BASIC BUSINESS STATISTICS

PREFACE

This book is the result of my working experience in the subject Basic Business Statistics for BBA and MBA students. It is designed to meet the requirements of Bachelors Masters and Ph D level students .The main highlight of the book is both theory instructor manual and solved problem approach & explanations added for numerical question problems framed by the author.This book has a large number of problems solved in all 11 chapters.

All theoretical concepts needed by students are covered by the author in this book.

I am extremely grateful to all my students who were attentive in the classes when I was conducting the classes.

SRINIVAS R RAO

EDUNXT CERTIFIED LEVEL 3 FACULTY FOR MBA

TRACKS INDIA INFOTECH LTD,UDUPI

SIKKIM MANIPAL UNIVERSITY,MANIPAL

ABOUT THE BOOK

This book is on Basic Business Statistics is a compulsory subject for Commerce students.The higher level students and bachelor level students can also read it as it contains a lot of numerical problems framed by me with full Instructor Manual.

Introduction to Statistics,Concepts of probability,Sampling methods and sampling distribution,Random variables and Probability Distribution,Descriptive Statistics,Inferential Statistics,F-Test and Analysis of Variance,Chi-square applications,Simple linear regression and correlation,Time Series Analaysis and Business Forecasting,Index Numbersare the 11 chapters with various sub-topics covered in this book.

I feel that this is a unique book as there are many theoretical Instructor Manual concepts and numerical problems solved with explanation.

HAPPY READING.

THANKS

REGARDS

AUTHOR

(SRINIVAS R RAO)

ABOUT THE AUTHOR

Author's name is Srinivas R Rao, born and done his school level education in Mangalore,Karnataka in a reputed private school Canara High School and PUC(+2) from Canara PUC in Science stream with PCMB as main subjects.

Later, pursuing LL.B(5 Years) course passed the degree in 1999 and done Diploma in Export Management ,Diploma in Customs and Central Excise , Diploma in Business Administration and some important IT subjects like MS-Office,Internet/Email,Visual Basic 6.0,C,C++,Java,Advanced Java,Oracle with D2K,HTML with Javascript,VBscript and Active Server Pages.

Joined as a FACULTY for students in a small computer Institute in 2002 July and later after 4 months worked in a company by name CRP Technologies(I) .P.Ltd as Branch Manager(Risk Manager) for Mangalore,Udupi and Kasargod areas from January 26 2003 to June 11 2007.In the year 2005 pursued MBA distance education course. Currently working as a FACULTY in Sikkim Manipal University , Udupi centre for BBA & MBA students and teaching numerical subjects like Statistics/Operations Research(Mgt Science/Quant. Techniques for Mgt)/Accounting and several numerical and difficult oriented subjects for distance education students in their weekend contact classes from July 2010 till present day.

Thanks

Regards

Author

(SRINIVAS R RAO)

BASIC BUSINESS STATISTICS

FOR BACHELORS,MASTERS AND PH.D-STATISTICS UNDER COMMERCE,ENGINEERING & MANAGEMENT SUBJECTS

CHAPTERS:

INSTRUCTOR'S MANUAL

CHAPTER 1

Introduction to Statistics

Learning Objectives

The study of this chapter should enable you to:

- ❖ Define the meaning of Statistics and other popular terms widely used in statistics
- ❖ Describe the types of statistics—descriptive and inferential
- ❖ Describe the sources of data, the types of data and variables
- ❖ Understand the different levels of measurement
- ❖ Describe the various methods of collecting data

Key Teaching Points

1.1 WHAT IS STATISTICS?

- 'Statistics' is a science that involves the efficient use of numerical data relating to groups of individuals (or trials).
- Related to the collection, analysis and interpretation of data, including data collection design in the form of surveys and experiments.
- Defined as the science of:
  - ¤ Collecting
  - ¤ Organizing
  - ¤ Presenting
  - ¤ Analyzing
  - ¤ Interpreting numerical data to efficiently help the process of making decisions
- A person who works with the applied statistics (the practical application of statistics), and is particularly eloquent in the way of thinking for the successful implementation of statistical analysis is called a 'statistician'.
- The essence of the profession is to measure, interpret and describe the world and patterns of human activity in it both in the private and public sectors.
- Those involved in marketing, accounting, quality control and others, such as consumers, sports players, administrators, educators, political parties, doctors, etc. on the other hand, tend to widely use the outcomes of various statistical techniques to help make decisions.
- Population size refers to a very large amount of data where making a census or a complete sampling of all of the population would be impractical or impossible.
- A sample is a subset of the population.
- Samples are collected and statistics are calculated from the samples in order to make conclusions about the population.

1.2 TYPES OF STATISTICS

- Two types of statistics:
  1. Descriptive statistics
  2. Inferential statistics
- Descriptive statistics explains the sample data whereas inferential statistics tries to reach conclusions that go beyond the existing data.

1.2.1 Descriptive Statistics

- Statistical methods used to describe the basic features of the data that have been collected in a study.
- Provide simple summaries about the data and the measures.
- Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.
- Use descriptive statistics simply to describe what's going on in our data.
- Used to present quantitative descriptions in a manageable form.
- Help us to facilitate large data in a way that easily makes sense.
- Each statistic reduces large data into a simple summary.

1.2.2 Inferential Statistics

- Methods used to find out something about a population based on a sample taken from that population.
- Also called statistical inference or inductive statistics.
- Most of the major inferential statistics come from a general family of statistical models known as the General Linear Model
  - Includes the *t*-test
  - Analysis of variance (ANOVA)
  - Regression analysis
  - Multivariate methods like factor analysis, multidimensional scaling, cluster analysis, discriminant function analysis, etc.

1.3 SOURCES OF DATA

- Two sources of data: 'primary data' and 'secondary data'.
- Researchers conduct various research projects using questionnaires addressed directly to respondents, and their responses are known as the primary data.
- Other studies involving the use of data collected by others, such as information from census and earlier findings are also used by researchers—called secondary data.
- Primary data offer information tailored to specific studies, but are usually more expensive and takes a longer period to process.
- Secondary data are usually less expensive to be acquired and can be analyzed in a shorter period.

1.3.1 Primary Data

- 'Primary data' is the specific information collected by person who is doing research (researcher).
- Researchers collect data through surveys, interviews, direct observations and experiments.
- Essential to all areas of study because it is the original data of an experiment or observation that has not been processed or altered.
- Primary data can be prospective, retrospective, interventional or observational in nature.
- Prospective data is collected from subjects in real time
- Retrospective data is collected from archival records.
- Retrospective primary data provides information on past circumstances or behaviours.

- Interventional primary data can be gathered after the interventions of interest have been prospectively delivered, manipulated or managed.
- Observational primary data is collected by monitoring an intervention of interest without intervening in the delivery of the intervention.

Advantages:

1. Researchers can decide the type of method they will use in collecting the data and how long it will take them to gather that particular data.
2. Researchers can focus the data collection on specific issues of their research and enable them to collect more accurate information.
3. Researchers would know in detail how the data were gathered and hence, will be able to present original and unbiased data.

Disadvantages:

1. Primary data collection consumes a lot of time, effort and cost; the researchers will not only need to make certain preparations, in addition, they will need to manage both their time and cost effectively
2. Researchers will have to collect large volumes of data since they will interact with different people and environments; also they will need to spend a lot of time checking, analyzing and evaluating their findings before using such data.

1.3.2 Secondary Data

- Any material that has been collected from published records, such as newspapers, journals, research papers and so on.
- Sources of secondary data may include information from the census, records of employees of a company, or government statistical information such as Malaysia gross national income (GNI) in different sectors and many others.
- Easily available and cheap.
- Available for a longer period of time.

Advantages:

1. Using data from secondary sources is more convenient as it requires less time, effort and cost.
2. Secondary data helps to decide what further researches need to be done.

Disadvantages:

1. Secondary data may have transcription errors (reproduction errors).
2. Data from secondary sources may not meet the user's specific needs.
3. Not all secondary data is readily available or inexpensive.
4. The accuracy of the secondary data can be questionable.

1.4 TYPES OF DATA AND VARIABLES

- 'Data' refers to qualitative or quantitative attributes of a variable or set of variables.
- A variable is defined as any measured attribute that varies for different subjects.
- Two basic types of data
  1. *Quantitative data*
  2. *Qualitative data*

1.4.1 Quantitative Data

- Data that measures or identifies based on a numerical scale.
- Can be analyzed using statistical methods
  - Values obtained can be illustrated using diagrams such as tables, graphs and histograms.

- Variable being studied can be reported numerically and is called a quantitative variable while the population is called a quantitative population.
- Quantitative variables can be further classified as either discrete or continuous.
- Discrete variables can assume only integer values (whole number such as 0, 1, 2, 3, 4, 5, 6, etc.).
- Discrete variables result from counting.
- Continuous variable can assume any value over a continuous range of possibilities.
  - For example:
    - ✓ Time (05:31:24 a.m)
    - ✓ Temperature (35.5 °C)
    - ✓ Weight (85.6 kilograms)
    - ✓ Height (167.5 cm)
    - ✓ Speed (183.7 km/h), etc.
- Continuous variables result from measuring something.

1.4.2 Qualitative Data

- Provide items in a variety of qualities or categories that may be 'informal' or even using features that is relatively obscure, such as warmth and taste.
- Although, the data that was originally collected as qualitative information, it can be quantitative if it is further simplified using the method of counting.
- Can include the obvious aspects such as gender, age or occupation.
- Can also be in the form of pass-fail, yes-no, or various other categories.
- If qualitative data uses categories based on ideas of subjective or non-existent, it is generally less valuable for scientific study than quantitative data.
- Sometimes it is possible to obtain quantitative estimates of the qualitative data.
  - For example:
    - ✓ People can be asked to rate their perceptions about their interest in a sport based on the Likert scale, that is, a rating or a psychometric scale commonly used in questionnaires.
    - ✓ If a 10-point scale is used, '1' would signify 'strongly agree' and '10' would indicate 'strongly disagree'.
- When the characteristics or *variable* being studied is non-numeric (categorical), it is called a *qualitative variable* or an *attribute*, while the population is called a qualitative population.
- When the data are qualitative, we are usually interested in:
  - How many?
  - What proportion fall in each category?
- Qualitative variables are measured according to their specific categories and are often summarized in charts.
  - For example:
    - ✓ Gender is measured as 'male' or 'female'.
    - ✓ Marital status is measured as 'single' or 'married', and so on.

1.5 LEVELS OF MEASUREMENT

- Can be classified into four categories:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

1.5.1 Nominal Level

- The most 'primitive', 'the lowest', or the most limited type of measurement.

- In this level of measurement,
  - ¤ Numbers or even words and letters are used to categorize the data.
- Suppose there are data about students who sat for an examination.
  - ¤ Hence, in a nominal level of measurement,
    - ✓ Students who passed the examination are classified as 'P'
    - ✓ Students who failed can be classified as 'F'

1.5.2 Ordinal Level

- Describes the relationship within a group of items in some specified order.
- For example,
  - ¤ For a student with the highest marks in a class—he will be placed as the first rank.
  - ¤ Then, a student who received the second highest marks will be placed as the second rank, and so on.
- This level of measurement indicates an approximate ordering of the measurements. The difference or the ratio between any two types of rankings is not always the same along the scale.

1.5.3 Interval Level

- Includes all the features of ordinal level (classification and direction).
- States that the distances between intervals are the same along the interval scale from low to high (constant size).
- A popular example of this level of measurement is temperature in Celsius.

1.5.4 Ratio Level

- Is the 'highest' level of measurement
- Has all the characteristics of interval level.
- Major differences between interval and ratio levels of measurement are:
  (1) Ratio-level data has a meaningful zero point
  (2) Ratio between any given two numbers is meaningful
- Divisions between the points on the scale have an equivalent distance between them
- Rankings assigned to the items are according to their size.
- Money is a good illustration,
  - ¤ Having zero ringgit means 'you have none'
- Weight is another ratio-level measurement.
  - ¤ If the dial on a scale is zero, there is a complete absence of weight.
  - ¤ If you earn $40 000 a year and Abu earns $10 000, you earn four times what he does.

1.6 METHODS OF COLLECTING DATA

- Data collection is an important aspect of any type of research study as inaccurate data collection can impact the results of a study and ultimately lead to invalid results.
- Investigator (researcher) must first of all, define the scope of his inquiry in every detail.
- The probable cost, time and labour required must next be estimated.
- If a complete coverage of information is not possible, for example, in market research, the sample size and method of sampling will have to be determined.
- Investigators collect primary data directly from the original sources.
- They can collect the necessary data appropriate for specific research needs, in the form they need.
- In most cases, primary data collection is costly and time-consuming.
- For some areas within social science research, such as socio-economic surveys, studies of social anthropology, market research, etc., necessary data are not always available from secondary sources, and they must be directly collected from the original or primary sources.

- In cases where the available secondary data are not suitable, again, the primary data should be collected.

1.6.1 Methods of Primary Data Collection

- 'Method' refers to a data collection mode or method
- 'Tool' is an instrument used to carry out the method.
- Some important methods of data collection:
  1. Observations
  2. Experimentation
  3. Simulation
  4. Interviewing
  5. Panel Method
  6. Mail Survey
  7. Projective Techniques
  8. Sociometry

1.6.2 Tools for Data Collection

- A number of different types of instruments or tools are used for data collection depending on the nature of the information to be gathered.

1. Types of Tools

✓ Observation schedule
✓ Interview guide and schedule
✓ Questionnaires
✓ Rating scale
✓ Checklists
✓ Data sheet
✓ Institution's schedule

2. Constructing Schedule and Questionnaire

✓ Schedule and questionnaire are the most common tools of data collection.
✓ These tools have many similarities and contain a set of questions related to the problem under study.
✓ Both these tools aim at retrieving information from the respondents.
✓ The content, structure, question words, question order, etc. are the same for all respondents.
✓ Each may use a different method; schedule is used for interviewing (the interviewer fills the schedule) and questionnaire is used for mailing (the respondents fill out questionnaires by themselves).
✓ Schedule and questionnaire are constructed almost in the same way.
✓ It consists of some main steps as below:
  (i) Identifying the research data
  (ii) Prepare 'dummy' tables
  (iii) Determine the level of the respondents
  (iv) Decide methods of data collection
  (v) Design instrument/tool
  (vi) Assessment of the design instrument
  (vii) Pre-testing
  (viii) Specification of procedures
  (ix) Planning format

3. Pilot Studies and Pre-Tests

- It is often difficult to design a large study without adequate knowledge of the problem; population to cover, level of knowledge, and so on.
  - What are the issues and the concepts related to the problem under study?
  - What is the best method of study?
  - How long will it take and what is the cost?
  - These and other related questions require a lot of knowledge about the subject matter.
- To obtain such pre-knowledge, a preliminary or pilot study should be conducted.
- Pilot study is a full-fledged miniature study of a problem
- Pre-test is a trial test of a specific aspect of the study such as method of data collection or instrument.
- Instrument of data collection is designed with reference to the data requirements of the study.
  - It cannot be perfected purely on the basis of a critical scrutiny by the designer and other researchers.
  - It should be empirically tested (should be tested using a collection of data). Hence, pre-testing of a draft instrument is rather indispensable.
- Pre-testing has several beneficial functions:
  - To test whether the instrument will get the responses needed to realize the objectives of the study.
  - To examine whether the content of the instrument is relevant and sufficient.
  - To test the questions whether the words are clear and in accordance with the understanding of the respondents.
  - To examine other qualitative aspects of the instrument such as the question structure and the sequence of questions.
  - To develop appropriate procedure to deal with the instrument in the field.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 1: INTRODUCTION TO STATISTICS
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 1.1 WHAT IS STATISTICS?
- **Slide 5** – 1.2 TYPES OF STATISTICS
- **Slide 6** – 1.3 SOURCES OF DATA
- **Slide 7** – *Advantages of Primary Data*
- **Slide 8** – *Disadvantages of Primary Data*
- **Slide 9** – *Advantages of Secondary Data*
- **Slide 10** – *Disadvantages of Secondary Data*
- **Slide 11** – 1.4 TYPES OF DATA AND VARIABLES
- **Slide 12** – 1.4 TYPES OF DATA AND VARIABLES (cont.)

