictions is concerned**?**

(4) Predict the quantity expected to be sold if the price of the doll is**:**

(a) L.E. **20** **(b)** L.E. **18** **Comment** on your results.

Question (4):

Four machines, **A** , **B** , **C** , and **D** are used to manufacture certain machine parts which are classified as first grade and second grade. The quality control engineer wants to test whether the quality of the product from the four machines is consistent. The data collected from **100** parts taken from the four machines are classified and tabulated as follows**:**

			Machines		
Grades	**A**	**B**	**C**	**D**	**Total**
First	20	15	14	15	64
Second	10	5	11	10	36
Total	30	20	25	25	100

Can we conclude That the quality of the two grades produced by all machines is the same**?**

South Valley University Faculty of Commerce	Date: 22/12/2002 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 4 Questions, 3 Pages

Question (1):

(1) Briefly define each of the following**:**

A. Descriptive statistics **B.** Population and sample

(2) A Dealer of Toyota cars sold **20,000** Toyota cars last year. He is interested to know if his customers are satistisfied with their purchases. **3000** questionnaires were mailed at random to the purchasers. **1600** responses were received. **1440** of these responses indicated satisfaction.

A. What is the population of interest**?**

B. What is the sample**?**

C. Is the percentage of satisfied customers a parameter or a statistic**?**

(3) Describe the application of statistics in the following fields**:**

A. Purchasing **B.** Production

Question (2):

(1) The following set of data represents the marks obtained by **30** students in Economics in the final exam.

66	61	65	59	87	61	58	70	77	94
45	80	58	78	49	72	75	92	84	82
79	75	68	88	75	90	78	85	69	57

A. (1) Construct a frequency distribution for these data with **5** classes starting at **45** as the lower limit of the first class.

B. Using the frequency distribution of **Part (1)** , what percentage of marks are**:**

- less than **70?** **-** between **65** and **85?**

(2) The department of Labour in a company wants to determine the pattern of daily wages paid to **100** workers. The data were individually collected and combined into a frequency distribution as follows**:**

Weekly Wages (L.E.)	**Number of Workers**
10 and less than **15**	1
15 ,, ,, ,, **20**	4
20 ,, ,, ,, **25**	10
25 ,, ,, ,, **30**	18
30 ,, ,, ,, **35**	28
35 ,, ,, ,, **40**	26
40 ,, ,, ,, **45**	13
Total	100

A. Without calculations , is this distribution skewed **?** If so, then in which direction**?** Explain.

B. Compute for these data**:**
(i) Mean **(ii)** Median **(iii)** Coefficient of variation

C. Find the coefficient of skewness for this distribution. Is the value of this coefficient consistent with your results in **Parts (1)** and **(2)?**

Question (3):

In economics, the demand for an item is often related to the price of the item. An Electronic company has come up with a new electronic toy for children and is trying to estimate the demand function for this new toy at various prices. In a pilot study, **eight** retail stores in different cities were selected and the following data were collected for a **30-day** period.

Price Per Unit (in dollars) :	12	18	13	16	20	14	17	10
Demand (100s of units):	9	2	6	5	1	5	3	9

(1) Find the **least square regression line** of demand on price.

(2) Interpret the meaning of the regression coefficient obtained in **Part (1)**.

(3) How does the quantity demanded change with lowering of price by one dollar each**?**

(4) Calculate the coefficient of correlation and intrepret the nature of the value as calculated.

(5) What proportion of the variation in demand can be explained by its relationship to price**?**

Question (4):

A behavioural scientist is conducting a survey to determine if the financial benefits in terms of the total salary influences the level of satisfaction of the employees or whether there are other factors such as the work environment, which are more important than money. A random sample of **100** employees of an organization are given a test to determine their level of satisfaction. Each employee's total salary is also recorded. The information is tabulated as follows**:**

Level of Satisfaction		**Annual Salary**	
	Under L.E. 3000	**L.E. 3000-6000**	**OverL.E.6000**
High	10	7	3
Medium	26	16	8
Low	14	10	6

Can you conclude that there is a relationship between salary and job satisfaction**?** Explain your results in the context of this issue.

South Valley University
Faculty of Commerce

Date: 19/1/2004
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 5 Questions, 4 Pages

Question (1):

(1) Briefly define each of the following**:**

(a) Simple random sample. **(b)** Inferential statistics.
(c) Continuous variable, give examples.

(2) The following data give the number of new cars sold at a dealership during a 5-day period.

8 5 10 6 9

Describe the meaning of**:**

(a) Variable **(b)** Element **(c)** Observation **(d)** Data set

with reference to these data.

Is the variable, in this case, discrete or continuous**?**

Question (2):

(1) The manager of famous restaurant has received complaints that the customers have to wait too long time in the lounge after they arrive at the restaurant and before they are actually served dinner. He selected a random sample of **30** customers and kept track of their waiting time. The following data recorded, with waiting time measured to the nearest minute.

29	40	25	31	60	68	39	42	60	43
28	52	30	32	48	15	40	21	31	30
51	72	22	29	19	43	43	36	50	32

(a) Develop a frequency distribution using 6 classes.
(b) Compute the average waiting time for the grouped data and compare it with that computed for the raw data. Comment.

(2) A group of **25** students were surveyed as to how much cash did they carry with them on a given day. The data collected is presented as follows**:**

Amount (L.E.)	10-	12-	14-	16-	18-20	Total
Number of Students	2	6	9	6	2	25

A. For the frequency distribution, find:

(i) Mean **(ii)** Median **(iii)** Standard deviation
(iv) Coefficient of skewness **(v)** First and third quartiles

B. What connection do you see between your answers to A.**(ii)** and **A.(v)** and the answer obtained in **A.(iv)**. Explain.

Question (3):

(1) Given that the mean and standard deviation of a set of figures are μ and σ respectively, write down the new values of the mean and standard deviation when**:**

(a) each figure is **increased** by a constant **c**.
(b) each figure is **multiplied** by a constant **k**.

(2) A group of students sat two examinations, one in statistics **(x)** and in mathematics **(y)**. In order to compare the results, the statistics marks were scaled linearly (that is, a mark of **x** became a mark of **ax + b** where **a** and **b** are constants) **so that** the means and standard deviations of the marks in both examinations became **the same**. The original means and standard deviations are shown as follows**:**

	Statistics (*x*)	**Mathematics (y)**
Mean	48	62
Standard deviation	12	10

(a) Find the values of **a** and **b**.

(b) The original marks of a particular student are **36** in statistics, **48** in mathematics. In what sense, has he done better in statistics than in mathematics**?**

(3) Answer the following**:**

A. Which measure of central tendency is affected by some extremely large values or extremely small values so that it ceases to become the representative measure**?**

B. Which measure of central tendency is defined as the value which appears most often in the data.

A. Under what circumstances would median be the most appropriate measure of central tendency**?**

B. Most women use a shoe size of **39**. Which measure of central tendency would most appropriately represent the average shoe size of women**?**

C. In a symmetric distribution, which measure of central tendency has the largest value, if any**?**

D. In a negatively skewed frequency distribution, which measure of central tendency is the largest**?**

E. In comparing the dispersion of two or more distributions, what is the most appropriate measure if they are in different units**?**

<u>Question (4)</u>:

The following table reports heights (in inches) for **10** married couples.

Couple	1	2	3	4	5	6	7	8	9	10
Husband	76	75	72	68	67	62	70	78	77	65
Wife	71	70	67	64	63	61	66	74	72	62

(1) Find the coefficient of correlation between husbands' and wives' heights. What would you say about the strength of the linear relationship between husbands' and wives' heights**?**

(2) What might you expect for the value of the rank correlation coefficient when compared with that obtained in **Part (1)?** **<u>Note:</u> <u>No calculations required</u>**.

(3) Find the regression line of wife's height on husband's height.

(4) What does the regression line tell you about the relationship between husbands' and wives' heights**?**

(5) What percentage of the variation in wives' heights is **unexplained** by the variation in husbands' heights**?**

(6) Predict the height of the wife of a man who is**:**
(a) 74 in tall (**b) 78** in tall

(7) Comment on these predictions.

Question (5):

Four brands of light bulbs are being considered for use in a large manufacturing plant. The director of purchasing asked for samples of **100** from each manufacturer. The numbers of acceptable and unacceptable bulbs from each manufacturer are shown below**:**

	Manufacturer			
	A	**B**	**C**	**D**
Acceptable	12	8	5	11
Unacceptable	88	92	95	89
Total	100	100	100	100

Can you conclude that there is a relationship between the manufacturer and the quality of the bulbs**?** Explain.

South Valley University
Faculty of Commerce

Date: 10/1/2005
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

<u>Answer the Following Questions</u>: 4 Questions, 3 Pages

<u>Question (1)</u>:

(1) Briefly explain the meaning of each of the following terms**:**

A. Random sample.

B. Quantitative variables, give examples.

C. Descriptive statistics.

(2) A sample of midterm grades for **5** statistics students showed the following results**:**

Student No.	**:**	1	2	3	4	5
Grade	**:**	72	65	82	90	76

A. What variable has been measured in this case**?** Is the variable quantitative or qualitative**?**

B. How many elements does this data set contain**?** What are they**?**

C. How many observations in this data set**?**

<u>Question (2)</u>:

(1) A psychologist developed a new test of adult intelligence. The test was administrated to **20** individuals, and the following data were obtained.

112	123	116	128	105	129	118	134	107	126
138	118	100	121	110	117	136	148	109	119

A. Organize the data into a frequency distribution using **5** classes of equal width.

B. Prepare a **"Less than"** cumulative frequency distribution for these data.

C. Use the cumulative frequency distribution constructed in **Part (A)** to determine the proportion of individuals who had a grade of:

(i) at least **112**. **(ii)** between **130** and **145**.

(2) The gross hourly earnings (in dollars) of **50** workers in a large industrial concern were organized into the following frequency distribution.

Hourly Earnings	10-	15-	20-	25-	30-35	Total
Number of Workers	18	12	8	7	5	50

A. Find the **mean** and **variance** of hourly earnings.

B. **(2)** On the basis of inspection only **(<u>Without performing any calculations</u>)**, Do you agree with the claim that the distribution of hourly earnings has a positive skewness**?** Explain.

C. On the basis of your answer to **Part (2)**, would you expect the median to be greater than the mean**?** Justify your answer.

<u>Note: No calculations required</u>.

D. Check your answer to **Part (3)** by computing the median. Comment on the degree to which the distribution of hourly earnings is skewed.

<u>Question (3)</u>:

(1) In order to determine a realistic price for a new product that a company wishes to market, the company's research department selected **10** sites thought to have essentially identical sales potential and offered the product in each at a different price. The resulting sales are recorded in the accompanying table.