- **Slide 13 –** 1.5 LEVELS OF MEASUREMENT
- **Slide 14 –** 1.6 METHODS OF COLLECTING DATA
- **Slide 15 –** 1.6.1 Methods of Primary Data Collection
- **Slide 16 –** 1.6.2 Tools for Data Collection
- **Slide 17 –** 1. Types of Tools
- **Slide 18 –** 2. Constructing Schedule and Questionnaire
- **Slide 19 –** Main Steps to Construct Schedule and Questionnaire
- **Slide 20 –** Four Crucial Decision Areas
- **Slide 21 –** 3. Pilot Studies and Pre-Tests

INSTRUCTOR'S MANUAL

CHAPTER

2 Concepts of Probability

Learning Objectives

The study of this chapter should enable you to:

- ❖ Define the meaning of probability and other key terms
- ❖ Understand the concepts of sample space and events
- ❖ Apply permutations and combinations rules to count sample points
- ❖ Calculate the probability of an event including conditional probability
- ❖ Define and apply additive rules, multiplicative rules, law of total probability and Bayes' rule

Key Teaching Points

2.1 INTRODUCTION

- Probability is a branch of mathematics that studies the possible outcomes of events with its possible likelihoods and relative distributions.
- The word 'probability' refers to the chance that a particular event (or series of events) will occur on a linear scale from 0 (impossibility) to 1 (certainty), or as a percentage (0 to 100%).
- Frequentists (classic approaches) see probability as a measure of the frequency of an event.
- Bayesians (evidential probabilities) considers probability as an estimating parameter for a set of observed distributions.

2.2 SAMPLE SPACE AND EVENTS

- A *sample space* is defined as a list of all possible outcomes of a random experiment or trial.
- An *event* is a set of outcomes or sample points.

2.2.1 Sample Space

- The word *experiment* is used by Statistician to describe any process that generates a set of data.
- A simple example of a statistical experiment is the tossing of a coin several times.
- There are only two possible outcomes, 'heads' or 'tails' and we are particularly interested in the uncertain observations every time the coin is tossed.
- The set of all possible outcomes of a statistical experiment is called the sample space and is usually represented by the symbol *S*.
- Each outcome in a sample space is called an element or a member of the sample space, or simply known as a *sample point*.
- The sample space *S*, of possible outcomes when a coin is tossed, may be written as $S = \{H, T\}$, where *H* and *T* correspond to 'heads' and 'tails', respectively.

2.2.2 Tree Diagram

- In some experiments it is helpful to list the elements of the sample space systematically through a tree diagram.
- The term 'tree diagram' refers to a graphic organizer used to list all possibilities of a sequence of events in a systematic way.

2.2.3 Events

- For any given experiment we may be interested in the occurrence of certain *events* rather than in the outcome of a specific element in the sample space.
- For instance, we may be interested in the event A, the outcome that is divisible by 3 when a dice is tossed.
  - This will occur if the outcome is an element of the subset $A = \{3, 6\}$.
  - To each event we assign a collection of sample points, which constitutes to a subset of the sample space.
  - Subset represents all of the elements for which the event is true while an *event* is a subset of a sample space.

2.3 COUNTING SAMPLE POINTS

- In many cases we shall be able to solve a probability problem by counting the number of sample points without actually listing each element.

2.3.1 Multiplication Rule

- The fundamental principle of counting is often referred to as the multiplication rule.

Theorem 2.3.1
If an operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, then the two operations can be performed in $n_1 n_2$ ways.

Theorem 2.3.2
Suppose that an operation can be performed in n_1 ways, and if for each one of these (from the first operation) a second operation can be performed in n_2 ways, and then for each one of these (from the second operation) a third operation can be performed in n_3 ways, and so on, then it can be shown that the sequence of k operations can be performed in $n_1 n_2, \ldots, n_k$ ways.

2.3.2 Permutations

- Frequently, we are interested in a sample space that contains elements with all possible orders or arrangements of a group of objects.
- The different arrangements are called *permutations*.
- Consider the letters a, b and c.
  - The six possible, distinct permutations are *abc*, *acb*, *bac*, *bca*, *cab* and *cba*.
  - Using Theorem 2.3.2, we could arrive at the answer 6 without actually listing the different orders.
  - n distinct objects can be arranged in $n(n-1)(n-2)\ldots(3)(2)(1) = n!$ ways.

Theorem 2.3.3
The number of permutations of n distinct objects is $n!$.
In general, n distinct objects taken r at a time can be arranged in

$$n(n-1)(n-2)\ldots(n-r+1) \text{ ways} = {}^{n}P_{r} = \frac{n!}{(n-r)!}.$$

Theorem 2.3.4

The number of permutations of n distinct objects arranged in a circle is $(n - 1)!$. So far we have considered permutations of distinct objects. Obviously, if the letters b and c are both equal to x, then the 6 permutations of the letters a, b, c become axx, axx, xax, xax, xxa and xxa, of which only 3 are distinct. Therefore, with 3 letters, 2 being the same, we have $3!/2! = 3$ distinct permutations.

Theorem 2.3.5

The number of different permutations of n objects of which n_1 objects of type 1, n_2 objects of type 2, ..., and n_k objects of type k is $\frac{n!}{n_1!\, n_2! \ldots n_k!}$.

Theorem 2.3.6

The number of ways of partitioning a set of n objects into r cells with n_1 elements in the first cell, n_2 elements in the second, and so forth, is

$$\binom{n}{n_1, n_1, \ldots, n_r} = \frac{n!}{n_1!\, n_2! \ldots n_r!}, \text{ where } n_1 + n_2 + \ldots + n_r = n.$$

2.3.3 Combinations

- In many problems we are interested in,
  - The number of ways of selecting r objects from n without considering its order.
- These selections are called *combinations*.
- A combination is actually a partition with two cells,
  - The one cell containing the r objects selected
  - The other cell containing the $(n - r)$ objects that are left
- The number of such combinations, denoted by,

  $\binom{n}{r, n-r}$, is shortened to, $\binom{n}{r}$

  since the number of elements in the second cell must be $n - r$.

Theorem 2.3.7

The number of combinations of n distinct objects taken r at a time is

$$\binom{n}{r} = \frac{n!}{r_1(n - r)}.$$

2.4 CALCULATING THE PROBABILITY OF AN EVENT

- The probability of an event is calculated as
  - The ratio of the outcomes (sample points) of the event divided by the total number of possible outcomes (all points in the sample space).

2.4.1 Probability of an Event

- In many experiments, such as tossing a coin or a dice,
  - All the sample points have the same chance of occurring and are assigned equal probabilities.
  - For points outside the sample space, that is, for simple events that cannot possibly occur, we assign a probability of zero.
- To find the probability of an event A, we sum all the probabilities assigned to the sample points in A.
  - The sum is called the probability of A and is denoted $P(A)$

    $0 \le P(A) \le 1$,

    $P(\varnothing) = 0$ and $P(S) = 1$

Theorem 2.4.1
If an experiment can result in any one of N different equally likely outcomes, and if exactly n of these outcomes correspond to event A, then the probability of event A is:
$P(A) = n/N$.

2.4.2 Additives Rule

- Several important laws that frequently simplify the computation of probabilities are as follows. The first, called the additive rules, applies to unions of events.

Theorem 2.4.2
If A and B are any two events, then
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$
If A and B are mutually exclusive, then
$$P(A \cup B) = P(A) + P(B)$$
If $A_1, A_2, A_3, \ldots, A_n$ are mutually exclusive, then
$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$
If $A_1, A_2, A_3, \ldots, A_n$ is a partition of a sample space S, then
$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n) = P(S) = 1.$$

Theorem 2.4.3
For three events A, B and C,
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - \mathrm{P}(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Theorem 2.4.4
If A and A' are complementary events, then:
$$P(A) + P(A') = 1.$$

2.4.3 Conditional Probability

- The probability of an event B occurring when it is known that some event A has occurred is called a conditional probability and is denoted by $P(B|A)$.
- The conditional probability of B, given the occurrence of A, denoted by $P(B|A)$, is defined by

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, \text{ if } P(A) > 0.$$

Independent Events

- $P(A|D)$ differs from $P(A)$.
- This suggests that the occurrence of A influenced D.
- However, consider the situation where we have events A and B and $P(A|B) = P(A)$;
  - In other words the occurrence of B had no impact on the odds of occurrence of A.
  - Here the occurrence of A is independent of the occurrence of B.
- Two events A and B are independent if and only if $P(B|A) = P(B)$ and $P(A|B) = P(A)$.
  - Otherwise, A and B are dependent.

2.4.4 Multiplicative Rules

- Multiplicative rules apply to intersections of events.

Theorem 2.4.5
If in an experiment the events A and B can both occur, then
$$P(A \cap B) = P(B \cap A) = P(A)P(B|A) = P(B)P(A|B).$$

Theorem 2.4.6

Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

Theorem 2.4.7

If, in an experiment, the events $A_1, A_2, A_3, \ldots, A_k$ can occur, then

$$P(A_1 \cap A_2 \cap \ldots \cap A_k)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3)\ldots. P(A_k|A_1 \cap A_2 \cap \ldots A_{k-1}).$$

If the events $A_1, A_2, A_3, \ldots, A_k$ are independent, then

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)P(A_2)P(A_3)P(A_4)\ldots. P(A_k).$$

2.4.5 Law of Total Probability

- Before we state and prove Bayes' rule (Section 2.4.6), it is important to state the law of total probability.
- The law of total probability is useful in proving Bayes' rule and in solving probability problems.
- In addition, this is a fundamental rule relating marginal probabilities to conditional probabilities.

Theorem 2.4.8

According to the Law of total probability, if the events $B_1, B_2, B_3, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

2.4.6 Bayes' Rule

- It is the result of probability theory.
- Relates the conditional and marginal probability distributions of random variables.
- Tells us how to revise beliefs in light of new evidence.
- The probability of occurrence of an event A conditional on the occurrence of another event B is different from the probability that B is conditional upon A
  - There is a definite relationship between these two, and Bayes' rule is the statement of that relationship.

Theorem 2.4.9

According to the *Bayes' Rule*, if the events $B_1, B_2, B_3, \ldots, B_k$ constitute a partition of the sample space S, where $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$,

$$P(B_i|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}, \text{ for } r = 1, 2, \ldots, k$$

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 2: CONCEPTS OF PROBABILITY
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 2.1 INTRODUCTION
- **Slide 5** – 2.2 SAMPLE SPACE AND EVENTS
- **Slide 6** – 2.2.1 Sample Space
- **Slide 7** – Examples
- **Slide 8** – 2.2.2 Tree Diagram
- **Slide 9** – 2.2.3 Events
- **Slide 10** – 2.3 COUNTING SAMPLE POINTS
- **Slide 11** – 2.3 COUNTING SAMPLE POINTS (cont.)
- **Slide 12** – 2.3.1 Multiplication Rule
- **Slide 13** – 2.3.2 Permutations
- **Slide 14–16** – 2.3.2 Permutations (cont.)
- **Slide 17** – 2.3.3 Combinations
- **Slide 18** – 2.3.3 Combinations (cont.)
- **Slide 19** – 2.4 CALCULATING THE PROBABILITY OF AN EVENT
- **Slide 20** – 2.4.1 Probability of an Event
- **Slide 21** – 2.4.2 Additive Rules
- **Slide 22** – 2.4.3 Conditional Probability
- **Slide 23** – Independent Events
- **Slide 24** – 2.4.4 Multiplicative Rules
- **Slide 25–27** – 2.4.4 Multiplicative Rules (cont.)
- **Slide 28** – 2.4.5 Law of Total Probability
- **Slide 29** – 2.4.5 Law of Total Probability (cont.)
- **Slide 30** – 2.4.6 Bayes' Rule
- **Slide 31** – 2.4.6 Bayes' Rule (cont.)

INSTRUCTOR'S MANUAL

CHAPTER 3 Sampling Methods and Sampling Distribution

Learning Objectives

The study of this chapter should enable you to:

- ❖ Distinguish various sampling methods (non-probability and probability)
- ❖ Define the bias in survey sampling
- ❖ Describe and apply the central limit theorem
- ❖ Define the sampling distribution of a sample statistic

Key Teaching Points

3.1 INTRODUCTION

- Sampling method refers to the way observations are selected from a population to be used as sample for a sample survey.
- Two types of sampling methods,
  - ¤ Non-probability sampling—does not involve random selection
  - ¤ Probability sampling—involves random selection
- Probability sample
  - ¤ A sample selected in such a way that each item or person in the population being studied has a known likelihood of being included in the sample.
- If probability sampling is done,
  - ¤ Each item in the population has a chance of being chosen.
- If non-probability methods are used,
  - ¤ Not all items have a chance of being included in the sample.
  - ¤ Results may be biased (the sample results may not be representative of the population).
- The sampling distribution,
  - ¤ Of the sample used to construct confidence intervals for the mean and for significance testing
  - ¤ With large samples leads to the *central limit theorem* (CLT).
    - ✓ CLT is one of the most remarkable results of the theory of probability as it justifies many procedures in applied statistics and quality control.

3.2 WHY SAMPLE THE POPULATION?

- It is often not feasible to study the entire population.
- Major reasons why sampling is necessary are:
  1. The destructive nature of certain tests.
  2. The physical impossibility of checking all items in the population.
  3. The cost of studying all the items in a population is often prohibitive.

4. The adequacy of sample results.
5. To contact the whole population would often be time-consuming.

3.3 SAMPLING METHODS: NON-PROBABILITY SAMPLING

- Non-probability sampling:
  - Does not use random selection
  - Does not rely on probability theory
- If a non-probability sample is used,
  - May be unable to represent a population well
  - Often find it difficult to assess how well we have done the sampling
- there are situations where probability sampling methods are not feasible or practical to employ
  - Consider various alternatives of non-probability sampling
  - Two types:
    - Accidental
    - Purposive—approach the problem of sampling with a specific purpose which has been planned in advance

3.3.1 Accidental, Haphazard or Convenience Sampling

- One of the most common methods of sampling goes under various titles.
- This includes the traditional 'man on the street' interviews conducted frequently by television news programs to get a quick reading of public opinion.
- The typical use of university students in much psychological research is primarily a matter of convenience.
- In the simplest research practices, we often use the respondents that are easily accessible.
- In the context of a more in-depth research, we usually collect samples using volunteers which are easily available.
- It is clear that in all these types of sampling, we do not have strong evidence to reflect that the samples may represent the populations being studied and hence contributes to doubting the samples in most cases.

3.3.2 Purposive Sampling

- Take samples with a purpose that had been specified in advance.
- Usually have set a specific group that to do the sampling.
- For example, you may have met with a group of interviewers who collect information by interviewing people passing by in front of them.
  - Most likely they're doing a purposive sampling and that may also be a market research.
  - They might be looking for respondents in a certain age group by means of assessing people passing by who fall into the category they are looking for, and then do the interviews.
  - The first thing they would do is to ensure that the respondents meet the established criteria.
- In situations where need to get samples quickly, and proportionality is not the main concern, purposive sampling methods can be very useful.
- Can easily obtain information from target population by using a purposive sample, but may also be biased towards certain groups of respondents because they are easily approachable.
- The methods listed below can be considered as subcategories of purposive sampling methods:
  1. Modal Instance Sampling
  2. Expert Sampling
  3. Quota Sampling
  4. Heterogeneity Sampling
  5. Snowball Sampling

- May do sampling for certain groups of people as in modal instance, expert or quota sampling.
- May also take samples for diversity as in heterogeneity sampling, or can take advantage of informal social groups to identify specific respondents who are hard to find as in snowball sampling.
- In these methods, we always know what we need—we perform sampling with a purpose.

3.4 SAMPLING METHODS: PROBABILITY SAMPLING

- Probability sampling or random sampling is a sampling technique in which the probability of getting any particular sample may be calculated.
- Any sampling method that utilizes some form of random selection is referred to as probability sampling.
- In establishing a method of random selection, we must establish some procedures to ensure that the different units in a population have equal probabilities to be selected.
- Various forms of random selection have been long practiced such as selecting numbers from a box in a lucky draw.
- As of today, with advanced technology, we tend to use computers as a mechanism to generate random numbers as the basis of random selection.
- For the various probability methods, here are some basic terms that must be defined:

  N = Number of cases in the sampling frame

  n = Number of cases in the sample

  ${}^{N}C_{n}$ = Number of combinations (subsets) of n from N

  $f = n/N$ = Sampling fraction
- It is always useful to remember that there is no one 'best' method of selecting a probability sample from a population of interest.
- All probability sampling methods have a similar goal, and that is to allow chance to determine the items to be included in the sample.

3.4.1 Simple Random Sampling

- A *simple random sample* is a sample formulated so that each item or person in the population has the same chance of being included.
- Suppose a population consists of 1 000 clients and then a sample of 100 clients needs to be selected.
- One way of ensuring that every employee has a chance of being chosen is to first write the name of each one on a small slip of paper and deposit all slips in a box.
- After they have been thoroughly mixed, the first selection is made.
- This process is repeated until the sample of size 100 is chosen.
- A more convenient method of selecting a random sample is to use the identification number of each client and a table of random numbers, for example Table A1 in Appendix (two-digit random numbers).
- For each number, the probability is the same and biased selection process can be eliminated.

3.4.2 Systematic Random sampling

- In a systematic random sample, the items or individuals of the population are arranged in some way (alphabetically) or by some other methods.
- A random starting point is selected, and then every kth member of the population is selected for the sample.
- The steps for a systematic random sample:

  (i) Number the units in the population from 1 to N

  (ii) Decide on the n (sample size) that we want

  (iii) $k = N/n$ = the interval size

(iv) Randomly select an integer between 1 to k
(v) Take every kth unit

3.4.3 Stratified Random Sampling

- Stratified random sampling method:
  - Divides a population into several homogeneous subgroups (called strata)i.e. non-overlapping subgroups N_1, N_2, ... N_i, where $N_1 + N_2 + ... + N_i = N$
  - Takes a simple random sample from each subgroup (stratum) by a fraction off = n/N
  - Referred to as proportional or quota random sampling
- After the population has been divided into strata, either a *proportional* or *non-proportional* sample can be selected.
- As the name implies, a proportional sampling procedure requires that the number of items in each stratum be in the same proportion as found in the population.
- In a non-proportional stratified sample,
  - The number of items studied in each stratum is disproportionate to the respective numbers in the population
  - Then weight the sample results according to the stratum's proportion of the total population.
- Regardless of whether a proportional or non-proportional sampling procedure is used, every item or person in the population has a chance of being selected for the sample.
- There are reasons why the stratified random sampling is usually preferred over simple random sampling.
  - Firstly, this is the only effective method that ensures that the samples taken are not only representing the whole population but also the main subgroups, especially the minorities.
    - ✓ If the sizes of some subgroups are very small, can use different sampling fractions (f) in different strata.
  - The second advantage is the stratified random sampling generally has more statistical precisions compared to simple random sampling, and this would be true only if the strata are homogeneous.
    - ✓ The variability within groups is lower than the variability for the population

3.4.4 Cluster (Area) random Sampling

- If we take a sample from a population that covers a wide geographic area using random sampling method, it is necessary to consider all parts of the area.
- *Cluster random sampling* is often employed to reduce the cost of sampling a population scattered over a large geographic area.
- Suppose we want to conduct a survey to determine the views of industrialists in a state with respect to environmental policies.
  - Selecting a random sample of industrialists in the state and personally contacting each one would be time-consuming and expensive.
  - Instead, we could employ cluster sampling by subdividing the state into small units, either districts or regions—often called *primary units*.
  - Suppose we divided the state into 12 primary units, then selected at random four regions, 2, 7, 4 and 12, and concentrated our efforts in these units.
  - We could take a random sample of the industrialists in each of these regions and interview them.
- The steps in cluster sampling are as follows:
  (i) Divide population into clusters (usually along geographic boundaries)
  (ii) Randomly sample clusters
  (iii) Measure all units within sampled clusters

3.4.5 Multi-Stage Sampling

- An approach that combines or incorporates a number of different sampling methods is called multi-stage sampling.

3.5 BIAS IN SURVEY SAMPLING

- The term 'bias' refers to the inclination of a sample statistic to be systematically different from a population parameter as a result of sampling procedure.

3.5.1 Bias Due to Unrepresentative Samples

- A good sample should be representative of the entire population.
- Each unit in the sample represents some of the elements of the population.
- Bias occurs when the sample cannot represent the population well—referred to as selection bias.
- Examples of selection bias are as follows:
  1. Undercoverage
  2. Non-response Bias
  3. Voluntary Response Bias

3.5.2 Bias Due to Measurement Error

- In any survey, the lack of appropriate measurement processes can lead to biased findings.
- A measurement process should take into account the environment in which the survey is being conducted, the structure of questions and type of respondents.
- Bias resulting from a measurement process is referred to as response bias.
- The following describes some response bias:
  1. Leading Questions
  2. Social Desirability

3.5.3 Sampling Error and Survey Bias

- A statistic calculated from a sample, called sample statistic, is used to estimate a population parameter.
- The sample is generated from a survey and if we repeat this survey several times, we will have a number of different samples and hence a number of different estimates of a statistic for the same population parameter.
- The average of all estimates calculated from all samples would be equal to the actual population parameter if the sample statistic is unbiased.
- This is true even though each of these estimates may differ from the population parameter and the variability among these estimates is called the sampling error.
- Sampling error can be reduced by increasing the sample size, i.e. a large sample size would reduce the variability of the sample statistic.
- However, survey bias cannot be reduced or eliminated by increasing the sample size.
- Survey bias is actually caused by the problems in sampling methodology (undercoverage, non-response bias, etc.) and this should be corrected first rather than the sample size.
- Large sample size cannot fix the problems in the methodology that lead to survey bias.

3.6 THE CENTRAL LIMIT THEOREM

- All tests of means are based on the Central Limit Theorem (CLT).
- This theorem provides a simple procedure to determine the mean, variance and shape of a distribution of sample means.
- All tests of hypotheses related to the means require the use of distributions of sample means.

- Therefore, the CLT should be clearly understood before the testing of hypotheses with means. The theorem could be expressed as follows:

  'When an infinite number of random samples is taken consecutively from a population, the distribution of the sample means calculated from each sample will be approximately normally distributed with mean, μ and standard deviation, $\sigma/\sqrt{N}$ ($\sim N(\mu, \sigma/\sqrt{N})$) as the sample size, N becomes larger regardless of the shape of distribution of the population.'
- The Central Limit Theorem consists of three main components—successive sampling from a population, increasing sample size and distribution of a population.
- Please note that this theory can only be employed for mean and no other statistics.

3.6.1 Successive Sampling

- To illustrate the successive sampling, we may use samples that successively drew from a uniform distribution.

3.6.2 Increasing Sample Size

- The second component of the CLT is the *sample size*.
- Generally, the sampling distribution of the mean becomes more normally distributed as sample size increases.

3.6.3 Population Distributions

- First, get samples from a population of a uniform distribution, then, calculate the mean for each sample and plot the values of the sample means.
- Although the actual distribution is completely flat, if we take an infinite number of successive random samples from the population, the distribution of the sample means becomes approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{N}$ ($\sim N(\mu, \sigma/\sqrt{N})$) as the sample size increases.
- This component of the central limit theorem means that the sampling distribution of the mean will be approximately normally distributed no matter what the actual shape of the population distribution.
- For instance, a Poisson distribution is often found when studying rare events where the shape of the distribution is positively skewed (skewed to the right).
- A normal distribution on the other hand reflects a variety of physical and psychological attributes, and the shape of the distribution is unimodal, symmetric and bell-shaped.

3.7 SAMPLING DISTRIBUTION

- The probability distribution of a statistic is called a *sampling distribution*.
- The probability distribution of $\bar{X}$ is called the *sampling distribution of the mean*.
- The sampling distribution of a statistic depends on the size of the population, the size of the samples and the method of choosing the samples.
- One should view the sampling distribution of $\bar{X}$ as the mechanism from which we will eventually use to make inferences on the parameter, μ.
- The sampling distribution of $\bar{X}$ with sample size, n is the distribution that results when an experiment is conducted over and over (always with sample size n) and produces many values of $\bar{X}$ result.
- This sampling distribution, then, describes the variability of sample averages, $\bar{x}$ around the population mean, μ.

3.7.1 The Sampling distribution of the Sample Mean

Theorem 3.7.1

The sampling distribution of $\bar{X}$ is the distribution of values of the sample mean over all possible samples of the same size from the same population.

From each sampling, we calculate and record the value of the sample mean, $\bar{X}$. Now we have an infinite number of $\bar{X}$'s, and then we calculate the mean and standard deviation of all $\bar{X}$'s. Next, we draw a histogram to check the shape of the distribution and it has already been made clear that we can use the three most important characteristics (mean, standard deviation and shape) to describe the sampling distribution of $\bar{X}$.

In practice, nobody is able or willing to take an infinite number of samples. Even so, the results can be obtained through the application of probability theory. For the sample mean, in general, the following results have been derived by theoretical statisticians.

Theorem 3.7.2

The Mean and Standard Deviation of $\bar{X}$:

For a random sample of size, n taken from a population with mean, μ and standard deviation,σ; the sample mean ($\bar{X}$), from an infinite size of population, will have a mean and a standard deviation of μ and $\sigma/\sqrt{n}$, respectively, or a standard deviation of $\sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$ if the population size is finite (size N).

Theorem 3.7.3

The Shape of the Distribution of $\bar{X}$*:*

The sample mean $\bar{X}$ will have a normal distribution regardless of n if the population is normal. However, if not normal but the sample size n is large, $\bar{X}$ will have an approximately normal distribution.

Note:

1. The mean of $\bar{X}$, denoted by $E(\bar{X})$ or $\mu_{\bar{X}}$, is known as the Expected Value of $\bar{X}$. The standard deviation of $\bar{X}$, denoted by $\sigma_{\bar{X}}$, is known as the *standard error* (SE).
2. As the sample size n approaches infinity, the distribution of $\bar{X}$ approaches normality. In practice, as agreed by most statisticians, the normal approximation is acceptable if $n = 30$.
3. From Theorem 3.7.2, the two expressions for the standard deviation of the sample mean $\bar{X}$ differ only by a factor $\sqrt{\frac{N-n}{N-1}}$ (N is the population size and n is the sample size). This is called the *Finite Population Correction Factor* (FPCF) and this factor is close to 1 if N is much larger than n, and hence it can be ignored. In practice, the FPCF can be ignored if n is less than 5% of N.
4. Population standard deviation, σ is often not known. If n is 'large', we can replace it with the sample standard deviation, s. Then the standard error (SE) can be estimated by $s/\sqrt{n}$ or $((s/\sqrt{n})(\sqrt{N-n})/(N-1))$

   With the knowledge of the sampling distribution of the sample mean, $\bar{X}$ we can now determine the probabilities of its possible values.

3.7.2 The Sampling distribution of the Sample Proportion

- There are also other sample statistics that have the sampling distributions such as median, standard deviation and proportion.

Theorem 3.7.4

The Central Limit Theorem for the Sample Proportion, $\bar{p}$:

A random sample of size n from a population with a sample proportion, $\bar{p}$ will have sample mean, $E(\bar{p}) = p$ and sample standard deviation, $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$ if the size of the population is infinite, or with standard deviation, $\sigma_{\bar{p}} = \sqrt{(N-n)/(N-1)}\ \sqrt{p(1-p)/n}$ if the size of the population is finite (N). If n is reasonably large, the sample proportion, $\bar{p}$ will have an approximately normal distribution.

This is an estimate or approximate distribution—as the sample size, n increases, and the accuracy of the estimation or approximation will also increase. In practice, determining how large the value of n to be acceptable for the approximation is different from what we used for the sampling distribution of the sample mean ($n \geq 30$). The value of n is considered to be large enough if both np and $n(1-p)$ are greater than or equal to 5.

From Theorem 3.7.4, the two expressions for the standard deviation of the sample proportion, $\bar{p}$ differ only by the FPCF, $\sqrt{(N-n)/(N-1)}$ (same as for the sample mean $\bar{X}$).

However, as before, this FPCF can be ignored if n is less than 5% of N.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 3: SAMPLING METHODS AND SAMPLING DISTRIBUTIONS
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 3.1 INTRODUCTION
- **Slide 5–6** – 3.1 INTRODUCTION (cont.)
- **Slide 7** – 3.2 WHY SAMPLE THE POPULATION?
- **Slide 8** – 3.2 WHY SAMPLE THE POPULATION? (cont.)
- **Slide 9** – 3.3 SAMPLING METHODS: NON-PROBABILITY SAMPLING
- **Slide 10** – 3.3 SAMPLING METHODS: NON-PROBABILITY SAMPLING (cont.)
- **Slide 11** – 3.3.1 Accidental, Haphazard or Convenience Sampling
- **Slide 12** – 3.3.1 Accidental, Haphazard or Convenience Sampling (cont.)
- **Slide 13** – 3.3.2 Purposive Sampling
- **Slide 14–16** – 3.3.2 Purposive Sampling (cont.)
- **Slide 17** – 3.4 SAMPLING METHODS: PROBABILITY SAMPLING
- **Slide 18** – 3.4 SAMPLING METHODS: PROBABILITY SAMPLING (cont.)
- **Slide 19** – 3.4.1 Simple Random Sampling
- **Slide 20** – 3.4.2 Systematic Random Sampling
- **Slide 21** – 3.4.3 Stratified Random Sampling
- **Slide 22** – 3.4.4 Cluster (Area) Random Sampling
- **Slide 23** – 3.4.5 Multi-Stage Sampling

- **Slide 24 –** 3.5 BIAS IN SURVEY SAMPLING
- **Slide 25 –** 3.5.1 Bias Due to Unrepresentative Samples
- **Slide 26 –** 3.5.1 Bias Due to Unrepresentative Samples (cont.)
- **Slide 27 –** 3.5.2 Bias Due to Measurement Error
- **Slide 28 –** 3.5.2 Bias Due to Measurement Error (cont.)
- **Slide 29 –** 3.5.3 Sampling Error and Survey Bias
- **Slide 30 –** 3.5.3 Sampling Error and Survey Bias (cont.)
- **Slide 31 –** 3.6 THE CENTRAL LIMIT THEOREM
- **Slide 32 –** 3.6 THE CENTRAL LIMIT THEOREM (cont.)
- **Slide 33 –** 3.6.2 Increasing Sample Size
- **Slide 34 –** 3.6.3 Population Distributions
- **Slide 35 –** 3.6.3 Population Distributions (cont.)
- **Slide 36 –** 3.7 SAMPLING DISTRIBUTION
- **Slide 37 –** 3.7 SAMPLING DISTRIBUTION (cont.)
- **Slide 38 –** 3.7.1 The Sampling Distribution of the Sample Mean
- **Slide 39–40 –** 3.7.1 The Sampling Distribution of the Sample Mean (cont.)
- **Slide 41 –** 3.7.2 The Sampling Distribution of the Sample Proportion

INSTRUCTOR'S MANUAL

CHAPTER

4 Random Variables and Probability Distribution

Learning Objectives:

The study of this chapter should enable you to:

- ❖ Define a random variable and a probability distribution
- ❖ Calculate the expected value of a random variable (discrete and continuous)
- ❖ Identify some specific probability distributions
- ❖ Calculate probabilities using normal approximation

Key Teaching Points

4.1 INTRODUCTION

- In an experiment, the outcomes are not necessarily in the form of numbers.
- For example, a coin is tossed and the outcome is either a 'head' or a 'tail'.
- For any experiment, a random variable is used to represent every outcome with a unique numerical value.
- As the experiment is repeated, the value of a random variable will vary from trial to trial.
- A probability distribution identifies:
  - ¤ The probability of each unique value of a random variable (discrete variable)
  - ¤ The probability of a value in a particular interval (continuous variable)
- It reflects the range of all possible values that a random variable can achieve and the probability that a value of a random variable is in any measurable subset from that range.
- The concept of probability distribution and random variables underlies the mathematical discipline of probability theory and the science of statistics.
- There is variability in almost any value that can be measured from a population (e.g. weight, strength, growth, sales, etc.).
- Almost all measurements have some intrinsic error.
- For these and many other reasons, simple numbers are often inadequate for describing a quantity, while probability distributions are often more appropriate.