Location	1	2	3	4	5	6	7	8	9	10
Price	14	15	16	17	18	20	21	22	23	24
Sales ($1,000s)	14	13	15	9	11	10	8	9	6	5

A. Find he **correlation coefficient**. Does it reflect a strong or weak relationship between prices and sales**?** Is it a direct or inverse relationship?

B. What percentage of variation in sales can be explained by prices**?**

C. Develop the least squares **regression line** of sales on prices.

D. What is the change in sales for each dollar increase in price**?**

E. Find the predicted sales for the **5th** location. How can you interpret the difference between the actual and predicted sales**?**

F. Predict the company sales if the price is **19** in one location. Would you feel comfortable about this prediction**?** Why or why not**?**

G. Determine the **standard error** of estimate.

(2) For the two variables **x** and **y**, given that**:**

A. **x** and **y** have the same **variance**.

B. The correlation coefficient between **x** and **y** equals **0.8**.

C. The mean of **y** is **10**.

D. The predicted value for **y** is **6** if **x** equals **2**.

Compare between the two variables **x** and **y** in regard to the coefficient of variation.

Question (4):

Data on the marital status of men and women aged 25 to 35 were obtained as a part of a national survey. The results from a sample of **300** men and **200** women follow.

Gender	**Marital Status**		
	Never Married	**Married**	**Divorced**
Men	240	50	10
Women	60	120	20

Does there appear to be a relationship between marital status and gender**?** If yes, to which extent**?**

South Valley University Faculty of Commerce	Date: 18/1/2006 Time Allowed: 3 hours
Principles of Statistics	
English Teaching Section	

Answer the Following Questions: 3 Questions, 3 Pages

Question (1):

(1) Briefly explain the meaning of each of the following terms**:**

A. Census and sample survey **B.** Simple random sample

C. Inferential statistics **D.** Continuous variable

(2) The data below give the ages for **5** persons**:**

18 22 19 24 17

Describe the meanings of a **variable**, an **element**, an **observation**, and a **data set** with reference to these data.

Question (2):

(1) The following data show the sales (in thousands of pounds) in a big firm during the last **20** months**:**

61	76	95	50	80	75	84	72	99	71
56	74	88	65	63	79	82	75	78	86

A. Group the above data into **5** classes of equal width.

B. Prepare a **"Less than"** cumulative frequency distribution for these data.

C. Based on your cumulative frequency distribution constructed in **Part (B)**, find the number of months with sales of at least L.E.**72500**.

(2) The following is a frequency distribution for the ages of a sample of **20** employees at a company.

Age	20 -	30 -	40 -	50 -	60 - 70
Number of Employees	4	5	6	3	2

A. Find the **mean**, **median**, and **standard deviation** of ages.

B. Based on your results of **Part (A)**, what would you expect for the type and degree of skewness for this distribution**?**

C. Calculate the coefficient of skewness, Is the value of the coefficient of skewness consistent with what you expected in Part **(B)?**

Question (3):

The following table gives the **monthly output** (in thousands of tons) and **labor cost** (in thousands of dollars) of a factory for a given 8-month period.

Month	1	2	3	4	5	6	7	8
Monthly Output (1,000 tons)	66	74	78	70	82	90	87	85
Labor Cost ($1,000s)	50	57	59	52	64	85	77	68

(1) Calculate the **correlation coefficient (r)** for the two variables shown. Interpret the nature of the value as calculated.

(2) How would your answer of Part **(1)** be affected if output were measured in tons, instead of tons, where **1 ton = 1.016 ton?** Justify your answer.

(3) Compute the **rank correlation coefficient (r_s)** between monthly output and labor cost.

(4) Compare between the values of r and **r_s** as obtained in **Parts (1)** and **(3)**, respectively. Comment on your results.

(5) Would it be fair to state that knowledge of output for a given month allows you to predict the labour cost for this month**?** Explain.

(6) Predict the **labor cost** you would expect for the following months**:**

A. The **4th** month **B.** A month in which output is a **$ 80,000**

Comment on your results of **Parts (a)** and **(b).**

(7) Find the **standard error** of estimate.

(8) Find the **standard score** for the output of the **5th** month, and explain what it means.

Question (4):

(1) Explain the relationship between the **mean**, **median**, and **mode** for **symmetric** and **skewed** distributions. Illustrate with graphs.

(2) The following data give the weights (in kilograms) for **10** persons**:**

8 7 2 5 3 84 7 8 2 4

A. Find the **mean** and **median** for these data.

B. Using the results of Part **(A)**, how can you investigate that these data contain outliers**?**

C. **(3)** If you dropped the outlier and the mean and median were recalculated. Which of the two measures is expected to be changed by a larger amount**?** **No calculations required**.

D. Is the mean or the median a better measure for these data**?** Why**?**

(3) A garage sells three types of new cars (A, B, and C). The following data show, for each type of car sold, the number requiring repair during the first 12 months.

	No. of Cars Sold		
Type of Car	**Requiring Repair**	**Not Requiring Repair**	**Total**
A	89	11	100
B	57	13	70
C	118	12	130

A. Do these figures indicate a difference in quality between different types of cars**?** Explain your conclusions.

B. **Repeat Part(A)** after **excluding** data about cars of **type B**.

South Valley University Faculty of Commerce	Date: 21/1/2007 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

<u>Answer the Following Questions</u>: 4 Questions, 4 Pages

<u>Question (1):</u>

(1) Briefly explain the meaning of each of the following terms**:**

A. Descriptive statistics **B.** Simple random sample

C. Quantitative variable, give examples

(2) A bank is concerned about its level of customer service and conducts a survey into the time which elapses from the moment a customer enters the bank to the moment they finish their transaction. The survey results (to the nearest minute) are as follows.

0	4	1	7	6	0	5	10	8	2
7	5	9	12	4	13	3	2	9	6
6	10	5	2	8	7	9	11	13	11

A. Develop a frequency distribution using **7** classes.

B. Comment on the shape of the distribution.

C. Based on your results in **Part (A)**, what percentage of customers have waiting times of less than **7** minutes**?**

<u>Question (2):</u>

(1) In a medium sized city there are **100** houses for sale of a similar size. The frequency distribution of the asking prices is

Price ($1000)	40-	50-	60-	70-	80-	90-100	Total
No. of Houses	6	8	10	20	30	26	100

A. On the basis of inspection only **(without performing any calculations),** what might you expect for the value of the mean compared with that of the median**?** Justify your answer.

B. Find the price value so that **50%** of houses have prices less than this value.

C. Find the **mean** and **coefficient of variation**.

D. Check your answers to **Parts (1), (2),** and **(3)** by calculating the **coefficient of skewness** for prices.

(2) Given that the mean and variance of a set of figures are **μ** and σ^2 respectively, write down the new values of the mean and variance, **justifying your answer**, when**:**

A. each figure is decreased by a constant **A**.

B. each figure is divided by a constant **K** and then added to a constant **m**.

Question (3):

The following data shows the **age** in years and the second-hand **price** of a sample of **10** cars advertised in a local paper.

Car	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Age (Years)	5	7	6	6	5	4	7	6	5	5
Price ($100)	76	45	58	55	70	88	43	56	69	70

(1) Determine which one is the **X** variable and which one is the **Y** variable. **Explain**.

(2) Calculate the **correlation coefficient**. Explain what it indicates in the context of this problem.

(3) Develop a **regression equation** that could be used to predict the price of a car given its age.

(4) Do you believe the regression equation developed in **Part (3)** would provide a good prediction of the price of car**? Explain**.

(5) Interpret the meaning of the **slope** of the equation obtained in **Part (3)**.

(6) Based on the regression equation in **Part (3)**, What would be the effect on prices if ages were **reduced** by **two** years per car**?**

(7) predict the price for the following cars**:**

(a) A car that is **8** years old **(b)** Car **F**

What would you say about your results in **Parts (a)** and **(b)** as far as the reliability of predictions is concerned**?**

(8) What **percentage of variation** in prices can be **explained** by ages**?**

(9) Find the **standard error** of estimate. **Explain**.

Question (4):

(1) The **scatter diagrams** of two data sets are shown in **Figures (1)** and **(2)**. What conclusions can you draw from these scatter diagrams as far as the simple linear correlation and simple linear regression are concerned**?**

Figure (1)

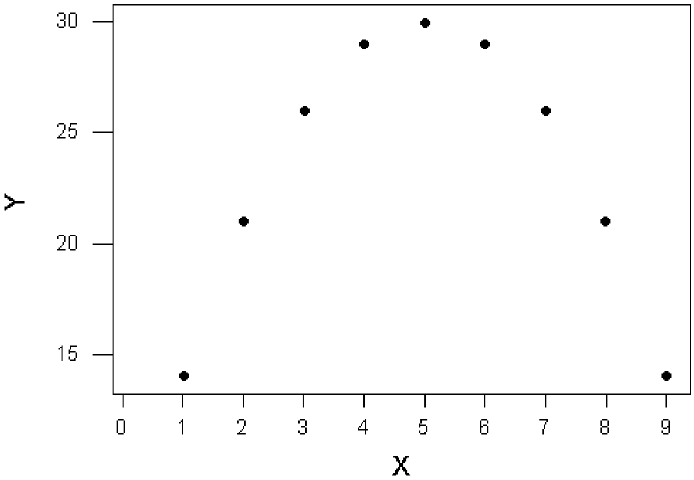
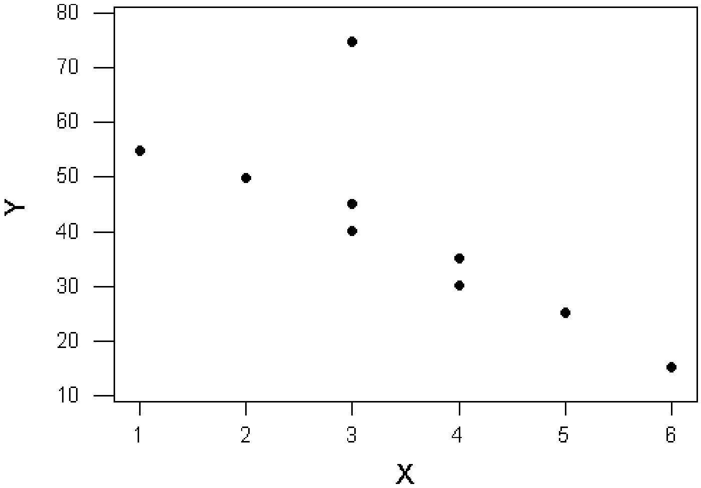


Figure (2)



(2) A convenience store is open 24 hours a day. The owner of the store is interested to find out if there is any relationship between the time a purchase is made and the amount of money spent on that purchase. **200** purchases are selected randomly from the store records. The following summarized data represent the number of purchases made as well as their time of purchase and amount of money spent on each purchase**:**

Time of Purchase	**Size of Purchase**		
	Less than $10	**$10-20**	**Over $20**
Morning	43	25	10
Evening	40	32	8
Night	20	18	4

Can you conclude that the amount of money spent on purchase depends upon the time of purchase**?** If yes, to what extent**?**

South Valley University Faculty of Commerce	Date: 15/1/2008 Time Allowed: 3 hours	

Principles of Statistics

English Teaching Section

Answer the Following Questions: 4 Questions, 3 Pages

Question (1):

(1) Briefly explain the meaning of each of the following terms**:**

A. Population and sample. **B.** Random sample.

C. Discrete variable, give examples.

D. Inferential statistics.

(2) The following data show the time (rounded to the nearest day) required to complete year-end audits for a sample of **20** clients of a small public accounting firm.

10	14	19	18	15	15	18	17	20	27
22	23	13	21	34	28	14	18	16	13

A. Display the data into a frequency distribution, using **five** classes and **10** as the lower limit of the first class.

B. On the basis of inspection only (**without doing calculations)**, is the distribution obtained in **Part (1)** symmetric or skewed**?** Explain.

C. Prepare a **"less than"** cumulative frequency distribution. Then, find the **proportion** of clients with a time of **at most 20** days**?**

Question (2):

(1) An auto insurance company reported the following information regarding the age of a driver and the number of accidents reported last year.

Age of Driver	20-	30-	40-	50-	60-70	Total
No. of Accidents	8	10	16	9	7	50

A. Find the **mean**, **median**, and **standard deviation** for ages.

B. Give the reasons why the values of the **mean** and the **median** should be approximately equal for this distribution.

C. Find the **coefficient of skewness** for the above distribution. Is the value obtained consistent with your answer in **Part (B)?**

D. Determine the **standard score** for the age **60**. Explain what it means.

(2) The accompanying table is the frequency distribution of the statistics scores of a group of students.

Score	0-	2-	4-	6-	8-10
No. of Students	1	2	**k**	2	1

A. For **k** ≥ **3**, explain how does the value of **k** affect the values of
(i) the mean**?** **(ii)** the median**?** **(iii)** the mode**?**

B. Is there a value for **k** such that the value of the **variance** of scores is **zero?** If yes, determine this value.

C. Given the variance of scores is **4.8**, find the value of **k**.

<u>Question (3)</u>:

A company wishes to investigate whether the amount it spends on advertising prior to the launch of a new product is related to the sales volume of the product in the first month. Data from the last **8** product launches is shown below.

Product Number	1	2	3	4	5	6	7	8
Advertising ($10,000)	50	25	20	65	30	17	25	40
Sales (1000 units)	18	14	13	20	14	12	13	16

(1) Find the **correlation coefficient** between advertising and sales. What does the value obtained suggest about the **Direction** and **strength** of the relationship between the two variables**?**

(2) What does the value of the correlation coefficient obtained in **Part (1)** tell you about how useful the advertising is for predicting sales. Explain.

(3) What proportion of the variability in sales **is not** explained by the variability in advertising expenditure**?**

(4) Using your results in **Part (1)**, find the **regression equation** for sales based on advertising.

(5) Based on the regression equation obtained in **Part (4),** what is the marginal effect on volume of sales if the amount spent on **sales** is **decreased** by **$10,000**.

(6) **Predict** the level of **sales** when **$400,000** is spent on advertising. What **proportion** of the error in this case**?** How can you interpret the occurring of this error**?**

(7) Find the **standard error** of estimate. Explain.

(8) For what reasons may the standard deviation be inappropriate for comparing the dispersion of advertising and sales. Suggest better alternative measure, and find it for each variable.

Question (4):

(1) Given are data for two variables, **x** and **y**.

X:	4	2	8	c	10
Y:	6	1	d	26	21

If you know that the two variables **x** and **y** are **perfectly related**.

A. Determine the **values** of **c** and **d**.

B. **(2)** **Without performing any calculations**, what is the **rank correlation coefficient** between **x** and **y?**

(2) A manufacturer of preassembled windows produced **50** windows yesterday. This morning the quality assurance inspector reviewed each window for all quality aspects. Each was classified as acceptable or unacceptable and by the shift on which it was produced. Thus, he reported two variables on a single item. The two variables are **shift** and **quality**. The results are reported in the following table.

		Shift		
	Day	**Afternoon**	**Night**	**Total**
Defective	10	6	24	40
Acceptable	80	54	26	160
Total	90	60	50	200

A. Is there evidence of a relationship between the quality of windows and the shift on which they were produced**?**

B. Discuss your results in the context of this issue.

C. Excluding the **night shift**, Repeat **Parts (A)** and **(B).**

South Valley University Faculty of Commerce	Date: 20/1/2009 Time Allowed: 3 hours
Principles of Statistics	
English Teaching Section	

Answer the Following Questions: 5 Questions , 4 Pages

Question (1): 16 Points

(1) Briefly explain the meaning of each of the following terms**:**

A. Applied statistics **B.** Continuous variable.

C. Simple random sample **D.** Representative sample

(2) A machine produces the following number of rejects in each successive period of five minutes.

16	21	26	24	11	24	17	25	26	13
27	24	26	05	27	23	24	15	22	22
12	22	29	21	18	22	28	25	07	17
22	28	19	23	23	22	06	19	13	31
23	28	24	09	20	34	30	23	20	12

A. Prepare a frequency distribution for these data using **six** classes of equal width. **Hence,**

B. Construct a **cumulative** frequency distribution.

C. Determine the **proportion** of periods that had **20** rejects at least.

D. Find the **number of rejects** so that **30%** of periods have less than this number.

Question (2): 32 Points

(1) The values of **100** properties handled by a property dealer over a six-month period are shown as follows:

Value of Property ($1000s)	15-	20-	25-	30-	35-40
No. of Properties	17	18	30	20	15

A. Find the **mean** and **median** for values of properties.

B. What value of property occurred most frequently**?**

C. Determine the **variance** for values of properties.

D. Compute the **semi-interquartile range (quartile deviation)** for this distribution and discuss why it is often preferred as a measure of variation over the range.

E. Based on your results in **Parts (1)** and **(3)**, what would you say about the **symmetry** of this distribution**?**

F. Confirm or contradict your answer in **Parts (5)** and **(6)** by finding the value of the **coefficient of skewness** for the above distribution.

(2) The following table represents the distribution of the annual incomes **(in thousands of dollars)** of **50** households.

Annual Income	Less Than 15	15-	25-	35-	45-55
No. of Housholds	15	20	8	5	2

If it is known that the **mean** annual income is equal to **$22,400**, determine the **lowest** possible annual income.

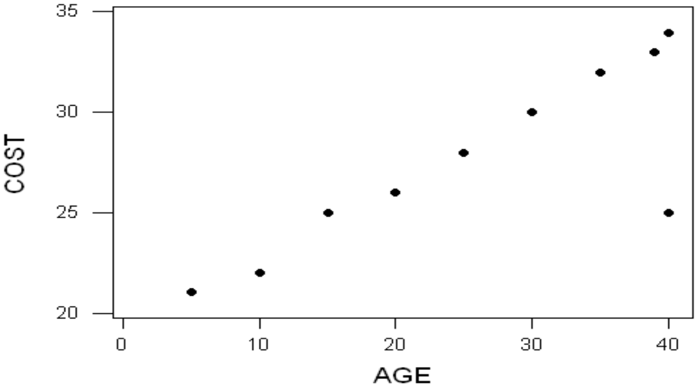
Question (3): 36 Points

The data in the following table relates the weekly maintenance **cost** (in **tens** of dollars) to the **age** (in months) of **ten** machines of similar type in a manufacturing company.

Machine	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age	5	10	15	20	25	30	35	40	40	39
Cost ($10s)	21	22	25	26	28	30	32	25	34	33

(1) Which variable has the greatest relative variability?

(2) The scatter diagram of weekly maintenance cost and age of the machines is shown as follows



From the scatter diagram shown above

A. What does the scatter diagram indicate about the relationship between age and cost**?**

B. Do the data contain outliers**?** If so, identify.

(3) Find the **correlation coefficient (r)** between **age** and **cost**, and describe what it tells you.

(4) How would your answer of **Part (3)** be affected if the **age** of each machine was **increased** by only **one** month. **Explain**.

(5) Find the rank correlation coefficient **(r_s)** between **age** of machine and maintenance **cost**.

(6) Compare between the values of **r** and **r_s** obtained in **(3)** and **(5)**, respectively. Comment on your result.

(7) Determine the **proportion of variation** in cost that would be **explained** by its relationship to age.

(8) Based on your results in **Part (7)**, would you feel comfortable using the age of a machine for predicting its maintenance cost**? Explain**.

(9) Develop a **linear regression equation** in which maintenance cost is to be predicted by age of machine. **Then,**

(10) What are the **fixed costs** of the company**?**

(11) Provide an interpretation for the **slope** of the estimated equation obtained in **(9)**.

(12) **Predict** the maintenance cost for the **4th** machine. Discuss the difference between the actual and predicted cost.

(13) Find the **standard error** of estimate. What does it tell you about the regression equation**?**

Question (4): 16 Points

(1) Given two variables, **x** and **Y** where

$$\mathbf{Y = \frac{x - 4}{2}}$$

Comment on the relationship between **x** and **Y** in regard to the**:**

A. Mean **B.** Median **C.** Variance

D. Rank correlation coefficient

E. Regression line of **Y** on **X**.

(2) A management behavior analyst has been studying the relationship between **male/female supervisory structures** in the workplace and the **level of employees' job satisfaction**. The results of a recent survey are shown here in the following table.

Level of Satisfaction	**Boss/Employee**			
	F/M	**F/F**	**M/M**	**M/F**
Satisfied	20	10	50	120
Dissatisfied	70	100	10	20

F = Female , M = Male

Can you conclude that the level of job satisfaction depends on the boss/employee gender relationship**?** Discuss your results in the context of this issue.

South Valley University Faculty of Commerce	Date: 6/1/2013 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the following questions: 4 Page, 3 Questions

Question (1): 25 Points

(1) Briefly explain the **meaning** of each of the following**:**

A. Simple random sample. **B.** Descriptive statistics.

C. Types of quantitative variables, give examples.

(2) The following data give the statistics score for 6 students.

76 82 95 68 43 57

Describe the **meaning** of each of the following, with **reference** to these data.

(1) Element **(2)** Observation

(3) Variable **(4)** Data set

(3) The following data give the daily wages (in Egyptian pounds) earned by a sample of **30** workers as shown below.

36 28 22 44 30 26 49 24 33 34 25 31 39 33 28

37 42 27 23 27 32 25 34 29 43 32 26 20 28 35

A. Prepare a **frequency distribution** for these data.
Hint: Use **Sturges rule** formula to determine the **number of classes**.