4.2 NUMERICAL EVENTS AND RANDOM VARIABLES

- For a scientist, engineer, or businessman, the events of main interest are those represented by numbers, referred to as numerical events.
- Since the value of a numerical event will vary from trial to trial (repeated sampling), it is referred to as a random variable.
- For each point in the sample space S, a real number will be assigned to denote the value of a numerical event.

- For some points, the same number may be assigned.
- However, the numbers will vary from one point to another.
- Therefore, a random variable that is a function of the sample points in the sample space has to be defined. If we let Y denote this variable, then $Y = a$, is the numerical event that contains all sample points assigned as 'a'.
- The sample space S can be partitioned into a number of mutually exclusive subsets in such a way that a subset consists of only the points that are assigned the same value of Y.
- Therefore, a random variable can be defined as 'a real-valued function defined over a sample space'.

4.3 THE EXPECTED VALUE OF A RANDOM VARIABLE

- The expected value of a random variable, also known as its mean value, is the weighted average of all the values that a random variable can take.

4.3.1 The Expected Value for a Discrete Random Variable

- 'Discrete' refers to countable numbers—integers or whole numbers.
- Therefore, a discrete random variable is one that can only assume the values 0, 1, 2, 3, 4, 5, 6, etc.
- The expected value of a discrete random variable is defined as the weighted average of all its possible values where the weights are the respective probabilities of the variable.
- Let Y be a discrete random variable with probability function $p(y)$. Then the *expected value* of Y, $E(Y)$, is defined to be:

$$E(Y) = \sum_{y} y \cdot p(y).$$

If $p(y)$ is an accurate characterization of the population frequency distribution, then $E(Y) = \mu$, the population mean.

Theorem 4.3.1
Let c be a constant, then,

$$E(c) = c.$$

Theorem 4.3.2
Let $g(Y)$ be a function of the random variable Y and let c be a constant. Then

$$E\,[cg(Y)] = cE[g(Y)].$$

Theorem 4.3.3
Let $g_1(Y), g_2(Y), \ldots, g_k(Y)$ be k functions of the random variable Y. Then

$$E\,[g_1(Y) + g_2(Y) + \ldots + g_k(Y)] = E\,[g_1(Y)] + E\,[g_2(Y)] + \ldots + E\,[g_k(Y)].$$

Theorem 4.3.4
The variance of a random variable Y can be determined by,

$$V(Y) = \sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2.$$

4.3.2 The Expected Value for a Continuous Random Variable

- A *continuous* random variable is one which takes an infinite number of possible values (real values).
- Hence, the *expected value* of a continuous random variable is the probability density-weighted integral of all possible values.

- The *expected value* of a continuous random variable Y is,

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f(y)dy.$$

provided the integral exists.

Theorem 4.3.5
Let $g(Y)$ be a function of Y. Then, the *expected value* of $g(Y)$ is:

$$E[g(y)] = \int_{-\infty}^{\infty} g(y)f(y)dy;$$

provided the integral exists.

Theorem 4.3.6
Let c be a constant and let $f(Y)$, $f_1(Y)$, $f_2(Y)$, …,$f_k(Y)$ be the functions of a continuous random variable Y. Then the important results can be listed as follows:

1. Expected value of a constant; $E(c) = c$
2. $E[c \cdot f(Y)] = c \cdot E[f(Y)]$
3. $E[f_1(Y) + f_2(Y) + \ldots + f_k(Y)] = E[f_1(Y)] + E[f_2(Y)] + \ldots + E[f_k(Y)]$
4. $V(Y) = \sigma^2 = E[(Y - \mu)^2] = E(Y^2) - \mu^2$

4.4 SPECIFIC PROBABILITY DISTRIBUTIONS

- Some experimental situations naturally give rise to specific probability distributions.
- In most cases, the distributions used are simply models of the observed phenomena.
- There are three important and widely used probability distributions such as binomial, Poisson and normal.

4.4.1 The Binomial Distribution

- In an experiment of repeated trials, each trial has two possible outcomes; success or failure.
- An example of the most obvious application is on a production line, where each item produced would be defective or non-defective.
- The trials are independent and the probability of a success remains the same from trial to trial.
- This type of process is called the Bernoulli process and each trial is called a *Bernoulli trial*.

1 Binomial Probability Distribution

- Each Bernoulli trial produces a success with probability p and a failure with probability $q = 1 - p$. Thus, the probability distribution of a binomial random variable X, representing the number of successes in n independent trials, may be written as:

$$P(X = x) = b(x; n, p) = \binom{n}{x} p^x q^{n-x},\ x = 0, 1, 2, \ldots, n \text{ successes}$$

2 Mean and Variance of Binomial

- By simply using the parameters p and n, the mean and variance of a binomial random variable X can be determined as follows:

$$\mu = E(X) = np$$
$$\sigma^2 = V(X) = np(1 - p).$$

4.4.2 The Poisson Distribution

- A second important discrete distribution is the Poisson distribution.
- While a binomial random variable counts the number of successes that occur in a fixed number of trials, a *Poisson random variable* counts the number of rare events (successes) that occur in a specified time interval or a specified region.

1 The Poisson Process

- A Poisson process possesses the following properties:

(a) For any interval, the number of successes does not depend on the number of successes in other intervals.

(b) The probability of a success in one interval is directly proportional to the size of the interval, and is the same for all intervals with the same size.

(c) As an interval becomes smaller, the probability of two or more successes in the interval will approach to zero.

- The Poisson model thus is applicable when the events of interest occur *randomly*, *independently* of one another and *rarely*.

2 Poisson Probability Distribution

- The *Poisson random variable* indicates the number of successes that occur during a given time interval or in specified region in a Poisson experiment.

- If X is a Poisson random variable, the probability distribution of X is given by

$$P(X = x) = p(x; \mu) = \frac{e^{-\mu}\,\mu^{x}}{X!},\ x = 0, 1, 2, \ldots \qquad (4.4.8)$$

where μ is the average number of successes occurring in the given time interval or region and e = 2.71828… is the base of the natural logarithms.

- Notice that, since μ (the average number of successes occurring in a specified interval) appears in the formula, we must obtain an estimate of μ (usually from historical data) before we can apply the Poisson distribution.
- The number of values of a Poisson random variable is infinite (no limit).
- Unlike the Binomial random variable, the Poisson random variable is discrete with infinitely many values.

3 Mean and Variance of Poisson

- If X is a Poisson random variable for which μ is the average number of successes that occur in a specified interval, the expected value (mean) and the variance of X have the same value.

$$E(X) = V(X) = \mu.$$

4 Poisson Approximation to Binomial Distribution

- Although Binomial and Poisson random variables have distinct distributions, the two distributions are related.
- If we imagine a Poisson random variable whose interval has been subdivided into n (where n is large) very small subintervals, the probability of a success in any subinterval is approximately $p = \mu/n$, and so we have an approximate Binomial random variable.
- Similarly, a Binomial distribution for which the number of trials n is large and the probability p of a success is very small can be approximated by a Poisson distribution.
- This approximation is useful because for large values of n, Binomial probability tables are often unavailable.
- The appropriate Poisson distribution that will be used for the approximation will have $\mu = np$, the mean for the Binomial distribution. In order for the approximation to be good one, p should be very small.
- Thus, it is conventional to suggest that at the least, we should have $p < 0.05$.

5 Comparison of Binomial and Poisson Probabilities

- If we compare the binomial and Poisson probabilities with mean $\mu = 1$, as shown in the table below, the values are not much different.

- Binomial and Poisson probabilities with mean $\mu = 1$.

x	Binomial probability ($n = 50$, $p = 0.02$)	Poisson probability ($\mu = np = 1$)
0	0.364	0.368
1	0.372	0.368
2	0.186	0.184
3	0.061	0.061
4	0.014	0.015
5	0.003	0.003
6	0.000	0.001

4.4.3 The Normal Distribution

- The normal distribution is symmetrical and bell-shaped curve.
- It is the most important continuous distribution.
- The normal distribution is important because of two reasons.
  - First, the normal distribution is considered to be the basis distribution of statistical inference, representing the distribution of possible estimates (from different samples) of a population parameter.
  - Second, the normal distribution gives a useful approximation for other distributions including discrete distributions such as Binomial.

1 Normal Probability Distribution

- A random variable X with mean, μ and variance, σ^2 is normally distributed if its probability density function is given by:

$$N(x; \mu, \sigma) = f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{-(1/2)\left[\frac{x-\mu}{\sigma}\right]^2}, -\infty < x < \infty,$$

where $\pi = 3.14159\ldots$ and $e = 2.71828\ldots$

- A random variable that is normally distributed is called a *normal random variable* and can take on any real value from $-\infty$ to $+\infty$.
- The normal probability density function $f(x)$ is also continuous and has a positive value for all values of x.
- However, the value of $f(x)$ does not represent the probability that the variable X is equal to x, instead it is an expression of the height of the curve at $X = x$.
- In addition, the total area under the curve $f(x)$ must equal to 1.
- It is clear that from the formula for the probability density function, a normal distribution is fully determined by the two parameters, μ and σ^2.
- That is, a whole family of different normal distributions exists, but one differs from another only in the location of its mean μ and in the variance σ^2 of its values, but the fact remains that all normal distributions have the same symmetrical and bell-shaped appearance.

2 Standard Normal Distribution

- After we determine that a situation can be modeled appropriately by using a normal distribution.
- The procedure for finding normal probabilities becomes crucial.

- The actual calculation of such an area (probability) is difficult, and so we can resort to the tabulated area provided by the *Standard Normal Probability Table*; given in the Appendix section (Table A4).
- Because each pair values for the parameter μ and σ^2 gives rise to a different normal distribution, there are infinitely many possible normal distributions, making it impossible to provide a table of areas for each one.
- Nevertheless, we can make do with just one table.
- The particular normal distribution for which Standard Normal Probability Table (Table A4) has been constructed is the normal distribution with $\mu = 0$ and $\sigma = 1$, called the *standard normal distribution*; $N(z; 0,1)$.
- The corresponding normal random variable, with a mean of 0 and a standard deviation of 1, is called the *standard normal random variable* and is denoted as Z.
- Thus, before using Standard Normal Probability Table, we must convert or transform our normal random variable, X into the standard normal variable, Z.
- Standard Normal Random Variable, $Z = \dfrac{X - \mu_x}{\sigma_x}$.
- The z-value corresponding to a given value x_0 has an important interpretation because $(x_0 - \mu)$ expresses how far x_0 is from the mean while the corresponding z-value, $z_0 = (x_0 - \mu)/\sigma$ tells us how many standard deviations, x_0 is from the mean.
- As we have just seen, we can obtain the desired probabilities for any normal distribution from probabilities tabulated for the standard normal distribution.

3 Normal Approximation to Binomial

- To approximate other probability distributions including the Binomial, the normal distribution can be used.
- For a Binomial distribution with a large number of trials ($n > 25$), the Binomial tables (Table A2) cannot be employed, instead the normal approximation to the Binomial becomes useful.
- Since the normal distribution is symmetrical, it would provide the best approximation when the Binomial is also symmetrical.
- Also, a Binomial distribution is symmetrical when p (the probability of a success) is equal to 0.5; hence the best approximation is when p is close to 0.5.
- The greater the difference between p and 0.5, the larger the number of trials n is needed for a good approximation.
- The normal approximation to the binomial distribution works best when only a very small probability exists, which results to the approximating normal random variable to assume a value that falls outside the binomial range ($0 \le X \le n$).
- Generally, this is satisfied if $np \ge 5$ and $n(1 - p) \ge 5$, so a conventional rule of thumb is that the normal distribution will provide an adequate approximation of a binomial distribution if $np \ge 5$ and $n(1 - p) \ge 5$.
- Recall that the Poisson distribution can be used to approximate binomial probabilities if p is small, say $p < 0.05$.
- Given a binomial distribution with n trials and probability p of a success on any trial, the mean and variance of the binomial distribution are:

$$\mu = np$$
$$\sigma^2 = np(1 - p)$$

- We therefore choose the normal distribution with $\mu = np$ and $\sigma^2 = np(1 - p)$ to be the approximating Binomial distribution.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2 –** CHAPTER 4: RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS
- **Slide 3 –** LEARNING OBJECTIVES
- **Slide 4 –** 4.1 INTRODUCTION
- **Slide 5 –** 4.1 INTRODUCTION (cont.)
- **Slide 6 –** 4.2 NUMERICAL EVENTS AND RANDOM VARIABLES
- **Slide 7 –** 4.2 NUMERICAL EVENTS AND RANDOM VARIABLES (cont.)
- **Slide 8 –** 4.3 THE EXPECTED VALUE OF A RANDOM VARIABLE
- **Slide 9 –** 4.3.1 The Expected Value for a Discrete Random Variable
- **Slide 10–12 –** 4.3.1 The Expected Value for a Discrete Random Variable (cont.)
- **Slide 13 –** 4.3.2 The Expected Value for a Continuous Random Variable
- **Slide 14 –** 4.3.2 The Expected Value for a Continuous Random Variable (cont.)
- **Slide 15 –** 4.4 SPECIFIC PROBABILITY DISTRIBUTIONS
- **Slide 16 –** 4.4.1 The Binomial Distribution
- **Slide 17–18 –** 4.4.1 The Binomial Distribution (cont.)
- **Slide 19 –** 4.4.3 The Poisson Distribution
- **Slide 20–25 –** 4.4.3 The Poisson Distribution (cont.)
- **Slide 26 –** 4.4.3 The Normal Distribution
- **Slide 27–31 –** 4.4.3 The Normal Distribution (cont.)

INSTRUCTOR'S MANUAL

CHAPTER 5

Descriptive Statistics: Describing, Exploring and Comparing Data

Learning Objectives

The study of this chapter should enable you to:

- ❖ Organize raw data into a frequency distribution
- ❖ Present a frequency distribution into graphic forms
- ❖ Describe and calculate different measures of central tendency
- ❖ Define and calculate different measures of dispersion and skewness

Key Teaching Points

5.1 INTRODUCTION

- Descriptive statistics aims to summarize quantitative data without using the probabilistic formulation and not to draw conclusions about the population.
- Although a data analysis uses inferential statistics to deduce the main conclusions, the features of descriptive statistics are usually highlighted at the same time.
- The computer has become an important tool in the presentation and analysis of data. Among the most popular and widely used are SAS, SPSS, Statgraphics and Minitab.
- The collected samples are referred to as raw data. This chapter discusses how to organize raw data into a frequency distribution and present it using graphic forms.
- The data are further explored using measures of central tendency and then compared using measures of dispersion and skewness.

5.2 FREQUENCY DISTRIBUTION

- A frequency distribution is a tabulation of values that contains one or more variables.
- It summarizes the distribution of values in the sample where each entry represents the frequency or count of the occurrences of values within a particular interval.
- It is much simpler to manage the frequency tabulated data than the raw data. With frequency tables, there are simple formulas to calculate the important statistics.

5.2.1 Frequency Distribution for Qualitative Data

- A frequency distribution exhibits how the frequencies are distributed over categories.
- A frequency distribution for qualitative data lists all categories and the number of elements that belong to each of the categories.

5.2.2 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 this form is called *grouped data*.
- An interval that includes all the values that fall within two numbers, the lower and upper limits, is called a *class*. The classes are non-overlapping.

5.2.3 Class Intervals and Midpoints (Quantitative Data)

- The *midpoint*, or *class mark*, is determined by going halfway between the lower and the upper class limits. It can be computed by adding these two limits and dividing the total by 2.
- The *class interval* for a frequency distribution can be determined by subtracting the lower limit of a class from the lower limit of the next higher class.

5.2.4 Suggestions on Constructing a Frequency Distribution (Quantitative Data)

- Use equal-size class intervals.
- Find the suggested class interval:
  - $\textit{Suggested class interval} = \dfrac{\textit{Highest value} - \textit{Lowest value}}{\textit{Number of classes}}$
  - $\textit{Suggested class interval} = \dfrac{\textit{Highest value} - \textit{Lowest value}}{1 + 3.322\ (\textit{logarithm of total freq.})}$
- Choose appropriate number of classes.
  - No fewer than 5 or more than 15 classes
  - Use the smallest integer k such that $2^k \leq n$; n is the no. of observations.

5.2.5 Relative Frequency Distribution

- It may be desirable to convert class frequencies to relative class frequencies to show the percentage of the total number of observations in each class.
- Each of the class frequencies is divided by the total number of frequencies.

5.2.6 Cumulative Frequency Distribution

- The *cumulative frequency* for each class is determined by summing the frequencies for the class and all prior classes.

5.3 GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION

- Most publications emphasize the importance of graphs or *charts* as they give the readers a quick view of the important facets of the statistical data.
- We will concentrate on four graphic forms: a *stem-and-leaf display*, a *histogram*, a *frequency polygon* and a *cumulative frequency polygon* (*ogive*).

5.3.1 Stem-and-Leaf Display

- The stem-and-leaf display technique is commonly used as it offsets the loss of information that occurs from summarizing raw data.
- Each value is divided into two portions; the stem is the leading digit and the leaf is the trailing digit. Stem is placed to the left of a vertical line and leaf is to the right.
- An advantage of a stem-and-leaf display over a frequency distribution is that we do not lose information on individual observations.
- A stem-and-leaf display is constructed only for quantitative data.

5.3.2 Histogram

- One of the most widely used charts and one of the easiest to understand.
- It describes a frequency distribution in terms of a series of adjacent bars, each used to represent the number of class frequencies in a particular class.
- The class frequencies are scaled on the vertical axis and either the class limits or midpoints are scaled on the horizontal axis.
- A bar is drawn for each class so that its height represents the frequency of that class.

5.3.3 The Frequency Polygon

- A *frequency polygon* is a graph formed by joining the midpoints of the tops of successive bars in histogram with straight lines.
- It consists of line segments connecting the points formed by the intersection of the class midpoints (X-axis) and the class frequency (Y-axis).

5.3.4 Cumulative Frequency Polygon (Ogive)

- It is sometime useful to determine the number of observations that fall above or below a certain value.
- This can be accomplished using the *cumulative frequency polygon* or *curve*, which is also known as *ogive*.
- *Less-than ogive* allows us to determine how many (or percentage) of the observations are equal to or less than a certain value.
- *More-than ogive* allows us to determine how many (or percentage) of the observations are equal to or more than a selected amount.

5.3.5 Other Graphic Representations of Data

- Line chart
  - Line chart is ideal for portraying the trend of data over a period of time.
  - It is constructed by connecting a series of data using straight line segments.
- Bar chart
  - A graph made of bars with heights that represent the frequencies of respective categories; horizontal bar chart, vertical bar chart and two-directional bar chart.
  - A bar chart can be used to depict any of the levels of measurement.
- Pie chart
  - A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories.
  - To construct a pie chart, we multiply 360 by the relative frequency for each category to obtain the size of the angle for the corresponding category.
  - A pie chart is useful for depicting a relative frequency distribution.

5.4 MEASURES OF CENTRAL TENDENCY

- A *measure of central tendency* is a single value that represents a set of data.
- It pinpoints the centre of the values and is commonly referred as an *average*.
- We will consider *four* measures of central tendency; arithmetic mean, median, mode and geometric mean.
- Any measurable characteristic of a *population*, such as the mean, is called a *parameter*; any measurable characteristic of a *sample* is called a *statistic*.

5.4.1 The Arithmetic Mean

- The arithmetic mean is a widely used measure of central tendency.

- Properties of Mean:
  - Every set of interval-level and ratio-level data has a mean.
  - All the values are included in computing the mean.
  - A set of data has only one mean (*unique*).
  - The mean is a useful measure for comparing two or more populations.
  - The arithmetic mean is the only measure of central tendency where the *sum of the deviations of each value from the mean will always be zero.*
- Mean for Ungrouped Data:

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

  - The mean of a sample, or any measure based on sample, is called a *statistic*.
- Mean for Grouped Data (Weighted Mean):
  - The mean of a sample organized in a *frequency distribution* is computed by:

$$\bar{x} = \sum_{i=1}^{k} f_i x_i / n$$

5.4.2 The Median

- For data containing one or two very large or very small values, the centre point can be better described using *median*.
- Properties of the Median:
  - Unique — there is only one median for a set of data.
  - It is determined by arranging the data from low to high, and find the middle value.
  - It is not affected by extremely large or small values.
  - It can be computed for an open-ended frequency distribution if the median does not lie in the open-ended class.
  - It can be computed for ratio-level, interval-level and ordinal-level data.
- Median for Ungrouped Data:
  - The median for ungrouped data is the *midpoint* of values after they have been arranged from the smallest to the largest, or the largest to the smallest.
  - There are as many values above the median as below it in the data array.
  - $Median = \begin{cases} Middle\ value,\ if\ n\ is\ odd; \\ Mean\ of\ the\ two\ middle\ values,\ if\ n\ is\ even. \end{cases}$
- Median for Grouped Data:
  - The median for the grouped data can be estimated by locating its class and then interpolating within that class to arrive at the median.

$$Median = \bar{x} = L_m + \left(\frac{\frac{N}{2} - \sum f_{m-1}}{f_m} \right) C_m$$

5.4.3 The Mode

- The mode of a set of measurements is the value that occurs most frequently.
- It is useful in describing nominal and ordinal levels of measurement.
- Mode for Ungrouped Data:
  - The mode has the advantage of not being affected by extreme values.
  - It can be used as a measure of central tendency for open-ended distributions.
  - Many data sets have no modes because no value appears more than once.
  - Some data sets have more than one mode; a data set with two modes is referred to as *bimodal*.

- Mode for Grouped Data:
  - The mode can be approximated by the midpoint of the class with the largest class frequency.
  - $Mode = \hat{x} = L_{\text{mode}} + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] C_{\text{mode}}$

5.4.4 The Geometric Mean

- The geometric mean is not so highly influenced by extreme values as is the arithmetic mean. It has two main uses:
  - To average percentages, indexes and relatives.
  - To determine the average percent increase in sales, etc.
- The geometric mean of a set of n positive numbers is defined as the nth root of the product of the n numbers. If one of the numbers is non-positive, the *G.M. cannot* be computed.
  - *Geometric Mean* = G.M. = $\sqrt[n]{(X_1)(X_2) \ldots (X_n)}$

5.4.5 Selecting an Average for Data in a Frequency Distribution

- For a symmetric distribution, the *three averages* (mean, median and mode) are located at the centre and are always equal.
- As the distribution becomes *asymmetrical*, or *skewed*, the relationship among the three averages changes.
- In a *positively skewed*, the mean is the largest of the three averages. The median is generally the next largest average and the mode is the smallest.
- In a *negatively skewed*, the mean is the lowest of the three averages. The median is greater than the mean and the modal value is the largest.
- If the distribution were highly skewed, the mean would not be a good average.
- An approximate relationship among the three averages:
  - If there is sufficiently large number of observations to suggest a smooth distribution and if the shape of the curve is only moderately skewed, *the median is approximately one third of the distance from the mean to the mode*.
  - If two averages of a moderately skewed frequency distribution are known, the third can be approximated.
    Mode = Mean – 3(Mean-Median)
    Mean = [3(Median) – Mode] / 2
    Median = [2(Mean) + Mode] / 3.

5.5 MEASURES OF DISPERSION AND SKEWNESS

- Several measures that describe the *dispersion*, *variability*, or *spread* of the data; range, mean deviation, variance, quartiles and percentiles.
- *Skewness* is a measure of the asymmetry of the probability distribution of a random variable. Skewness value can be positive, negative or undefined.

5.5.1 Why Study Dispersion?

- A *small* value for a measure of dispersion (spread) indicates that the data are clustered closely around the mean. The mean is therefore considered quite representative of the data (a reliable average).
- A *large* value for a measure of dispersion indicates that the mean is not very reliable.

5.5.2 Measures of Dispersion—Ungrouped Data

- The Range:
  - Range is the simplest measure of dispersion.

- It is the difference between the highest and lowest values in a set of data.
- *Range = Highest value – Lowest value.*

- Mean Deviation:
  - A serious defect of the range is that it is based only on two values.
  - The *mean deviation* measures the mean amount by which the values in a population, or sample, vary from their mean (mean absolute differences).
  - *Mean Deviation = M.D.* $= \dfrac{\sum_{i=1}^{n} |X_1 - \bar{X}|}{n}$.

- Variance and Standard Deviation:
  - Variance and standard deviation are also based on the deviations from the mean.
  - *Variance* is the arithmetic mean of the squared deviations from the mean while *standard deviation* is the positive square root of the variance.
  - The sample variance for ungrouped data: $s^2 = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \dfrac{\sum_{i=1}^{n} X_i^2 - \dfrac{\left[\sum_{i=1}^{n} X_i\right]^2}{n}}{n-1}$.

5.5.3 Measures of Dispersion—Grouped Data

- Range for Grouped Data:
  - *Range = Highest limit of the largest class - Lowest limit of the smallest class.*
- Mean Deviation for Grouped Data:
  - *Mean Deviation = M.D.* $= \dfrac{1}{n}\left(\sum_{i=1}^{k} |x_1 - \bar{x}| f_i\right)$
- Standard Deviation for Grouped Data:
  - $s = \sqrt{\dfrac{1}{n-1}\left(\sum_{i=1}^{k} f_i (x_i - \bar{x})^2\right)} = \sqrt{\dfrac{1}{n-1}\left(\sum_{i=1}^{k} f_i x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{k} f_i x_i\right)^2\right)}$.

5.5.4 Interpretation and Uses of the Standard Deviation

- The standard deviation is commonly used as a measure to compare the spread in two or more sets of observations.
- A large value of standard deviation indicates that the values of the data are widely dispersed from the mean.
- Conversely, for a small value of standard deviation, the values are spread close to the mean. This is clearly stated by Chebyshev's theorem and the empirical rule.
- Chebyshev's theorem:
  - For a set of values, regardless of the shape of the distribution, the proportion of the values within k standard deviations of the mean is at least $1\text{-}1/k^2$, $k>1$.
- The Empirical Rule: For a set of values with symmetrical and bell-shaped frequency distribution, the following can be concluded:
  - Approximately 68% of the values are within one standard deviation, of the mean;
  - About 95% of the values are within two standard deviation of the mean;
  - Practically 99.7% of the values are within three standard deviation of the mean.

5.5.5 Some Other Measures of Dispersion

- Interquartile Range:
  - The distance between the third quartile and the first quartile. The first and third quartiles, Q_1 and Q_3, respectively, locate the point below which 25% and 75% of the observations are located.

- *Interquartile range* = $Q_3 - Q_1$
- $Q_1 = L_{Q_1} + \dfrac{\left(\sum_{i=1}^{k} f_i / 4\right) - \sum f_{Q_1 - 1}}{f_{Q_1}} \cdot C_{Q_1}$; $Q_3 = L_{Q_3} \dfrac{3\sum_{i=1}^{k} f_i / 4 - \sum f_{Q_3 - 1}}{f_{Q_3}} \cdot C_{Q_3}$

- Quartile Deviation:
  - *Quartile deviation* is half the distance between the third and the first quartiles.
  - $QD = (Q_3 - Q_1)/_2$
- Percentile Range:
  - The 99 percentiles divide a distribution into 100 parts. The 10-to-90-percentile range is the distance between the 10th and 90th percentiles.
  - *Percentile range* = $P_{90} - P_{10}$
  - $P_{10} = L_{P_{10}} + \dfrac{\frac{10}{100} n - \sum f_{P_{10} - 1}}{f_{P_{10}}} \cdot C_{P_{10}}$; $P_{90} = L_{P_{90}} + \dfrac{\frac{90}{100} n - \sum f_{P_{90} - 1}}{f_{P_{90}}} \cdot C_{P_{90}}$
- **Box-Plots**
  - A graphical display based on quartiles—that helps us to picture a set of data.

5.5.6 Relative Dispersion (Coefficient of Variation)

- A direct comparison of two or more measures of dispersion (e.g. standard deviations) is impossible especially when the data are in different units.
- We need to convert each of the measures into a *relative* value, called *coefficient of variation*. This is a very useful measure when:
  - The data series are completely in different measurement units.
  - The data series in the same units, but the means are far apart.
- Coefficient of variation allows us to compare the relative variability of two data sets.
- Coefficient of variation (CV):
  - The ratio of the standard deviation to the mean, expressed as a percent.
  - $CV = \dfrac{s}{x}(100)$

5.5.7 Skewness

- Another characteristic that can be measured is the degree of *skewness*.
- If a frequency distribution is *symmetrical*, the skewness is zero.
- If one or more observations are extremely large, the distribution is *positively skewed*; and if one or more are extremely small, the distribution is *negatively skewed.*
- Pearson's Coefficient of Skewness:
  - $r = \dfrac{3(\bar{x} - \tilde{x})}{s}$ or $r = \dfrac{3(\bar{x} - \hat{x})}{s}$.
  - r lies between –3 and +3 (positively or negatively skewed).