B. Construct a **cumulative frequency distribution**.

C. On the **basis** of the **cumulative distribution** obtained in **Part (B)**, find the **percentage** of workers with a daily wage of **at least** L.E. **27**.

Question (2): 35 Points

(1) The following data give the ages for **6** persons.

65 82 92 86 5 90

A. Find the **mean** and the **median** for these data.

B. Using the results of **Part (A)**, how can you **investigate** that these data contain **outliers? Explain**.

C. If you **dropped** the **outlier** and the values of the mean and median were **recalculated**, which of the two measures is expected to be **changed** by a **larger amount? <u>No calculations required</u>**.

D. Is the mean or the median a better measure for these data**? Explain**.

(2) Given that the **mean** and **variance** of a set of figures are **μ** and **σ^2**, respectively.

What are the **new values** of the **mean** and **variance** when**:**

A. Each figure is **decreased** by a constant **A**.

B. Each figure is **multiplied** by a constant **B**.

Justify your answers.

(3) The following table presents the distribution of the monthly incomes (in **thousands** of Egyptian pounds) of **50** households.

Monthly Income (L.E. 1000)	25-	30-	35-	40-	45-50	Total
Number of households	2	3	5	22	18	50

A. On the bases of **inspection** only **(<u>without performing any calculations</u>)**, do you agree with the claim that the distribution of monthly incomes is **positively** skewed**? Justify** your answer.

B. Calculate the values of the **mean** and **standard deviation** for the monthly incomes.

C. Find the value of the **monthly income** such that **50%** of the households have incomes **less than** this value.

D. Find the **coefficient of skewness** for this distribution. **Comment**.

E. Is the value of the coefficient of skewness obtained in **Part (D)** consistent with your answer in **Part (A)?**

<u>Question (3)</u>: 40 Points

(1) The table given below shows the age (in years) and the second-hand price (in **thousands** of Egyptian pounds) of **8** cars advertised in a local paper.

Car No.	1	2	3	4	5	6	7	8	Total
Age	6	5	4	7	5	8	9	4	48
Price	58	75	85	55	66	50	44	95	528

A. For what reasons may the **standard deviation** be **inappropriate** for comparing the dispersion of the two variables. Suggest better alternative measure.
Note: No calculations required.

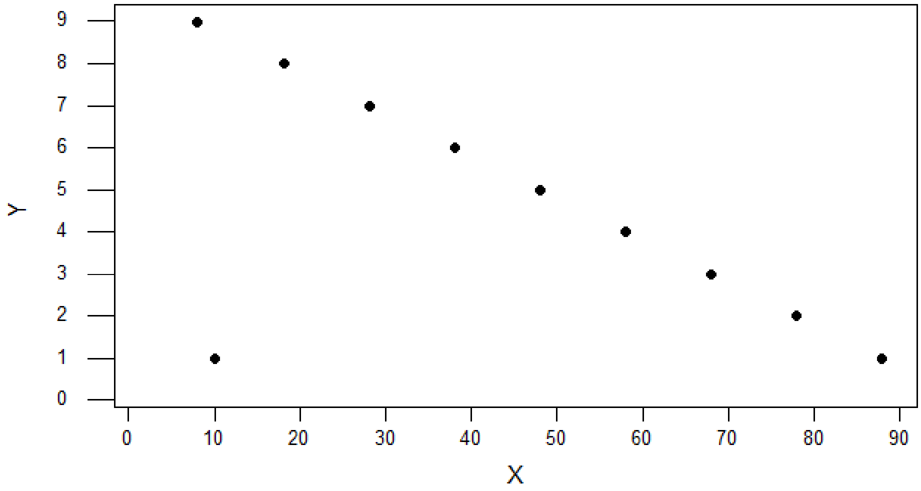
B. Calculate the value of the **correlation coefficient** for the two variables. **Explain** your result in the context of this question.

C. Using the **least squares method**, find the **equation** of the **regression line of price on age**. Would it be fair to state that the knowledge of the age of a given car allows you to predict the price for this car**? Explain**.

D. Predict the price you would expect for the **6th** car. **Comment.**

E. Find the **standardized value** for the **price** of the **4th** car. **Explain** what it means.

(2) The scatter diagram of the two variables **X** and **Y** is shown as follows.



What conclusions can you draw from the scatter diagram shown above as far as the **linear correlation, linear**

regression, and existing of outliers are concerned**?** If outliers exist, **identify**.

(3) Indiscipline and violence have become a major problem in schools in Egypt. A random sample of **200** adults were selected and they were asked if they favor giving more freedom to school teachers to punish students for indiscipline and violence. The three-way classification of responses of these adults is presented in the following table.

Opinion / **Gender**	**In Favor**	**Against**	**No Opinion**	**Total**
Men	96	18	6	120
Women	32	44	4	80
Total	128	62	10	200

Can you conclude that there is a relationship between gender and opinion on this issue**?** If the answer is **'yes'**, to what extent the opinion on this issue depends upon whether the adult is a man or a woman**? Explain** in the context of this issue.

South Valley University
Faculty of Commerce

Date: 5/1/2014
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 3 Questions , 4 Pages

Question (1): 15 Points

(1) Briefly explain the meaning of each of the following terms**:**

A. Inferential statistics. **B.** Random sample.

C. Qualitative variable, give examples.

(2) A bank is concerned about its level of customer service and conducts a survey into the time which elapses from the moment a customer enters the bank to the moment they finish their transaction. The survey results (to the nearest minute) for **20** clients are as follows.

5	12	11	9	10	15	6	14	10	12
17	7	11	9	13	5	12	19	10	8

A. Develop a **frequency distribution** using **5 equal** classes.

B. Based on your results in **Part (A)**, what **proportion** of customers have waiting times of **less than 10** minutes**?**

Question (2): 40 Points

(1) In a medium sized city there are **50** houses for sale of a similar size. The frequency distribution of prices is as follows.

Price ($10,000)	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**Total**
No. of Houses	14	18	12	4	2	50

On the basis of **inspection** only **(without performing any calculations),**

A. what might you expect for the value of the **mean compared** with that of the **median? Justify** your answer.

B. Find the **price** value so that **50%** of houses have prices **less than** this value.

C. Find the **mean** and **coefficient of skewness** of prices.

D. What connection do you see between your answers to **Parts (A)**, **(B)**, and **(C)?**

(2) Given that the **mean** and **variance** of a set of figures are **μ** and σ^2 respectively, write down the **new values** of the **mean** and **variance**, **justifying** your answer, when**:**

A. Each figure is **increased** by a constant **B**.

B. Each figure is **multiplied** by a constant **K** and then **subtracted** from a constant **m**.

(3) Given the following frequency distribution.

Age	2 - 4	4 - 6	6 - 8	8 - 10	10 and more
No. of Persons	1	2	3	K	1

If the values of the **mean** and **variance of the distribution of ages** are **7.5** and **8.25**, respectively, find the following:
A. The **value** of **K**. **B.**The **value** of the **highest age?**

Question (3): 45 Points

(1) Data in the following table relates the weekly maintenance **cost (in tens of L.E.)** to the **age (in years)** of **5** machines of similar type in a manufacturing company.

Machine	1	2	3	4	5	Total
Age (Years)	6	8	5	9	7	35
Cost (L.E. 10s)	4	7	2	7	5	25

A. Determine which one is the **independent** variable and which one is the **dependent** variable. **Explain**.

B. Calculate the **correlation coefficient**. **Explain** what it indicates in the context of this problem.

C. Develop a **regression equation** that could be used to **predict** the maintenance **cost** of a machine **given** its **age**.

D. Does the **regression equation** developed in **Part (C)** provide **good prediction** of the maintenance **cost** of machine**? Explain**.

E. How would your answers of **Parts (B)** and **(C)** be affected if costs were measured in Egyptian pound **(L.E.) instead of** tens of Egyptian pounds **(L.E. 10s)?**

F. Based on the regression equation in **Part (C)**, What would be the **effect** on **costs** if **ages** were **reduced** by **one** year per car**?**

G. **Predict** the **cost** you would expect for the following machines**:**
(i)A machine that is **4** years old.
(ii) Machine number **2**.
What would you say about your results in **Parts (i)** and **(ii)** as far as the **reliability** of predictions is concerned**?**

H. What **percentage of variation** in **costs** can be **explained** by **ages?**

(2) Given the two variables **x** and **Y.** If:

- **64%** of the **variation** in **X** is **explained** by the **variation** in **Y**.
- For each **one increase** in **X**, **Y** is **increased** by **4**.
- The **mean** of **X** is equal to **10**.
- The **two variables** have the **same coefficient of variation**.

Find the **regression equation of Y on X**.

(3) A convenience store is open 24 hours a day. The owner of the store is interested to find out if there is any relationship between the time a purchase is made and the amount of money spent on that purchase. **100** purchases are selected randomly from the store records. The following summarized data represent the number of purchases made as well as their time of purchase and amount of money spent on each purchase**:**

Time of Purchase	**Size of Purchase (in dollars)**			**Total**
	Less than 10	**From 10 to 20**	**Over 20**	
Morning	35	10	5	50
Evening	28	8	4	40
Night	7	2	1	10
Total	70	20	10	100

A. Can you conclude that the amount of money spent on purchase depends upon the time of purchase**?** If yes, to what extent**?**

B. Repeat **Part (A)** after excluding data about the two categories **"Over $20"** and **"Night"**.

C. Compare between your results in **Parts (A)** and **(B)**, **justifying** your answer.

South Valley University
Faculty of Commerce

Date: 4/1/2015
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 3 Questions , 3 Pages

Question (1): 40 Points

(1) Management of a restaurant is concerned with the time a patron must wait before being seated for dinner. Listed below is the **wait time** in minutes, for the **25** tables seated last Friday night.

28 39 21 21 37 32 56 40 66 50 51 45 44

68 43 55 46 41 34 44 48 65 44 35 53

A. Construct a **frequency distribution** for these data using **5** equal classes

B. Prepare a **"Less than"** cumulative frequency distribution. Then, find the **number** of patrons with wait time of **at least 48** minutes.

(2) The following frequency distribution reports the electricity cost (in dollars) of a sample of **50** two-bedroom apartments in a city.

Electricity Cost	30-	40-	50-	60-	70-	80-	Total
Number of Apartments	3	8	12	16	7	4	50

A. Find the **mean, median,** and **standard deviation** of cost.

B. Find the **coefficient of skewness** for this distribution. **Comment**.

(3) Consider the following frequency distribution:

Class	5-15	15-25	25-35	35-45	45-55
Frequency	3	A	6	B	2

Given that: **Mean = 29.5 , Median = 30 , Median Class is 25-35**
Determine the **values** of **A** and **B**.

Question (2): 45 Points

(1) A random sample of **10** homes currently listed for sale provided the following information on **size** (hundreds of square feet) and **price** (thousands of dollars).