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 5: DESCRIPTIVE STATISTICS: DESCRIBING, EXPLORING AND COMPARING DATA
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 5.1 INTRODUCTION
- **Slide 5** – 5.1 INTRODUCTION (cont.)
- **Slide 6** – 5.2 FREQUENCY DISTRIBUTION
- **Slide 7** – 5.2.1 Frequency Distribution for Qualitative Data
- **Slide 8** – 5.2.2 Frequency Distribution for Quantitative Data
- **Slide 9** – 5.2.3 Class Intervals and Midpoints (Quantitative Data)
- **Slide 10** – 5.2.4 Suggestions on Constructing a Frequency Distribution (Quantitative Data)
- **Slide 11** – 5.2.5 Relative Frequency Distribution
- **Slide 12** – 5.2.6 Cumulative Frequency Distribution
- **Slide 13** – 5.3 GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION
- **Slide 14** – 5.3.1 Stem-and-Leaf Display
- **Slide 15** – 5.3.2 Histogram
- **Slide 16** – 5.3.3 The Frequency Polygon
- **Slide 17** – 5.3.4 Cumulative Frequency Polygon (Ogive)
- **Slide 18** – 5.3.5 Other Graphic Representations of Data
- **Slide 19** – 5.3.5 Other Graphic Representations of Data (cont.)
- **Slide 20** – 5.4 MEASURES OF CENTRAL TENDENCY
- **Slide 21** – 5.4.1 The Arithmetic Mean
- **Slide 22** – 5.4.1 The Arithmetic Mean (cont.)
- **Slide 23** – 5.4.2 The Median
- **Slide 24–25** – 5.4.2 The Median (cont.)
- **Slide 26** – 5.4.3 The Mode
- **Slide 27–28** – 5.4.3 The Mode (cont.)
- **Slide 29** – 5.4.4 The Geometric Mean
- **Slide 30** – 5.4.5 Selecting an Average for Data in a Frequency Distribution
- **Slide 31** – 5.4.5 Selecting an Average for Data in a Frequency Distribution (cont.)
- **Slide 32** – 5.5 MEASURES OF DISPERSION AND SKEWNESS
- **Slide 33** – 5.5.1 Why Study Dispersion?
- **Slide 34** – 5.5.2 Measures of Dispersion—Ungrouped Data
- **Slide 35–36** – 5.5.2 Measures of Dispersion—Ungrouped Data (cont.)
- **Slide 37** – 5.5.3 Measures of Dispersion—Grouped Data
- **Slide 38** – 5.5.3 Measures of Dispersion—Grouped Data (cont.)
- **Slide 39** – 5.5.4 Interpretation and Uses of the Standard Deviation
- **Slide 40** – 5.5.4 Interpretation and Uses of the Standard Deviation (cont.)
- **Slide 41** – 5.5.5 Some Other Measures of Dispersion
- **Slide 42–44** – 5.5.5 Some Other Measures of Dispersion (cont.)
- **Slide 45** – 5.5.6 Relative Dispersion (Coefficient of Variation)
- **Slide 46** – 5.5.6 Relative Dispersion (Coefficient of Variation) (cont.)
- **Slide 47** – 5.5.7 Skewness

INSTRUCTOR'S MANUAL

CHAPTER 6

Inferential Statistics: Estimation and Hypothesis Testing

Learning Objectives

The study of this chapter should enable you to:

- ❖ Define the estimators of population parameters
- ❖ Describe the interval estimation and construct the confidence intervals of population parameters
- ❖ Define and apply various types of hypothesis tests

Key Teaching Points

6.1 INTRODUCTION

- Inferential statistics attempt to reach conclusions that go beyond the data.
- It determines the overall opinion of a population about a specific issue, or to test a statistical difference between groups.
- There are two types of statistical inferences; estimation of population parameters and hypothesis testing.
- Estimation theory is used to estimate the values of population parameters based on randomly selected empirical data.
- A confidence interval provides an interval estimate for a population parameter and thus demonstrates the reliability of an estimate. It is determined by the degree of confidence or confidence level.
- Hypothesis testing is one of the most important tools of application of statistics to real life problems. It is a form of statistical inference that uses data from a sample.

6.2 ESTIMATION THEORY

- A point estimate of some population parameter, θ is a single value $\hat{\theta}$ of a statistic, $\hat{\Theta}$.
- The value $\bar{x}$ of the statistic $\bar{X}$ computed from a sample n, is a point estimate of μ.
- An *estimator* is not expected to estimate the population parameter without error. We do not expect $\bar{X}$ to estimate μ exactly, but certainly hope that it is not far off.
- For a particular sample it is possible to obtain a closer estimate of μ by using the sample median $\tilde{X}$ as an estimator. Not knowing the true μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as estimator.

6.2.1 Unbiased Estimator

- Let $\hat{\Theta}$ be an estimator whose value $\hat{\theta}$ is a point estimate of some unknown population parameter θ.

- A statistic $\hat{\Theta}$ is said to be an *unbiased estimator* if its expected value is equal to θ.

6.2.2 Variance of a Point Estimator

- If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two unbiased estimators of the same parameter θ, we would choose the estimator whose sampling distribution has the smaller variance.
- Hence, if $\sigma^2_{\hat{\Theta}_2} < \sigma^2_{\hat{\Theta}_2}$, $\hat{\Theta}_1$ is a more efficient estimator of θ than $\hat{\Theta}_2$.
- If we consider all unbiased estimators of θ, the one with the smallest variance is called the most efficient estimator, or *minimum variance unbiased estimator* (MVUE).
- For normal populations, one can show that both $\bar{X}$ and $\tilde{X}$ are unbiased estimators of the population mean μ, but the variance of $\bar{X}$ is smaller than the variance of $\tilde{X}$.
- Both $\bar{x}$ and $\tilde{x}$ will, on the average, equal the population mean μ, but $\bar{x}$ is likely to be closer to μ for a given sample, and thus $\bar{X}$ is more efficient than $\tilde{X}$.

6.3 INTERVAL ESTIMATION

- An interval estimate of θ may be written as $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on a random variable $\hat{\Theta}$ for a particular sample and its sampling distribution.
- The standard error $\sigma^2_{\hat{\Theta}_2} = \sigma^2/n$ decreases as the sample size n increases, and thus the estimate will be closer to the parameter θ, producing a shorter interval.
- Different samples would produce different interval estimates.
- If $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$, $0 < \alpha < 1$, then the probability of selecting a random sample that produces an interval containing θ is $1 - \alpha$.
- The interval estimate $\hat{\theta}_L < \theta < \hat{\theta}_U$ is called a $(1-\alpha)100\%$ confidence interval with $(1-\alpha)$ degree of confidence.
- When $\alpha = 0.05$, we have a 95% confidence interval, and when $\alpha = 0.01$ we obtain a wider 99% confidence interval.
- The wider the confidence interval is, the more confident we can be that the given interval contains the unknown parameter.

6.3.1 Single Sample: Estimating the Mean

- The best estimator of the population mean, μ is the sample mean.
- The sampling distribution of the sample mean is centered at μ with smaller variance than other estimators. Thus, the point estimate for μ is the sample mean.
- Confidence Interval of μ: σ Known:
  - If $\bar{X}$ is the mean of a random sample of size n from a population with known variance σ^2, a $(1 - \alpha)100\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Theorem 6.3.1

If $\bar{x}$ is used as an estimate of μ, we can then be $(1 - \alpha)100\%$ confident that the error will not exceed $z_{\alpha/2}\sigma/\sqrt{n}$.

Theorem 6.3.2

If $\bar{x}$ is used as an estimate of μ, we can be $(1 - \alpha)100\%$ confident that the error will not exceed a specified amount e when $n = (z_{\alpha/2}\sigma/e)^2$.

- Confidence Interval for μ; σ Unknown:
  - If $\bar{x}$ and s are the mean and standard deviation of a random sample of size n from a normal population with unknown σ^2, a $(1-\alpha)100\%$ confidence interval for μ

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

6.3.2 Two Samples: Estimating the Difference between Two Means

- Confidence Interval for $\mu_1 - \mu_2$; σ_1 and σ_2 Known

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Confidence Interval for $\mu_1 - \mu_2$; $\sigma_1 = \sigma_2$ but Unknown

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - n_2}$, $t_{\alpha/2,\upsilon}$ is the t-value with degrees of freedom $\upsilon = n_1 + n_2 - 2$

- Confidence Interval for $\mu_1 - \mu_2$; $\sigma_1 \neq \sigma_2$ and Unknown

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$t_{\alpha/2,\upsilon}$ is the t-value with degrees of freedom $\upsilon = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$

- Confidence Interval for $\mu_D = \mu_1 - \mu_2$ for Paired Observations

$\bar{d} - t_{\alpha/2}\dfrac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{\alpha/2}\dfrac{s_d}{\sqrt{n}}$, $t_{\alpha/2}$ is the t-value with degrees of freedom $\upsilon = n - 1$

6.3.3 Single Sample: Estimating a Proportion

- In a Binomial experiment, the proportion p is estimated by the statistic, $\hat{P} = X/n$; the point estimator of p where X is the number of successes in n trials.
- By CLT, for n sufficiently large, $\hat{P}$ is approximately normally distributed with mean

$$\mu_{\hat{p}} = E(\hat{P}) = \left[\frac{X}{n}\right] = \frac{np}{n} = p, \text{ and variance } \sigma_{\hat{p}}^2 = \sigma_{x/n}^2 = \frac{\sigma_x^2}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}$$

- Large–Sample Confidence Interval for p
  - If $\hat{p}$ is the proportion of successes in a random sample of size n, and $\hat{q} = 1 - \hat{p}$, a $(1 - \alpha)100\%$ confidence interval for the binomial parameter p is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Theorem 6.3.3

If $\hat{p}$ is used as an estimate of p, we can be $(1 - \alpha)100\%$ confident that the error will not exceed $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

Theorem 6.3.4

If $\hat{p}$ is used as an estimate of p, we can be $(1-\alpha)100\%$ confident that the error will be less than a specified amount e when the sample size is $n = z_{\alpha/2}^2\hat{p}\hat{q}/e^2$.

Theorem 6.3.5

If $\hat{p}$ is used as an estimate of p, we can be *at least* $(1 - \alpha)100\%$ confident that the error will not exceed a specified amount e when the sample size is $n = z_{\alpha/2}^2/4e^2$.

6.3.4 Two Samples: Estimating the Difference between Two Proportions

- By choosing independent samples from the two populations, the variables $\hat{P}_1$ and $\hat{P}_2$ will be independent, and we can conclude that $\hat{P}_1 - \hat{P}_2$ is approximately normally distributed with mean, $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$.
- Large–Sample Confidence Interval for $p_1 - p_2$
  - If $\hat{p}_1$ and $\hat{p}_2$ are the proportion of successes in random samples of size n_1 and n_2, $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$, an approximate $(1 - \alpha)100\%$ confidence interval for the difference of two binomial parameters $p_1 - p_2$, is given by,

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

6.3.5 Single Sample: Estimating the Variance

- Consider a normal population with σ^2. When a random sample of size n is selected from the population, the calculated variance s^2 will be used as the estimator of σ^2.
- Confidence Interval for σ^2
  - If s^2 is the variance of a random sample of size n from a normal population, (1-α) 100% confidence interval for σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \ \chi^2_{\alpha/2} \text{ and } \chi^2_{1-\alpha/2} \text{ are } \chi^2 \text{ – values with degree of freedom } \upsilon = n - 1$$

6.3.6 Two Samples: Estimating the Ratio of Two Variances

- If σ_1^2 and σ_2^2 are the variances of normal populations, we can establish an interval estimate of $\sigma_1^2/\ \sigma_2^2$ by using the statistic $F = \sigma_2^2 S_1^2/\ \sigma_1^2 S_2^2$. The random variable F has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.
- Confidence Interval for $\sigma_1^2/\ \sigma_2^2$
  - If s_1^2 and s_2^2 are the variances of independent samples of size n_1 and n_2 from normal populations, then a $(1 - \alpha)$ 100% confidence interval for $\sigma_1^2/\ \sigma_2^2$ is

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(v_1, v_2), \text{ where } v_1 = n_1 - 1 \text{ and } v_2 = n_2 - 1$$

6.4 TESTS OF HYPOTHESIS

- Hypothesis is a proposition, statement, or assumption based on some previous observations about the value of a population parameter for testing purposes.
- Hypothesis testing is a process based on some sample data and probability theory to conclude whether the hypothesis is reasonable and should not be rejected.
- To test the validity of the proposition, we must select a sample from the population, calculate the sample statistics, and based on certain decision rules, accept or reject the hypothesis.

6.4.1 Five-Step Procedure for Testing a Hypothesis

- Step 1: The Null Hypothesis and the Alternate Hypothesis
  - State the hypothesis to be tested—the null hypothesis, H_0
  - Either reject or 'fail to reject' the null hypothesis
  - Alternate hypothesis H_1 describes what we will conclude if we reject H_0
- Step 2: The Level of Significance
  - Probability of rejecting the null hypothesis when it is actually true
  - Traditionally, 0.05 level is selected for consumer research projects, 0.01 for quality assurance, and 0.10 for political polling

- Type-I-error: Rejecting H_0 when it is actually true. P(Type-I-error) = α
- Type-II-error: Accepting H_0 when it's actually false. P(Type-II-error)= β
- Decisions the researcher could make and the possible consequences:

Null Hypothesis	Researcher	
	Accepts H_0	Rejects H_0
If H_0 is true and	Correct decision	Type I error
If H_0 is false and	Type II error	Correct decision

- Step 3: The Test Statistic
  - There are many test statistics; z, t, F and χ^2.
  - The *test statistic* is a value, determined from sample information, used to determine whether or not to reject the null hypothesis.
- Step 4: The Decision Rule
  - A declaration of the situations under which H_0 is rejected and not rejected.
  - The *rejection region* describes the location of all those values that are so small or large that the probability of their occurrence under a true H_0 is rather slim.
  - The dividing point between the regions where H_0 is rejected and not rejected is called the *critical value*.
- Step 5: Making a Decision
  - Reject or not to reject H_0 based on rejection region and critical value.

6.4.2 One-Tailed and Two-Tailed Tests of Significance and *p*-Value

- Two types of tests of significance; a *one-tailed test* looks for an increase or decrease in parameter, and a *two-tailed* looks for any change in parameter.
- One-Tailed and Two-Tailed Tests
  - The region of rejection is only in the right (upper) tail of the curve.
  - To determine the location of the rejection region is to look at the direction in which the inequality sign in the alternate hypothesis is pointing (< or >).
  - A test is one-tailed when the alternate hypothesis states a direction, for example, H_1: μ < 70 (left tail) or H_1: μ > 70 (right tail).
  - If no direction is specified under the alternate hypothesis, a *two-tailed test* is being applied, for example, H_1: μ ≠ 70.
  - In two-tailed test, the region of rejection is in both tails (divided equally).
- *p*-Value in Hypothesis Testing
  - *p*-value is the smallest level of significance, α, for which the observed data indicates that the null hypothesis should be rejected.
  - If *p*-value is smaller than α, H_0 is rejected; if it is larger than α, H_0 is not rejected.

6.4.3 Testing for the Population Mean: Population Variance Known

- *Assumption*: A sample of size n from a normal population with unknown mean but known variance.
- H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$.
- Test Statistic: $z = \dfrac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$. Rejection Region: $z > z_\alpha$, $z < -z_\alpha$, or $|z| > z_{\alpha/2}$.

6.4.4 Testing for the Population Mean: Large Sample, Population Variance Unknown

- *Assumption*: A sample of size n (large) from a normal population with unknown mean and unknown variance.
- H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$.
- Test Statistic: $z = \dfrac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$. Rejection Region: $z > z_\alpha$, $z < -z_\alpha$, or $|z| > z_{\alpha/2}$.

6.4.5 Testing for the Population Mean: Small Sample, Population Variance Unknown

- *Assumption*: A sample of size n (small) from a normal population with unknown mean and unknown variance.
- H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$.
- Test Statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$. Rejection Region: $t > t_\alpha$, $t < -t_\alpha$, or $|t| > t_{\alpha/2}$; degrees of freedom = $n-1$.

6.4.6 Testing for the Two Population Means: Population Variances Known

- *Assumption*: Two samples of size n_1 and n_2 from two normal populations with unknown means but known variances.
- H_0: $\mu_1 - \mu_2 = d_0$ versus H_1: $\mu_1 - \mu_2 > d_0$, $\mu_1 - \mu_2 < d_0$, or $\mu_1 - \mu_2 \neq d_0$.
- Test Statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$. Rejection Region: $t = z > z_\alpha$, $z < -z_\alpha$, or $|z| > z_{\alpha/2}$.