Home No.	1	2	3	4	5	6	7	8	9	10
Size	27	28	34	30	29	35	32	40	23	22
Price ($000)	29	31	32	33	36	41	49	55	25	19

A. Find the **correlation coefficient** between **size** and **price**. What would you say about the **relationship** between the two variables**?**

B. How would your answer of **Part (A)** be affected if price is measured in hundreds of dollars instead of thousands of dollars?

C. What **proportion of variability** in **price** that is **not explained** by **size**?

D. What do you conclude from the result in **Part (A)** as far as the reliability of predictions is concerned?

E. Find the **predicted price** for each of the following**:**
(i) Home **number 4**. **(ii)** A home of **size 2500 square feet**.

F. Comment on these predictions.

G. Determine the **standard error of estimate.**

H. Find the **standard score** for the **price** of the **8th** home. **Explain** what it means.

(2) For the two variables **x** and **y**, given that**:**

- $y_i = \hat{y}_i$ for all values of i.
- For each unit **increase** in **x**, **y** is **decreased** by **2**.
- The **mean** and **variance** of **x** are **8** and **4** respectively.
- For **x = 5**, **y = 10**.

A. Compare between **x** and **y** in regard to**:**
(i) Mean **(ii)** Variance **(iii)** Coefficient of variation
(iv) Rank correlation Coefficient

B. Predict the value of **y** for **x = 2**.

Question (3): 15 Points

A study regarding the relationship between age and the amount of pressure sales personnel feel in relation to their jobs revealed the following sample information.

Age (years)	**Degree of Job Pressure**			**Total**
	Low	**Medium**	**High**	
Less than 25	70	20	10	100
25 - 40	70	50	80	200
40 - 60	10	30	160	200
Total	150	100	250	500

(1) Does the sample information provide evidence to conclude that the degree of job pressure depends upon age**?** Explain your results in the context of this issue.

(2) Repeat **Part (1) excluding** the category of **"medium"** for the degree of job pressure and the category **"25 up to 40"** for age. Comment on your answer.

South Valley University Faculty of Commerce	Date: 17/1/2016 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 3 Question , 3 Pages

Question (1): 15 Points

(1) Briefly **explain** the **meaning** of each of the following:

A. Qualitative variable, give examples. **B.** Descriptive statistics.

(2) Suppose the following data are the ages of Internet users for a sample.

39	15	31	25	24	23	21	22	22	18
19	16	23	27	34	24	19	20	29	17

A. (1) Organize these data into a **frequency distribution** using **5** as a **width** for each class.

B. Based upon the frequency distribution obtained in **Part (1):**

(i) Prepare a **"Less than" cumulative** frequency distribution.

(ii) Find the **percentage** of Internet users with age of **at least 32** years.

Question (2): 35 Points

(1) The following frequency table gives the distribution of bonus payments (in hundreds of L.E.) made to **100** employees in a company.

Monthly Bonus (L.E. 00)	30-	40-	50-	60-	70-80	Total
Number of Employees	12	20	36	20	12	100

A. On the basis of inspection only (no calculations required), do you agree with the claim that the **distribution** of monthly bonus is **positively skewed**? **Explain**.

B. Find the **mean** and **standard deviation** of monthly **bonus**.

C. Find the **bonus value** so that **75%** of employees have bonuses **less than** this value.

D. **Based on** your results of **Parts (A), (B), and (C)**, determine the **bonus value** so that **75%** of employees have bonuses **greater than** this value.

(2) Consider the following frequency distribution**:**

Score	Less than 10	10 - 20	20 - 30	30 - 40
Number of students	4	B	8	2

If: Mean = 19.6 and **Variance = 68.64,**
determine the following**:**

a. The value of **B**. **b.** The **lowest score**.

<u>Question (3)</u>: 50 Points

(1) The following table presents the salaries (in hundreds of L.E.) and years of experience for a sample of **8** workers.

Worker No.	1	2	3	4	5	6	7	8	Total
Years of Experience	7	2	5	8	9	7	6	4	48
Salary (L.E. 00)	20	6	15	18	22	19	16	12	128

A. For what reasons may **standard deviation** be **inappropriate** for comparing the dispersions of the two variables? Suggest **better** alternative measure.

B. Compute the **correlation coefficient (r)** between **years of experience** and **salary**. **Describe** what it tells you.

C. Find the **coefficient of determination** and **interpret** its meaning.

D. How would your answer of **Part (B)** be **affected** if the salary of each worker is **multiplied** by **2** and then **subtracted** from **40**? **Justify** your answer.

E. What **percentage of variation** in **salary** is **explained** by **years of experience**?

F. How useful do you think the years of experience is as a predictor of salary? **justify** your answer.

G. Find the **rank correlation coefficient (r_s)** between **years of experience** and **salary**.

H. Compare between the values of r and **r_s. Comment**.

I. Find the **predicted salary** for each of the following**:**

(i) The worker **number 6**. **Comment**.

(i) A worker with experience of **3** years. Would you feel comfortable about this prediction? **Justify** your answer.

J. **Interpret** the value of the **regression coefficient** obtained in **Part (I)**.

(2) For the two variables **x** and **y**, given**:**

- The two variables have the **same coefficient of variation**.
- **$y_i = \hat{y}_i$** for all values of i.
- The two variables are **positively related**.
- **$\sigma_y = 2\sigma_x$** .

A. **Compare** between **x** and **y** in regard to**:**

(i) Mean **(ii)** Variance

(iii) Rank correlation coefficient

B. **Predict** the value of **y** for **x = 4**.

(3) Data on the **social class** and the **number of children** in a family were obtained as a part of national survey. The results from a sample of **200** families follow.

Number of Children	**Social Class**			**Total**
	Lower	**Middle**	**Upper**	
1 or 2	12	24	84	120
More than 2	64	12	4	80
Total	76	36	88	200

A. **(1)** To what extent you can say that the **number of children** in a family **depends upon** the **social class** of this family**? Explain** your result.

B. **Repeat Part (A) excluding** the **category** of **"middle"** for the social class.

C. **Compar**e between your results in **Part (A)** and **Part (B)**. **Comment**.

South Valley University Faculty of Commerce	Date: 15/1/2017 Time Allowed: 3 hours
Principles of Statistics	
English Teaching Section	

Answer the Following Questions: 3 Questions , 4 Pages

Question (1): 20 Points

(1) Briefly **explain** the **meaning** of each of the following:

A. Simple random sample. **B.** Inferential statistics.

(2) The data below give the ages, to the nearest year, for **5** children**:**

8 2 5 4 6

Describe the meaning of**:**

(a) Variable **(b)** Element **(c)** Observation

With reference to these data. Is the variable, in this case, **discret**e or **continuous?**

(3) The following data (in thousands of dollars) represent the net **annual income** for a sample of **20** taxpayers**:**

41	34	20	39	29	38	42	30	37	44
26	32	30	43	33	39	32	35	33	31

A. Construct a **frequency distribution** for these data having **5** classes.

B. Based upon the **frequency distribution** obtained in **Part (A):**

(i) Prepare a **'Less than' cumulative** frequency distribution.

(ii) Determine the value of the **annual income** so that **40%** of taxpayers will earn **at least** this value.

Question (2): 40 Points

(1) The income distribution of the middle management in a large organization is tabulated below**:**

Income ($ hundreds)	20-	25-	30-	35-	40-	45 - 50	Total
Number of Managers	4	5	6	15	12	8	50

A. **On the basis of inspection only (without performing any calculations)**, **compare** and **contrast** the values of the **mean** and **median** you would **expect** for the income distribution.
Justify your answer.

B. Find the **mean** and **standard deviation** of **income**.

C. Find the **income value** so that **50%** of managers have incomes **less than** this value.

D. Determine the **type** and **degree** of **skewness** of the income distribution.

E. Is your answer of **Part (4) consistent** with what you **expected** in **Part (1)**? **Explain**.

(2) The accompanying table is the frequency distribution of the statistics scores of a group of students.

Score	4-6	6-8	8-10	10-12	12-14
No. of Students	2	5	k	5	2

A. For **k > 5**, **explain** how does the value of **k** affect the values of
(i) Mean**?** **(ii)** Median**?** **(iii)** Standard deviation?

B. Is there a value for **k** such that the value of the **variance** of scores is **zero?** If **yes**, determine this **value**.

C. For **k > 5**, if the value of the **third quartile (**Q_3) is equal to **10.8**, **use this value** to **determine** the value of the **first quartile (**Q_1**)**. **Explain**.

D. Given the **variance** of scores is **5.2**, find the **value of k**.

Question (3): 40 Points

(1) A new-car dealer is interested in the **relationship** between the **number of salespeople** working in a month and the **number of cars sold**. Data were gathered in **5** consecutive months**:**

Month	1	2	3	4	5	Total
Number of Salespeople	5	6	4	2	8	25
Number of Cars sold	22	20	14	9	25	90

A. What would be the **advantage** of **plotting** these data using the **scatter diagram**? <u>**Hint**</u>**: Do not draw the scatter diagram**.

B. Calculate the values of the **correlation coefficient (r)** and the **coefficient of determination**. What do these values tell you about the relationship between the two variables**?**

C. How much of **the variation** in the **number of cars** sold is **not explained** by the **number of salespeople**?

D. Compute the **rank correlation coefficient (r_s)** between the **number of salespeople** and the **number of cars sold**.

E. Compare between the values of **r** and **r_s** as obtained in **Parts (B)** and **(D)**, respectively**. Comment** on your results.

F. Would it be fair to state that knowledge of the number of salespeople for a given month allows you to predict the number of cars sold for this month**? Justify** your answer.

G. Predict the **number of cars sold** you would **expect** for the following months**:**
(i) The **2nd** month. **(ii)** A month with **7** salespeople.

Comment on your predictions of **Parts (i)** and **(ii).**

H. If the **number of salespeople** was **reduced** by **2** per month, what would be the effect on**:**
(i) Value of the correlation coefficient?
(ii) Number of cars sold?

I. Find the **standard score** of the **number of cars sold** for the **5th** month, and **explain** what it means.

(2) For the **two variables x** and **y**, given**:**

- The **correlation coefficient** between **x** and **y** equals **0.6**.
- For x **= 4,** the **predicted value of y** is **5** .
- The **mean of x** is **8**. - **$S_x = 0.8S_y$** .

A. Determine the **regression line** of **y on x**.
B. Find the **mean** for the variable **y**.

(3) An insurance company is interested in determining whether there is a **relationship** between automobile **accident**

frequency and **cigarette smoking**. It randomly sampled **50** policyholders and came up with the following data**:**

Cigarette Smoking	**Number of Accidents**			**Total**
	0	**1**	**2 or more**	
Smokers	8	6	4	18
Nonsmokers	12	10	10	32
Total	20	16	14	50

A. Does the sample provide sufficient information to conclude that there is a relationship between automobile accident frequency and cigarette smoking? **If yes**, to **what extent**? **Discuss** your results in the **context** of this issue.