6.4.7 Testing for the Two Population Means: Large Samples, Population Variances Unknown

- *Assumption*: Two samples of size n_1 and n_2 (large) from two normal populations with unknown means and unknown variances.
- H_0: $\mu_1 - \mu_2 = d_0$ versus H_1: $\mu_1 - \mu_2 > d_0$, $\mu_1 - \mu_2 < d_0$, or $\mu_1 - \mu_2 \neq d_0$.
- Test Statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$. Rejection Region: $z > z_\alpha$, $z < -z_\alpha$, or $|z| > z_{\alpha/2}$.

6.4.8 Testing for the Two Population Means: Small Samples, Population Variances Unknown

- *Assumption*: Two independent samples of size n_1 and n_2 (small) from a normal population with unknown mean and unknown variance, i.e. $\sigma_1 = \sigma_2 = \sigma$.
- H_0: $\mu_1 - \mu_2 = d_0$ versus H_1: $\mu_1 - \mu_2 > d_0$, $\mu_1 - \mu_2 < d_0$, or $\mu_1 - \mu_2 \neq d_0$.
- Test Statistic: $t = \dfrac{(x_1 - x_2) - d_0}{s_p\sqrt{((1/n_1) + (1/n_2))}}$, $s_p = \sqrt{\dfrac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{n_1 + n_2 - 2}}$.
- Rejection Region: $t > t_\alpha$, $t < -t_\alpha$, or $|t| > t_{\alpha/2}$, degrees of freedom = $n_1 + n_2 - 2$.

6.4.9 Testing for the Two Population Means: Paired Observations

- H_0: $\mu_D = d_0$ versus H_1: $\mu_D > d_0$, $\mu_D < d_0$, or $\mu_D \neq d_0$.
- Test Statistic: $t = \dfrac{(\bar{d} - d_0)}{s_d/\sqrt{n}}$, $s_d = \sqrt{\dfrac{\sum d^2 - (\sum d)^2/n}{n - 1}}$.
- Rejection Region: $t > t_\alpha$, $t < -t_\alpha$, or $|t| > t_{\alpha/2}$, degrees of freedom = $n - 1$.

6.4.10 Choice of Sample Size for Testing Mean (and The Power of a Test)

- Suppose we wish to test the hypothesis H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$, with a significance level α when the variance is known.
- For a specific alternative, say $\mu = \mu_0 + \delta$, the power of our test is

$$1 - \beta = P(\bar{X} > a, \text{ when } \mu = \mu_0 + \delta)$$

- Therefore, $\beta = P(\bar{X} < a, \text{ when } \mu = \mu_0 + \delta) = P\left[\dfrac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} < \dfrac{a - (\mu_0 + \delta)}{\sigma/\sqrt{n}}\right]$.

$\beta = \left[Z < \dfrac{a - \mu_0}{\sigma/\sqrt{n}} - \dfrac{\delta}{\sigma/\sqrt{n}}\right] = P\left(Z < z_\alpha - \dfrac{\delta}{\sigma/\sqrt{n}}\right)$, from which we conclude that

$-z_\beta = z_\alpha - \dfrac{\delta}{\sigma/\sqrt{n}}$, and hence the choice of sample size $n = (z_\alpha + z_\beta)^2\, \sigma^2/\delta^2$, a result that is also true when the alternative hypothesis is $\mu < \mu_0$.

- In the case of a two-tailed test we obtain the power 1–β for a specified alternative when $n \approx (z_{\alpha/2} + z_\beta)^2\, \sigma^2/\delta^2$.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 6: INFERENTIAL STATISTICS: ESTIMATION AND HYPOTHESIS TESTING
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 6.1 INTRODUCTION
- **Slide 5** – 6.1 INTRODUCTION (cont.)
- **Slide 6** – 6.2 ESTIMATION THEORY
- **Slide 7** – 6.2.1 Unbiased Estimator
- **Slide 8** – 6.2.2 Variance of a Point Estimator
- **Slide 9** – 6.3 INTERVAL ESTIMATION
- **Slide 10** – 6.3 INTERVAL ESTIMATION (cont.)
- **Slide 11** – 6.3.1 Single Sample: Estimating the Mean
- **Slide 12–13** – 6.3.1 Single Sample: Estimating the Mean (cont.)
- **Slide 14** – 6.3.2 Two Samples: Estimating the Difference between Two Means
- **Slide 15–17** – 6.3.2 Two Samples: Estimating the Difference between Two Means (cont.)
- **Slide 18** – 6.3.3 Single Sample: Estimating a Proportion
- **Slide 19–21** – 6.3.3 Single Sample: Estimating a Proportion (cont.)
- **Slide 22** – 6.3.4 Two Samples: Estimating the Difference between Two Proportions
- **Slide 23** – 6.3.5 Single Sample: Estimating the Variance
- **Slide 24** – 6.3.6 Two Samples: Estimating the Ratio of Two Variances
- **Slide 25** – 6.4 TESTS OF HYPOTHESES

- **Slide 26 –** 6.4.1 Five-Step Procedure for Testing a Hypothesis
- **Slide 27–30 –** 6.4.1 Five-Step Procedure for Testing a Hypothesis (cont.)
- **Slide 31 –** 6.4.2 One-Tailed and Two-Tailed Tests of Significance and p-Value
- **Slide 32–33 –** 6.4.2 One-Tailed and Two-Tailed Tests of Significance and p-Value (cont.)
- **Slide 34 –** 6.4.3 Testing for the Population Mean: Population Variance Known
- **Slide 35 –** 6.4.4 Testing for the Population Mean: Large Sample, Population Variance Unknown
- **Slide 36 –** 6.4.5 Testing for the Population Mean: Small Sample, Population Variance Unknown
- **Slide 37 –** 6.4.6 Testing for the Two Population Means: Population Variances Known
- **Slide 38 –** 6.4.7 Testing for the Two Population Means: Large Samples, Population Variances Unknown
- **Slide 39 –** 6.4.8 Testing for the Two Population Means: Small Samples, Population Variances Unknown
- **Slide 40 –** 6.4.9 Testing for the Two Population Means: Paired Observations
- **Slide 41 –** 6.4.10 Choice of Sample Size for Testing Mean (and The Power of a Test)
- **Slide 42 –** 6.4.10 Choice of Sample Size for Testing Mean (and The Power of a Test) (cont.)

INSTRUCTOR'S MANUAL

CHAPTER 7

F-Test and Analysis of Variance (ANOVA)

Learning Objectives

The study of this chapter should enable you to:

- Describe the *F*-distribution and apply the *F*-test for comparing two population variances
- Define the analysis of variance (ANOVA)
- Apply ANOVA to compare several population means simultaneously

Key Teaching Points

7.1 INTRODUCTION

- A statistical test that follows an *F*-distribution under the null hypothesis is referred to as an *F*-test.
- It is often used while comparing statistical models to decide the most appropriate model which fits the population based on the available sample data.
- The *F*-distribution is also a probability distribution as it is used as the test statistic for a number of situations:
  - It is used to test whether two samples are from populations with equal variances.
  - It is also applied in a simultaneous comparison of two or more population means, called *analysis of variance* (ANOVA).
- In both situations, the populations are normal.

7.2 THE *F*-DISTRIBUTION

- *F*-distribution is a continuous distribution.
- It is frequently used as the null distribution of a test statistic, particularly in likelihood-ratio tests, and most frequently in ANOVA.

7.2.1 Characteristics of the *F*-Distribution

- There are a variety of *F*-distributions; the degrees of freedom in numerator and in denominator determine a particular *F*-distribution. These two parameters also determine the shape of distribution.
- The total area under the curve is 1.
- The values of *F* are always greater than or equal to zero.
- The curve representing an *F*-distribution is *positively skewed*.
- It is asymptotic, with a range value from 0 to ∞. As *X* increases, the *F* curve approaches *X*-axis; similar to normal probability distribution.

7.2.2 Comparing Two Population Variances and Validating Assumptions

- *F*-distribution is used to test the hypothesis that the variances of two normal populations are equal.
- This is useful for determining whether one normal population has more variation than another.
- *F*-test can also be used to validate assumptions for certain statistical tests; for instance, the assumption 'two population variances are equal' used in a *t*-test.
- Whether to establish that one population has more variation than another, or to verify a statement with respect to a statistical test, first the null hypothesis has to be stated:
  - H_0: The variances of two (or more) normal populations are equal, vs.
  - H_1: The variances are different; H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_1: $\sigma_1^2 \neq \sigma_2^2$.
- To conduct the test, a random sample of n_1 is obtained from one population, and a sample of n_2 is obtained from second population.
- The test statistic is s_1^2 / s_2^2 (the ratio of sample variances).
- If the null hypothesis is true, the test statistic follows an *F*-distribution with $(n_1 - 1)$ degrees of freedom in numerator and $(n_2 - 1)$ degrees of freedom in denominator.
- The sample with larger variance is considered as the first sample and its variance is placed in the numerator. The test statistic (*F* ratio) is always more than 1.
- The rejection region is determined by the upper-tail critical value *of* the *F* distributions using α/2 and two values of degrees of freedom.

7.3 ANALYSIS OF VARIANCE (ANOVA)

- *F*-distribution is also used for testing the equality of more than two means using a technique called ANOVA (*analysis of variance*).
- It consists of statistical models and procedures wherein the sample variance is partitioned into components arising from the various sources of variation.
- ANOVA makes available a statistical test regardless of whether or not the means of some groups are all equal, and hence generalizes a *t*-test for more than two groups.
- ANOVA has an obvious advantage over the two-sample *t*-test. For comparing three or more means, we have to perform multiple *t*-tests that would lead to a higher probability of a type I error, unlike ANOVA which performs comparisons at once.
- ANOVA models can be divided into three classes; fixed-effects, random-effects, and mixed-effects.
- Fixed-effects models are employed to situations in which one or more treatments are applied to the test subjects to see if there are changes in the dependent variable.
- Random-effects models are applied when the treatments are not predetermined. Random effects occur when various factors are sampled from a large population. Since the factors themselves are random variables, several assumptions and the method of contrasting the treatments vary from fixed-effects model.
- Most of the random-effects or mixed-effects models cannot be used in deducing inferences about the specific sampled factors.
- If there is concern in the realized value of the random effects, then the best linear unbiased prediction could be employed to get a 'prediction'.

7.3.1 Assumptions of ANOVA

- The term 'treatment' is used to identify the different populations being examined.
- A treatment is defined as a cause, or specific source, of variation in a set of data.
- There are a number of approaches to ANOVA. The most common is to use a linear model that relates the response to treatments.

- Even if the statistical model is non-linear, it can be approximated by a linear model for which an ANOVA might be suitable.
- Many consider a linear model to perform the ANOVA, thus created the assumptions about the probability distribution of responses:
  - Independence: The samples are independent and randomly selected from the populations; simplifies the analysis.
  - Normality: The populations being studied are normally distributed.
  - Equality (or homogeneity) of variances: The populations have equal variances. Model-based approaches usually assume that the variance is constant.

7.3.2 Analysis of Variance Procedure

- The Test Statistic is $F = \dfrac{\text{Estimated population, variance based on variation between sample means}}{\text{Estimated population, variance based on variation within samples}}$
- Common terminology for numerator is 'between-sample variance'; for denominator, it is 'within-sample variance'.
- Numerator has k-1 degrees of freedom and denominator has $N - k$ degrees of freedom, where k is the number of treatments and N is the total number of observations.
- ANOVA table:

Source of variation	(1) – Sum of squares	(2) – Degrees of freedom	(3) – Mean square (1)/(2)
Between treatments	SST	$k - 1$	SST/($k - 1$) = MSTR
Error (within treatments)	SSE	N – k	SSE/($N - k$) = MSE
Total	SS Total		

F = MSTR/MSE

- MSTR is the *mean square between treatments*.
- MSE is the *mean square due to error* (also called *mean square within treatments*).
- 'Mean square' refers to sum of squares divided by the degrees of freedom, exactly how a variance is calculated.

$$\text{SST} = \sum\left[\frac{T_c^2}{n_c}\right] - \frac{(\sum X)^2}{N}, \quad \text{SSE} = \sum X^2 - \sum\left[\frac{T_c^2}{n_c}\right], \quad \text{SS Total} = \sum X^2 - \frac{(\sum X)^2}{N}$$

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2 –** CHAPTER 7: *F*-TEST AND ANALYSIS OF VARIANCE (ANOVA)

- **Slide 3 –** LEARNING OBJECTIVES
- **Slide 4 –** 7.1 INTRODUCTION
- **Slide 5 –** 7.2 THE *F*-DISTRIBUTION
- **Slide 6 –** 7.2.1 Characteristics of the *F*-Distribution
- **Slide 7 –** 7.2.2 Comparing Two Population Variances and Validating Assumptions
- **Slide 8 – 9 –** 7.2.2 Comparing Two Population Variances and Validating Assumptions (cont.)
- **Slide 10 –** 7.3 ANALYSIS OF VARIANCE (ANOVA)
- **Slide 11–13 –** 7.3 ANALYSIS OF VARIANCE (ANOVA) (cont.)
- **Slide 14 –** 7.3.1 Assumptions of ANOVA
- **Slide 15 –** 7.3.1 Assumptions of ANOVA (cont.)

INSTRUCTOR'S MANUAL

CHAPTER

8 Chi-square Applications

Learning Objectives

The study of this chapter should enable you to:

- ❖ Describe the Chi-square test
- ❖ Apply the Chi-square test for homogeneity
- ❖ Apply Chi-square test of goodness-of-fit
- ❖ Apply Chi-square test of independence between two variables

Key Teaching Points

8.1 INTRODUCTION

- The Chi-square test is one of the simplest and most widely accepted *non-parametric* tests in various fields.
- It does not require any assumptions about the population—wider applicability.
- It is a statistical measure for homogeneity, goodness-of-fit, or independence.
- The Chi-square test can be applied to:
  - ¤ Establish whether a sample was drawn from a normal population
  - ¤ Conclude whether two random variables are independent
  - ¤ Verify whether or not categories of a variable are represented in the same proportions in two or more populations.
- The value of χ^2 represents the magnitude of discrepancy between observed and expected.
- If the $\chi^2 = 0$, there is no significant difference between observed and expected.
- The higher the value of χ^2, the higher it would be for the discrepancy between observed and expected frequencies.
- Conditions of χ^2 test: Chi-square test have the following conditions and if fail to satisfy, leads to many rejections of null hypothesis:
  - ¤ Sample observations should be independent.
  - ¤ Sample observations should be drawn randomly.
  - ¤ Total frequency should at least contain '50' observations.
  - ¤ Expected frequency in each cell should be more than '5'.
- Applications: Basically, Chi-square test has the following applications:
  - ¤ Test of Homogeneity.
  - ¤ Test of Goodness-of-fit.
  - ¤ Test of Independence.