B. Does it make difference to your results of **Part (A)** if the category **'2 or more'** is **excluded**? **Explain**?

South Valley University
Faculty of Commerce

Date: 26/12/2017
Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

Answer the Following Questions: 3 Questions , 3 Pages

Question (1): 14 Points

(1) Briefly **explain** the **meaning** of each of the following:

A. Descriptive statistics. **B.** Discrete variable, give examples.

(2) Before admission into a college, the students have to take Basic Skills Test in fundamentals of mathematics. The scores of **25** students are recorded below out of a total **maximum** of **40** points.

Hint: Pass Mark = 60% of Full Mark.

15 12 15 22 28 30 19 25 24 28 10 23 16

20 26 22 18 20 27 14 12 19 21 24 32

A. Prepare a **frequency distribution** for these data using **5** as **width** for each class.

B. Using your result in **Part (A),** prepare a **'Less than' cumulative** frequency distribution. Then, find the **proportion of students** who **passed** the exam.

Question (2): 32 Points

(1) The following table presents the distribution of the hourly wages of **20** workers in a certain factory.

Hourly Wage	20 - 25	Less Than 30	Less Than 35	Less Than 40	Less Than 45
No. of workers	2	A	14	18	20

If the value of the **median** is **32.5,** find the **value** of **:**

(i) A **(ii)** The **Mode**

(2) The time taken for the weekly maintenance of a group of machines in a workshop over the past **25** weeks is shown in the following table.

Maintenance Time (hours)	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10	Total
Number of Weeks	2	6	10	5	2	25

A. Find the **mean**, **variance** and **coefficient of variation** for **maintenance time**.

B. Determine the **maintenance time below** which **10** weeks will lie.

C. On the basis of inspection only (without any computations), can you say that this distribution **is extremely skewed to the right**? **Explain**.

<u>Question (3)</u>: 54 Points

(1) The following table represents the daily production (units) and the number of workers assigned for each of the **8** days.

Day	1	2	3	4	5	6	7	8	Total
Number of Workers	5	7	6	8	9	7	8	6	56
Production	12	14	13	17	19	16	16	13	120

A. For what reasons may **standard deviation** be **inappropriate** for comparing the dispersion of the two variables? Suggest better alternative measure and use it for such a comparison.

B. Find the **correlation coefficient (r)** between the two variables. **Explain** what it indicates in the context of this problem.

C. How would your answer of **Part (B)** be affected if the **number of workers** for **each day** is**:**
(i) decreased by 2. **(ii) multiplied** by **2**.

D. Find the **coefficient of determination**. **Explain** its meaning.

E. The value of the **rank correlation coefficient (r_s)** was found to be **0.97**. What would you say about this value when **compared** to that of **the correlation coefficient (r)** obtained in **Part (B)**?

F. Predict the **number of units produced** you would **expect** for**:**

i. Day no. **4**. How can you **interpret** the **difference** between the **actual** and **predicted** number of units produced?

ii. A day in which **10** workers were assigned. **Comment**.

G. Interpret the value of the **regression coefficient** obtained in **Part (6)**.

H. Find the **standard error** of the **estimate**. **Comment**.

(2) Given are data for **two variables x** and **y**.

X	4	A	2	7
y	4	2	0	B

If you know that the **two variables** are **perfectly related**.

A. Determine the values of **A** and **B**.

B. Compare between the two variables in regard to the **coefficient of variation**.

C. Without any computations, what is the value of the **rank correlation coefficient (r_s)**?

(3) A sample of **200** units produced by a machine were classified as good or defective and by the shift on which they were produced. The results are reported in the following table.

Quality	**Shift**			**Total**
	First	**Second**	**Third**	
Good	76	64	40	180
Defective	4	6	10	20
Total	80	70	50	200

A. Is there evidence of a **relationship** between the **quality** of the units produced and the **shift** on which they were produced? **If yes**, to **what extent**?
Explain your conclusion in the **context** of this issue.

B. Excluding the **second shif**t, **repeat Part (A)**.

C. Compare between your **results** of **Parts (A)** and **(B)**

South Valley University Faculty of Commerce	Date: 9/1/2019 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

You have 25 questions. Please mark all your answers on the answer sheet provided to you. You have to submit both questions' papers and answer's sheet.

- **Make sure that the answer sheet form matches the questions form.**
- **Choose the best answer for each of the following questions.**

<u>Hint</u>: <u>Five points are assigned for each of the following six questions (Q1-Q6)</u>.

Q1. The following data show the time (rounded to the nearest day) required to complete year-end audits for a sample of **20** clients of a small accounting firm.

10	24	19	18	25	15	18	17	20	27
22	32	13	21	34	24	14	26	16	13

The **frequency distribution** for these data using **5** equal classes is:

(A)

Time (day)	10-15	15-20	20-25	25-30	30-35
Number of Clients	3	8	5	3	1

(B)

Time (day)	10-15	15-20	20-25	25-30	30-35
Number of Clients	4	7	4	3	2

(C)

Time (day)	10-15	15-20	20-25	25-30	30-35
Number of Clients	4	6	5	3	2

The time taken for the weekly maintenance of a group of machines in a factory over the past 25 weeks is shown in the following table:

Maintenance Time (hours)	0 -	2-	4-	6-	8-10	Total
Number of Weeks	2	6	10	5	2	25

Answer the following six questions (Q2 - Q7):

Q2. The **mean** of the maintenance **time** is

(A) 5.12 (B) 4.92 (C) 4.85 (D) 4.55

Q3. The of the **median** of this distribution is equal to

(A) 4.8 (B) 5.2 (C) 5.4 (D) 4.9

Q4. The **standard deviation** of the time is

(A) 2.8 (B) 2.1 (C) 3.2 (D) 2.4

Q5. What is the **maintenance time** so that **80%** of weeks have time **less than** this value?

(A) 6.6 (B) 7.2 (C) 6.8 (D) 7.4

Q6. The **number of weeks** that have maintenance **time** of **at least 9** hours is

(A) 1 (B) 3 (C) 2 (D) 4

Hint: Three points are assigned for each of the following six questions (Q7-Q12).

Q7. On the basis of **inspection** only **(no need to waste time for calculations)**, the distribution of maintenance time is

(A) Highly skewed to the right (B) Almost symmetric
(C) Highly skewed to the left (D) Symmetric

Q8. Which **measure of central tendency** is **affected** by **extreme** values?

(A) Mean (B) Median (C) Mode (D) First quartile

Q9. Most women use a shoe size of 39. Which statistical measure would most appropriately represent the average shoe size of women?

(A) Mean (B) Median (C) Mode (D) First quartile

Q10. In a **symmetric** distribution, which measure of the central tendency has the largest value, if any?

(A) Mean (B) Median (C) Mode (D) None of the above

Q11. For a **symmetric distribution**, if **Mean = 50** and **First quartile (Q_1) = 20**, the value of the **quartile range ($Q_3 – Q_1$)** is

(A) 80 (B) 60 (C) 40 (D) 50

Q12. The value that has half of the observations above it and half of the observations below it is called

(A) Median (B) Mean (C) Mode (D) Third quartile

Hint: Four points are assigned for each of the following 13 questions (Q13-Q25).

For the two variables x and y, if each value of y is obtained by multiplying the value of x by -2 and then adding 50. If the mean and variance of the variable x are 20 and 64 respectively.

Answer the following two questions (Q13 and Q14):

Q13. The **mean** of **y** is

(A) 8 (B) 15 (C) 12 (D) 10

Q14. The **standard deviation** of **y** is

(A) 12 (B) 18 (C) 16 (D) 10

The following data give information of the age (in years) and the number of breakdowns during the past month for 5 machines in a small company.

Machine No.	1	2	3	4	5	Total
Age (Year)	7	2	4	8	9	30
Number of Breakdowns	5	1	2	5	7	20

Answer the following six questions (Q15- Q20):

Q15. What is the value of the **correlation coefficient (r)** between the two variables?

(A) 0.98 (B) 0.95 (C) 0.88 (D) 0.85

Q16. The value of the **rank correlation coefficient (r_s)** is

(A) 0.975 (B) 0.895 (C) 0.925 (D) 0.875

Q17. The value of the **coefficient of determination** is

(A) 0.72 (B) 0.90 (C) 0.77 (D) 0.96

Q18. The **regression line** of **number of breakdowns (y) on age (x)** is

(A) $\hat{y}$ = - 0.82 + 0.92x **(B) $\hat{y}$ = - 0.92 + 0.82x**

(C) $\hat{y}$ = 0.82 + 1.2x **(D) $\hat{y}$ = -1.2 + 0.82x**

Q19. What is the **predicted** number of **breakdowns** for a machine of **5** years old?
<u>Round your answer to the nearest integer</u>.

(A) 2 **(B) 4** **(C) 1** **(D) 3**

Q20. What is the **standard value** for the **age** of the **4th machine**?
(A) 0.77 **(B) 0.85** **(C) 0.96** **(D) 0.82**

For the two variables x and y, given:
- The two variables have the same coefficient of variation.
- y_i = $\hat{y}_i$ for all values of i**.** **-** x and y are **negatively correlated.**
- The **mean** of **x** is **8**. **- s_y = 0.5s_x.**

Answer the following <u>two Questions</u> (Q21 and Q22)
Q21. What is the **predicted** value of **y** for **x = 10**?
(A) 1.5 **(B) 2.5** **(C) 3** **(D) 2**

Q22. The value of The **rank correlation coefficient** (**r_s**) is
(A) -0.9 **(B) -1** **(C) 0.8** **(D) -0.95**

Q23. The percent of **total variation** of the **dependent** variable **y explained** by the **independent** variable **x** is measured by
(A) Coefficient of determination
(B) Coefficient of correlation
(C) Rank correlation coefficient
(D) Coefficient of skewness

A company sells three types of new cars (A, B and C). The following data show, for each type of car sold, the number requiring repair during the first two years.

Number of Cars sold	**Type of Car**			**Total**
	A	**B**	**C**	
Requiring Repair	20	60	80	160
Not Requiring Repair	10	10	20	40
Total	30	70	100	200

Answer the following two questions (Q24 and Q25):

Q24. The value of the **contingency coefficient (Cramer's Coefficient)** is

(A) 0.18 (B) 0.25 (C) 0.15 (D) 0.32

Q25. Excluding the **type B** of cars, what is the value of **the coefficient of association (Yule's coefficient)?**

(A) 0.28 (B) 0.36 (C) 0.42 (D) 0.33

	South Valley University Faculty of Commerce	Date: 9/1/2020 Time Allowed: 3 hours	

Principles of Statistics

English Teaching Section

You have 30 questions. Please mark all your answers on the answer sheet provided to you. You have to submit both questions papers and answer sheet.

Please:

- **Make sure that the answer sheet form matches the questions form.**
- **Choose the best answer for each of the following questions.**

Hint: Two points are assigned for each of the following 10 questions (Q1- Q10).

Q1. In a **symmetric** distribution, if $\mathbf{Q_3 - Q_1 = 20}$ and **median = 15**, then $\mathbf{Q_3}$ is equal to

(A) 20 (B) 25 (C) 10 (D) 15

Q2. The **variance** is **zero** if observations are the

(A) Different (B) Squares (C) Square roots (D) Same

The **mean** and **variance** of a set of numbers is **20** and **4** respectively. If each number is **multiplied** by **2** and then **increased** by **5**.
Answer the following two questions (Q3 - Q4):

Q3. The **mean** of **new numbers** is

(A) 30 (B) 36 (C) 45 (D) 40

Q4. The **standard deviation of new numbers is**

(A) 10 (B) 4 (C) 6 (D) 5

Q5. In a set of observation, the **variance is 50**. All the observations are **increased** by **100%**. The **variance** of the **increased observations** will become

(A) 200 (B) 240 (C) 180 (D) No change

Q6. If the **correlation coefficient** between ***x*** and **y** equals **0.75**, then the **correlation coefficient** between **u = 2*x*** and **v = 2y** is

(A) 0 (B) - 0.75 (C) 1.5 (D) 0.75

Q7. The **signs** of the **regression coefficient** and **correlation coefficient** are always

(A) Different (B) Positive (C) Same (D) Negative

Q8. If the **percent** of **total variation** of the **dependent** variable **y unexplained** by the **independent** variable **x** is **24%**. The **correlation coefficient (r)** between **x** and **y** is

(A) 0.49 (B) 0.85 (C) 0.76 (D) 0.87

Q9. If the value of any **regression coefficient** is **zero**, then the two variables are

(A) Correlation (B) Independent
(C) Dependent (D) Quantitative

Q10. If the **coefficient of determination** is a **positive** value, then a **regression equation**

(A) Must have a positive slope
(B) Must have a negative slope
(C) Could have either a positive or a negative slope
(D) Must have a positive y- intercept

Hint: Three points are assigned for each of the

Q11. The following data show sales (in thousands of dollars) in a firm during the last **20** months.

61	76	95	50	80	75	84	72	99	71
56	74	88	65	63	79	82	75	78	86

The **frequency distribution** for these data using **5** equal classes is

Sales		50- 60	60 -70	70- 80	80-90	90-100	Total
Number of Months	A	2	3	8	5	2	20
	B	2	4	7	4	3	20
	C	2	3	9	4	2	20
	D	None of the above is correct					

Q12. The **mean** of wages is

(A) 18.24 (B) 17.64 (C) 16.85 (D) 17.75

Q13. What is the **variance** of wages?
(A) 35.3725 (B) 36.1875 (C) 38.1284 (D) 37.6825

Q14. Determine the **wage** below which **50%** of workers will be.
(A) 17.5 (B) 16.8 (C) 17.2 (D) 16.4

Q15. What is the **standard value** for the wage **18**?
(A) 0.12 (B) 0.05 (C) 0.04 (D) 0.08

Q16. Find the **number of workers** who have daily **wages** of **at least $12**.
(A) 12 (B) 14 (C) 18 (D) 16

Q17. What is the **coefficient of skewness** for a frequency distribution that has**: Mean = 10** , **Median = 8** , **Coefficient of variation = 25%**?
(A) 2.4 (B) 2.6 (C) 3.2 (D) 2.8

For the two variables **x** and **Y**, Given that
- The **two variables** have the **same variance** and the **same coefficient of variation**.
- The **correlation coefficient** between **x** and **Y** is equal to **0.5**.
- The **mean** of **Y** is **10**.
- The **predicted value** of **Y** when **x = 6** equals **10**.
Answer the following two questions (Q18-Q19):

Q18. The **regression coefficient** of the regression line of **Y** on **X** is
(A) 0.4 (B) 0.5 (C) 0.3 (D) 0.6

Q19. The **mean** of **X** is
(A) 6 (B) 8 (C) 12 (D) 10

Q20. If $y_i = \hat{y}_i$ for all values of i, then the two variables have

(A) Perfect positive relationship
(B) Perfect negative relationship
(C) Could have either a perfect positive or a perfect negative relationship.
(D) Weak relationship

Hint: Five points are assigned for each of the following 10 questions (Q21-Q30).

A cost accountant has derived the total cost (in thousands of dollars) against the output (in thousands of units) of a certain product over a period of 5 weeks, yielding the following data:

Output	4	2	5	8	6
Total Cost	25	20	30	40	35

Answer the following six questions (Q21- Q26):

Q21. The value of the **correlation coefficient (r)** between the two variables is

(A) 0.86 (B) 0.94 (C) 0.99 (D) 0.88

Q22. The value of the **rank correlation coefficient (r_s)** is

(A) 1 (B) 0.95 (C) 0.97 (D) 0.98

Q23. The value of the **coefficient of determination** is

(A) 0.74 (B) 0.88 (C) 0.98 (D) 0.77

Q24. The **regression coefficient** of the **regression line** of **cost on output** is

(A) 2.65 (B) 3.28 (C) 2.80 (D) 3.50

Q25. What is the **predicted cost** for producing **8** units?

(A) 39.8 (B) 40.5 (C) 42.4 (D) 38.6

Q26. Find the **standard error of estimate**.

(A) 1.29 (B) 1.32 (C) 2.16 (D) 2.24

Given are data for two variables x and y:

X:	16	15	12	k	11
Y:	24	22	16	12	14

If you know that the **two variables** are **perfectly related.**
Answer the following two Questions (Q27 and Q28):

Q27. Determine the **value** of **k**.

(A) 14 (B) 9 (C) 13 (D) 10

Q28. The value of the **rank correlation coefficient** (**r_s**) is

(A) 0.96 (B) -1 (C) 1 (D) Difficult to know

Q29. The owner of a store is interested to find out if there is any relationship between the time of a purchase is made and the amount of money spent on that purchase. The following table presents the number of purchases for size and time of each purchase.

Size of Purchase	**Time of Purchase**			**Total**
	Morning	**Evening**	**Night**	
$10-20	45	10	5	60
Over $20	5	10	25	40
Total	50	20	30	100

The value of the **contingency coefficient (Cramer's Coefficient)** is

(A) 0.66 (B) 0.72 (C) 0.68 (D) 0.75

Q30. The value of **the coefficient of association (Yule's coefficient)** for two attributes was found to be - **0.78**. Then, the **relationship** between these attributes is

(A) Moderate **(B) Strong**

(C) Moderate negative **(D) Strong negative**

South Valley University Faculty of Commerce	Date: 9/3/2021 Time Allowed: 3 hours

Principles of Statistics

English Teaching Section

You have 25 questions. Please mark all your answers on the answer sheet provided to you. You have to submit both questions papers and answer sheet. <u>Please</u>:

- Make sure that the answer sheet form matches the questions form.

- Choose the best answer for each of the following questions

<u>**Hint**</u>**:**<u>**Two points are assigned for each of the following 10 questions (Q1-Q10)**</u>**.**

Q1. Which of the following measures of central tendency can have more than one value in a single data set?

(A) Mean (B) Median (C) Mode (D) First quartile

Q2. If the mean of a frequency distribution is 100 and coefficient of variation is 45%, then standard deviation is

(A) 45 (B) 0.45 (C) 4.5 (D) 450

> The distribution of scores on a certain statistics test is strongly skewed to the left.
> Answer the following <u>two</u> questions (Q3-Q4):

Q3. Which set of measures of central tendency and dispersion are more appropriate for the distribution of scores?

(A) Mean and standard deviation
(B) Median and quartile deviation
(C) Mean and interquartile range
(D) Median and atandard deviation

Q4. What does this suggest about the difficulty of the test?

(A) It was an easy test (B) It was a hard test
(C) It wasn’t too hard or too easy (D) It is impossible to tell

Q5. Which one of the following statements is FALSE?

(A) The only way the standard deviation can be Zero is when all observations have the same value.
(B) If you interchange the independent variable and the dependent variable, the correlation coefficient remains the same.

(C) If the correlation coefficient between two variables is zero, that means that there is no possible relationship between the two variables.

(D) If the correlation coefficient r equals one, the value of r_s must equal one.

Q6. Of the following Z-score values, which one represents the location closest to the mean?

(A) +0.5 (B) +1 (C) -1.5 (D) -0.3

Q7. If Var (X) = 25 and Y = (2x + 5)/2, then the standard deviation of Y is equal to

(A) 7.5 (B) 5 (C) 25 (D) 10

Q8. The regression and correlation coefficients of the two variables will be the same if their ……. are same.

(A) Means (B) Standard deviations

(C) Variances (D) Either (B) or (C)

Q9. The elimination of extreme values at the top of the data set has the effect of

(A) No effect (B) Raising the mean

(C) Lowering the mean (D) Difficult to tell

Q10. How much variation is not explained by a correlation of o.9?

(A) 0.19 (B) 0.81 (C) 0.90 (D) None of these

Hint:<u>**Five points are assigned for each of the following 10 questions (Q11- Q20)**</u>.

Q11. If x and y are related by y = 2x + 5 and the standard deviation and mean of x are known to be 5 and 10 respectively, then the coefficient of variation of y is

(A) 30% (B) 40% (C) 20% (D) 25%

Q12. The marks obtained by 15 students on a statistics exam are given below.

54 65 40 56 74 69 73 58 63 80
67 76 62 89 79

The frequency distribution for these data using 5 equal classes is

Mark		40-50	50-60	60-70	70-80	80-90	Total
Number of Students	A	1	3	4	4	3	15
	B	1	2	6	3	3	15
	C	1	4	5	3	2	15
	D	1	3	5	4	2	15

Q13. If 5 is subtracted from each observation of some certain data set then its coefficient of variation is 10%, and if 5 is added to each observation then the coefficient of variation is 6%. Find the original coefficient of variation.

(A) 8% (B) 4% (C) 5.4% (D) 7.5%

The income distribution of the middle management in a large organization is tabulated below:

Income ($ hundreds)	20-	25-	30-	35-	40-45
Number of Managers	6	9	16	12	7

Answer the following five questions (Q14-Q17):

Q14. The mean of income is

(A) 31 (B) 34 (C) 33 (D) 32

Q15. What is the standard deviation of income?

(A) 6.02 (B) 6.08 (C) 6.04 (D) 7.03

Q16. Determine the value of the income so that 50% of managers will have greater than this value.

(A) 32.625 (B) 33.125 (C) 32.416 (D) 32.224

Q17. Find the number of managers who have income of less than 30.

(A) 18 (B) 15 (C) 31 (D) 20

Q18. What is the coefficient of skewness for a frequency distribution that has: First Quartile (Q_1) = 10 , Median = 18 , Third Quartile (Q_3) = 30?

(A) 0.10 (B) 0.25 (C) 0.20 (D) 0.28

Q19. A student obtained the mean and standard deviation of 100 observations as 40 and 5.1 respectively. It was later discovered that he had wrongly copied down an observation as 50 instead of 40. The correct standard deviation is

(A) 5 (B) 6 (C) 3 (D) 7

Q20. Following is the cumulative frequency distribution of the electricity cost (in dollars) of 20 two-bedroom apartments.

Electricity Cost	Less Then 40	Less than 50	Less than 60	Less than 70	Less then 80
Number of Apartments	3	7	13	17	20

If the lowest electricity cost is $30, what is the value of the mean?

(A) 62 (B) 65 (C) 52 (D) 55

Hint: Six points are assigned for each of the following 5 questions (Q21-Q25).

The following table presents advertising expenditures ($1000s) and revenue ($1000s) for 5 companies.

Company No.	1	2	3	4	5	Total
Advertising (x)	8	6	5	9	2	30
Revenue (y)	45	40	35	50	10	180

Computations provided the following:

$\sum xy = 1245$, $\sum x^2 = 210$, $\sum y^2 = 7450$

Answer the following four questions (Q21- Q24):

Q21. Find the value of the correlation coefficient (r) between the two variables.

(A) 0.94 (B) 0.97 (C) 0.96 (D) 0.95

Q22. Which of the following is the most likely value for the rank correlation coefficient (r_S)? Hint: You are not in need to make any calculations.

(A) 0.75 (B) 0.65 (C) 1 (D) 0.82

Q23. What is the error of predicting the revenue of the second company?

(A) 6 (B) 5 (C) 3 (D) 4

Q24. Find the standard error of the estimate.

(A) 4.6 (B) 3.8 (C) 3.6 (D) 4.8

Q25. Data on the social class and the number of children in a family were obtained as a part of national survey. The results from a sample of 200 families follow.

Number of Children	Social Class			Total
	Lower	Middle	Upper	
1 or 2	12	24	84	120
More than 2	8	16	56	80
Total	20	40	140	200

Based on the value of the Cramer's Coefficient (V), which one of the following statements is correct?

(A) The number of children in a family depends to a large extent on its social class.

(B) The relationship between the number of children in a family and its social class is neither strong nor weak.

(C) The number of children in a family depends to some extent on its social class.

(D) The number of children in a family does not depend at all on its social class.

South Valley University Faculty of Commerce	Date: 2/2/2022 Time Allowed: 3 hours
Principles of Statistics	
English Teaching Section	

You have 30 questions. Please mark all your answers on the answer sheet provided to you. You have to submit both questions papers and answer sheet.

<u>Please</u>:

- Make sure that the answer sheet form matches the questions form.

- Choose the best answer for each of the following questions

<u>Hint:</u> <u>Two points are assigned for each of the following 10 questions</u> <u>(Q1-Q10)</u>: (Time: 20 Minutes)

Q1. Which of the following measures can have more than one value for a set of data?

(A) Mode (B) Median (C) Mean (D) None of these

Q2. If the coefficient of determination is a positive value, then the regression equation ……

(A) Must have a positive slope (B) Must have a negative slope

(C) Could have either a positive or a negative slope

(D) Must have a positive y-intercept

> The mean and standard deviation of a set of numbers are 4 and 2 respectively. If each number is multiplied by 2 and then subtracted from 10.
>
> Answer the following <u>two</u> questions (Q3 - Q4):
>
> **Q3.** The mean of new numbers is ……
>
> (A) 5 (B) 2 (C) 4 (D) 6
>
> **Q4.** The standard deviation of new numbers is ……
>
> (A) 6 (B) 2 (C) 3 (D) 4

Q5. Let the correlation coefficient between X and Y be 0.6. Random variables W and Z are defined as Z = X + 4 and W = Y/2. What is the correlation coefficient between W and Z?

(A) 0.4 (B) 0.6 (C) 0.8 (D) None of these

Q6. If a test was generally very easy, except for a few students who had very low scores, then the distribution of scores would be ……

(A) Positively skewed (B) Negatively skewed
(C) Not skewed at all (D) Normal

Q7. The goal of ……. is to focus on using sample results to make decisions about the entire population from which the sample was drawn.

(A) Inferential statistics (B) Descriptive statistics
(C) None of the above (D) All of the above

Q8. If the percent of total variation of the dependent variable Y unexplained by the independent variable X is 36%. The correlation coefficient (r) between X and Y is ……

(A) 0.49 (B) 0.36 (C) 0.80 (D) 0.90

Q9. If the value of any regression coefficient is one, then the two variables are ……

(A) Independent (B) Moderately related
(C) Perfectly related (D) It is impossible to tell

Q10. The statistics score of a student is 2 standard deviations above the mean. If the mean and standard deviation of scores of all students are 70 and 5 respectively, what is the statistics score of this student?

(A) 80 (B) 82 (C) 60 (D) 86

Hint: Four points are assigned for each of the following 10 questions (Q11-Q20): (Time: 85 Minutes)

Q11. The following data show the time (rounded to the nearest day) required to complete year-end audits for a sample of 20 clients of a small accounting firm.

20	34	29	28	35	25	28	27	30	37
32	42	23	31	44	34	24	36	26	23

The frequency distribution for these data using 5 equal classes is ……

Time (day)		20-25	25-30	30-35	35-40	40-45	Total
Number of Clients	A	3	8	5	3	1	20
	B	4	7	4	3	2	20
	C	4	5	6	3	2	20
	D	None of the above is correct					

In a city, there are 25 houses for sale of similar size. The frequency of prices (in $1,000) is as follows:

Price ($1,000)	10 -	20 -	30 -	40 -	50 - 60	Total
Number of Houses	1	3	8	7	6	25

Answer the following five questions (Q12- Q16):

Q12. The mean of prices is ……

(A) 41.8 (B) 40.8 (C) 40.6 (D) 41.5

Q13. What is the variance of prices?

(A) 132.25 (B) 120.64 (C) 132.84 (D) 121.25

Q14. Find the number of houses that have prices of at least $35 thousands.

(A) 9 (B) 6 (C) 10 (D) 8

Q15. The value of the median is equal to ……

(A) 40.7 (B) 41.8 (C) 42.4 (D) 40.9

Q16. What does the coefficient of skewness tell you about this distribution?

(A) Moderately skewed to the left

(B) Approximately symmetric

(C) Moderately skewed to the right

(D) Highly skewed to the left

The following table gives the experience (in years) and the number of items which were rejected as unsatisfactory last week for 8 workers at a small factory.

Worker	A	B	C	D	E	F	G	H
Years of Experience (x)	8	5	10	3	7	2	9	12
Number of Rejects (y)	9	15	5	19	11	21	7	1

Computations provided the following:

$\sum x = 56$, $\sum y = 88$, $\sum xy = 448$, $\sum x^2 = 476$, $\sum y^2 = 1304$

Answer the following four questions (Q17- Q20):

Q17. **What is the value of the correlation coefficient (r) between the two variables?**

(A) – 0.985 (B) – 912 (C) – 1 (D) 1

Q18. **The value of the rank correlation coefficient (r_S) is ……**

(A) – 1 (B) – 0.95 (C) – 0.97 (D) 1

Q19. **The regression coefficient of the regression line of number of rejects on years of experience is ……**

(A) – 2.4 (B) – 3.0 (C) 2.4 (D) – 2.0

Q20. **Predict the number of rejects for a worker of 4 years of experience?**

(A) 15 (B) 16 (C) 17 (D) 18

Hint: Four points are assigned for each of the following 10 questions (Q21- Q30): (Time: 75 minutes)

For two variables X and Y, given:

$Y = 26 - 2X$, $\bar{X} = 5$, $S_X = 2$

Answer the following two questions (21–22):

Q21. **The coefficient of variation of Y is ……**

(A) 25% (B) 18% (C) 24% (D) 20%

Q22. **W hat is the correlation coefficient between the two variables?**

(A) 1 (B) -1 (C) 0 (D) Difficult to tell

Q23. For computing the value of Cramer's coefficient, if $O_{ij} = E_{ij}$ for all values of i andj(The observed and expected frequencies are equal for each cell of the contingency table). The relationship between the two attributes is ……

(A) Fairly strong (B) Moderate (C) Weak (D) No relationship

Q24. If $\hat{y}_i = y_i$ for all values of i, then the two variables have ……

(A) Perfect positive relationship

(B) Perfect negative relationship

(C) Could have either a perfect positive or a perfect negative relationship.

(D) Weak relationship

Given are data for two variables X and Y:

X:	4	2	k	3
Y:	2	6	10	4

If you know that the two variables are perfectly related.
Answer the following two Questions (Q25 - Q26):

Q25. Determine the value of k.

(A) 1 (B) 6 (C) 5 (D) 0

Q26. The value of the rank correlation coefficient (r_s) is ……

(A) 0.96 (B) -1 (C) 1 (D) Difficult to know

Data on the marital status of men and women aged 25 to 35 were obtained as a part of a national survey. The results from a sample of 120 men and 80 women follow.

Gender	Marital Status			Total
	Never Married	Married	Divorced	
Men	80	30	10	120
Women	10	50	20	80
Total	90	80	30	200

Answer the following two questions (Q27 - Q28):

Q27. **Based on the value of the Cramer's coefficient (V), which of the following statements is correct?**

(A) The marital status depends to a large extent on gender.

(B) The relationship between the marital status and gender is weak.

(C) The marital status moderately depends on gender.

(D) The marital status does not depend at all on gender.

Q28. **Repeating Q27 with excluding the category 'Divorced', the degree of association between the two attributes changes to be ……**

(A) Higher (B) Lower (C) Same (D) Difficult to tell

Given the following frequency distribution:

Age	2 - 4	4 - 6	6 - 8	8 – 10	10 and over
Number of Persons	1	2	5	K	1

The values of the mean and standard deviation of this distribution are 6.9 and 2.3, respectively.

Answer the following two questions (Q29 - Q30):

Q29. **The value of K is ……**

(A) 2 (B) 3 (C) 1 (D) 4

Q30. **The highest age is ……**

(A) 12 (B) 14 (C) 16 (D) 15

References

1. Anderson, R. A., Sweeney, D. J. and Williams, T. A. (2002). Statistics for Business and Economics (8th edition). South - Western Thomson Learning.

2.Erichson, B. H. and Nosanchuk. T. A. (1977). Understanding Data. McGraw-Hill.

3. Lind, D. A., Marchal, W. G. and Wathen, S.A. (2008). Basic Statistics for Business and Economics (13th edition). McGraw-Hill.

4. Mann, P. S. (2007). Introductory Statistics (2007). John Wiley.