8.2 TEST OF HOMOGENEITY

- Chi-square test for homogeneity is a test used to determine whether several populations are similar or 'homogeneous' with respect to some characteristics.
- The χ^2 test is applied when dependent variable is dichotomous (has only two categories) while *t*-test is applied when dependent variable is continuous.
- Test statistic: $\chi^2 = \sum \frac{(O - E)^2}{E}$

8.3 TEST OF GOODNESS-OF-FIT

- The Chi-square *test of goodness-of-fit* is used to test if a sample of data came from a population with a specific distribution.
- It enables us to ascertain how well a specific distribution fits the sample data.
- The null hypothesis usually states that the sample is drawn from the assumed distribution.
- A Chi-square of zero means that the model is a perfect fit of the observations to the expected frequencies.
- The expected frequency for a Poisson distribution: $Expected\ Frequency = N \cdot \frac{e^{-m} \cdot m^x}{x!}$

8.4 TEST OF INDEPENDENCE

- The *test of independence* is used to test the independence of two variables.
- If variables are uncorrelated, they are said to be independent; if correlated, they are said to be dependent.
- The null hypothesis is that the two variables are independent.
- This test tells whether or not any dependence relationship exists but does not provide either degree or direction of dependency.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 8: CHI-SQUARE APPLICATION
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 8.1 INTRODUCTION
- **Slide 5–7** – 8.1 INTRODUCTION (cont.)
- **Slide 8** – 8.2 TEST OF HOMOGENEITY
- **Slide 9** – 8.3 TEST OF GOODNESS-OF-FIT
- **Slide 10** – 8.4 TEST OF INDEPENDENCE

INSTRUCTOR'S MANUAL

CHAPTER 9

Simple Linear Regression and Correlation

Learning Objectives

The study of this chapter should enable you to:

- ❖ Define, plot and interpret a scatter diagram
- ❖ Describe simple linear regression and develop a regression equation using method of least squares
- ❖ Construct confidence intervals and prediction intervals for values of a dependent variable
- ❖ Define correlation analysis and calculate coefficient of correlation, coefficient of determination and rank correlation

Key Teaching Points

9.1 INTRODUCTION

- Many times we come across various phenomena where two variables tend to move either in the same or opposite direction. Such a relation between any two variables is called *simple correlation*.
- It is essential to have a forecast for an unknown variable. A forecast can be made through a *Regression analysis*—one of the most widely used and acceptable statistical techniques for analyzing observational data.
- It is possible to study the relationship between two or more variables and develop an equation that allows us to estimate an unknown variable based on the existing data.

9.2 SCATTER DIAGRAM

- A useful first step in looking at the relationship between two variables is to portray the information in a *scatter diagram*.
- For a scatter diagram, it is a common practice to put the dependent variable on the vertical axis (Y) and the independent variable on the horizontal axis (X).

9.3 SIMPLE LINEAR REGRESSION

- In regression analysis, we develop an equation (*regression equation*) to express the relationship between two variables, and estimate the value of the dependent variable Y based on a selected value of the independent variable X.

9.3.1 Method of Least Squares

- Judgment can be eliminated by determining the regression line using a mathematical method called *least squares method*.

- This method gives the 'best-fitting' straight line. It minimizes the sum of the squares of the vertical deviations along the line.
- The general form of the regression equation is $Y' = a + bX$.
- Regression equation is just an estimate of the relationship between the two variables.
- The values of a and b are usually referred to as the *regression coefficients.*

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2},$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n}.$$

9.3.2 The Standard Error of Estimate

- In the scatter diagram, all of the points do not lie on the regression line; perfect prediction is practically impossible.
- What is needed is a measure that would indicate how precise the prediction of Y is based on X—called the *standard error of estimate*.
- The standard error of estimate, $s_{y.x}$, measures the dispersion about an average line, the regression line.

$$s_{y.x} = \sqrt{\frac{\sum(Y - Y')^2}{n - 2}}$$

- For a large number of observations, $s_{y.x} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}$ may be used.

9.3.3 Linear Regression Assumptions and Empirical Rule

- Four assumptions of linear regression:
  (i) There is a group of *normally distributed* Y values for each X.
  (ii) The normal distributions of Y values have *means* that lie on the regression line.
  (iii) These normal distributions have *equal standard deviations*.
  (iv) The values of Y are independent.
- The Empirical Rule states that if the values are normally distributed:
  $\bar{X} \pm 1s$ encompasses approximately the middle 68 percent of the values.
  $\bar{X} \pm 21s$ encompasses approximately the middle 95.5 percent of the values.
  $\bar{X} \pm 31s$ encompasses approximately the middle 99.7 percent of the values.
  (If the distribution is highly skewed, these relationships will not hold.)
- The same relationships exist between the average predicted value, Y', and the standard error of estimate, $s_{y.x}$.
- If the scatter about the regression line is normally distributed and the sample is large, then:
  $Y^1 \pm 1s_{y.x}$ encompasses the middle 68 percent of the observed values.
  $Y^1 \pm 2s_{y.x}$ encompasses the middle 95.5 percent of the observed values.
  $Y^1 \pm 3s_{y.x}$ encompasses the middle 99.7 percent of the observed values.
- Standard deviation s measures the spread of Xs around the mean whereas standard error of estimate $s_{y.x}$ measures the spread of points around a regression line.

9.3.4 Significance Test (Linearity)

- The significance of variable X with Y can be checked using t-test.
- H_0: $\beta = 0$ versus H_1: $\beta \neq 0$; let level of significance be 5%.

$$t = \frac{b}{SE(b)} \qquad SE(b) = \sqrt{\frac{\sum(Y - Y')^2}{(n - 2)\sum(X - \bar{X})^2}}$$

9.4 CONFIDENCE INTERVAL AND PREDICTION INTERVAL

- We are interested in providing interval estimates of two types:
  - *Confidence interval*—reports the *mean* value for a given X
  - *Prediction interval*—reports the range of values for a particular X.

9.4.1 Confidence Interval of an Estimate

- The confidence interval for the mean value of Y for a given X:

$$Y' \pm t_{\alpha/2,v}\,(s_{y.x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - [(\sum X)^2/n]}}$$

9.4.2 Prediction Interval of an Estimate

- The prediction interval for a particular value of Y for a given value of X:

$$Y' \pm t_{\alpha/2,v}\,(s_{y.x})\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - [(\sum X)^2/n]}}$$

- There is an important distinction between a confidence interval and a prediction interval; a confidence interval refers to all cases with a given value of X whereas a prediction interval refers to a particular case for a given value of X.
- The prediction interval will have the wider ranges a result of the extra 1.

9.5 CORRELATION ANALYSIS

- *Correlation analysis* is a group of statistical techniques used to measure the strength of the association between numerical variables.
- Correlation is defined as the degree of linear relationship between two or more variables and also referred to as *covariation*.
- The correlation between two variables is sometimes called a *simple correlation*.
- A correlation is also used to represent the strength of a relationship between two factors which is referred to as a statistical index.
- A correlation does not provide information on cause and effect.
- The degree of linear relationship between one (dependent) variable and several other (independent) variables is called *multiple-correlation*.
- *Partial correlation* is the degree of linear relationship between two variables after excluding the effects of other factors.

9.5.1 Types of Correlation

- A simple correlation can be further divided into positive and negative correlations.
- *Positive correlation*—the values of two variables are increasing or decreasing in the same direction.
- *Negative correlation*—the values of two variables are moving in opposite directions.
- If there is no relation exists between, the variables are said to be *uncorrelated*.
- There are also other types of correlation; if the change in one variable is in a *constant ratio* with the change in the other variable, it is referred to as a *linear correlation*. Otherwise, if the change is *not constant*, it is referred to as a *non-linear correlation*.

9.5.2 Simple Correlation and Statistical Relationship

- In correlation analysis, we will develop some statistical measures to portray and explain more precisely the relationship between these two variables.
- Correlation analysis essentially seeks to determine how strong the relationship between two variables is.

- Relationship can be measured using the *coefficient of correlation* on a scale –1 to +1.
- The scatter diagram is a useful first step when looking at the relationship between two variables.

9.5.3 Coefficient of Correlation

- Coefficient of correlation describes the strength of the relationship between two sets of variables—Pearson's *r*, the *Pearson product-moment coefficient of correlation*.
- If there is absolutely no relationship between two sets of variables, Pearson's $r = 0$.
- A coefficient of correlation *r* close to 0 (say, 0.08) shows that the relationship is quite weak; the same conclusion for $r = -0.08$.
- Coefficients of –0.91 and +0.91 have equal strength; both indicate very strong correlation between the two sets of variables.
- Thus, the strength of the correlation does not depend on the direction.
- If the correlation is weak, there is considerable scatter about a straight line drawn through the centre of the data.
- For a strong relationship, there is very little scatter about the straight line.

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

9.5.4 Coefficient of Determination

- A measure that has a more exact meaning is the *coefficient of determination*—computed by 'squaring' the coefficient of correlation.
- Coefficient of determination is the proportion of the total variation in the dependent variable *Y* that is explained by the variation in the independent variable *X*.

9.5.5 Rank Correlation

- To study the relationship between sets of *ranked data*, a measure called *Spearman's coefficient of rank correlation*, r_s, is used.

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

- The coefficient of rank correlation can assume any value from –1.00 to +1.00.
- A value of –1.00 indicates perfect negative correlation and a value of +1.00 indicates perfect positive correlation among the ranks.
- A rank correlation of 0 indicates that there is no association among the ranks.
- A rank correlation of –0.84 and +0.84 both indicate a strong association.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 9: SIMPLE LINEAR REGRESSION AND CORRELATION
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 9.1 INTRODUCTION
- **Slide 5** – 9.1 INTRODUCTION (cont.)
- **Slide 6** – 9.2 SCATTER DIAGRAM
- **Slide 7** – 9.3 SIMPLE LINEAR REGRESSION
- **Slide 8** – 9.3 SIMPLE LINEAR REGRESSION (cont.)
- **Slide 9** – 9.3.1 Method of Least Squares
- **Slide 10** – 9.3.1 Method of Least Squares (cont.)
- **Slide 11** – 9.3.2 The Standard Error of Estimate
- **Slide 12** – 9.3.3 Linear Regression Assumptions and Empirical Rule
- **Slide 13–14** – 9.3.3 Linear Regression Assumptions and Empirical Rule (cont.)
- **Slide 15** – 9.3.4 Significance Test (Linearity)
- **Slide 16** – 9.4 CONFIDENCE INTERVAL AND PREDICTION INTERVAL
- **Slide 17** – 9.4.1 Confidence Interval of an Estimate
- **Slide 18** – 9.4.2 Prediction Interval of an Estimate
- **Slide 19** – 9.5 CORRELATION ANALYSIS
- **Slide 20** – 9.5 CORRELATION ANALYSIS (cont.)
- **Slide 21** – 9.5.1 Types of Correlation
- **Slide 22** – 9.5.1 Types of Correlation (cont.)
- **Slide 23** – 9.5.2 Simple Correlation and Statistical Relationship
- **Slide 24** – 9.5.3 Coefficient of Correlation
- **Slide 25–29** – 9.5.3 Coefficient of Correlation (cont.)
- **Slide 30** – 9.5.4 Coefficient of Determination
- **Slide 31** – 9.5.5 Rank Correlation
- **Slide 32** – 9.5.5 Rank Correlation (cont.)

INSTRUCTOR'S MANUAL

CHAPTER 10 Time Series Analysis and Forecasting

Learning Objectives

The study of this chapter should enable you to:

- ❖ Understand the concepts of trend, cyclical variation, seasonal variation and irregular variation
- ❖ Describe a linear trend, use least square and moving-average methods to analyze trend
- ❖ Describe seasonal variation, determine a seasonal index and use deseasonalized data to forecast
- ❖ Forecast using exponential smoothing with trend and when appropriate, applying a seasonal effect

Key Teaching Points

10.1 INTRODUCTION

- A time series is a sequence of data points measured successively at uniform time intervals.
- An analysis of a time series can be used by management to make current decisions, long-term forecasting and planning.
- It is important to obtain long-term forecasts to allow sufficient time, for instances, for developing new plants, planning of raw materials, and obtaining financial supports.
- Time series analysis consists of methods for analyzing time series data to extract important statistics and other features of the data.
- Time series data have a natural sequential order of observations which makes the analysis different from general data analysis.

10.2 ELEMENTS OF TIME SERIES

- Four *elements* (*component factors*) to a time series; trend, cyclical variation, seasonal variation and irregular (random) variation.
- The *classical multiplicative time series model* states that any observed value in a time series is the *product* of these factors.
- When the time series data are recorded *annually*, an observation Y_i for the year i may be expressed as $Y_i = T_i \cdot C_i \cdot I_i$; T_i, C_i, I_i are trend, cyclical and irregular components.
- When the time series are recorded either *quarterly* or *monthly*, an observation Y_i for the time period i may be expressed as $Y_i = T_i \cdot S_i \cdot C_i \cdot I_i$; S_i is the seasonal components.
- The first step in a time series analysis is to plot the data and observe their tendencies over time - whether there is a *trend* or whether the series *oscillates* over time.

10.2.1 Trend

- The secular (long-term) trend can be recognized by a smooth long-term direction.
- The trends may move upward, decline or remain the same over a period of time.

10.2.2 Cyclical Variation

- Another important time series component.
- The business cycle typically consists of *four periods* in sequence; prosperity, recession, depression and recovery.
- During a recession, business and economic time series are below their long-term trends, and during prosperity, these series are above their long-term trends.
- The fluctuations of a business are the results of a cyclical variation—has to go through these four periods consecutively.

10.2.3 Seasonal Variation

- Another component of a time series.
- Many sales, production and other series fluctuate with the seasons.
- Most business and economic series have recurring seasonal patterns.

10.2.4 Irregular Variation

- Also referred to as 'residual variation' since it represents what is left after the trend, cyclical and seasonal variations.
- Irregular fluctuation occurs due to unforeseen events such as natural disasters.
- Most analysts divide the irregular variation into episodic and residual variation.
- The episodic fluctuations can be identified but cannot be predicted.
- When the episodic fluctuations have been eliminated, what remains is the residual variation.
- The residual fluctuations cannot be identified and predicted.
- Both episodic and residual variations cannot be projected into the future.

10.3 TREND ANALYSIS

- The concept of gathering information and spotting a pattern (trend).
- Even though it is frequently applied to foresee future events, it could also be applied to estimate uncertainties based on past events.

10.3.1 Linear Trend

- The long-term trend of business and economic time series frequently approximates a straight line. The equation of the straight line may be written as $Y' = a + bt$.

10.3.2 Estimation of Trend Analysis by Least Squares Method

- Common method of constructing a straight line equation through data points to obtain 'best-fitting' line is called the *least squares method.*
- It uses calculus to determine the minimum sum of squares of the vertical differences of each point from the suggested straight line.
- Two unknown parameters (a and b) that give the least squares equation are determined using the following equations

$$b = \frac{\sum tY - (\sum Y)(\sum t)/n}{\sum t^2 - (\sum t)^2/n}, \; a = \frac{\sum Y}{n} - b\left(\frac{\sum t}{n}\right)$$

10.3.3 Estimation of Trend Analysis by Moving-Average Method

- Moving average method is one of the most popular approaches for smoothing out time series data; also the basic technique for measuring the seasonal fluctuation.
- The moving-average smoothes out the fluctuations by 'moving' the arithmetic means through the time series.
- In practice, moving-average method would not be able to produce a straight line for the trend.

10.4 SEASONAL VARIATION

- Another component of a time series. Many business and economic series have periods of above- and below-average activities during the year.
- In the manufacturing industry, the main reason for analyzing the seasonal variation is to ensure an adequate supply of raw materials and manpower.
- Seasonal variation analysis over a period of years can be exploited to assess the current sales. An index can be used to represent the sales of a particular product.

10.4.1 Determining a Seasonal Index

- For monthly data, there are 12 indexes (12-month period). For quarterly data (every three months), there are 4 seasonal indexes (four typical seasons).
- Each index is written as a percent and the average is equal to 100.0. Each index indicates the level of a value in relation to the average (100.0).
- A number of approaches have been developed to compute the typical seasonal pattern in a time series. The most common is the *ratio-to-moving-average method*.
- The method removes the trend, cyclical and irregular components from the data.
- The resulting numbers are called the *typical seasonal indexes*.

10.4.2 Deseasonalizing Data

- The typical indexes are important in adjusting a time series for seasonal variation.
- The resulting series is referred to as *deseasonalized* or *seasonally adjusted series*.
- 'Deseasonalizing' aims at removing the seasonal variation so that the cycle and trend can be studied.

10.4.3 Using Deseasonalized Data to Forecast

- The seasonally adjusted forecasts can be produced by combining the process for identifying trend and the seasonal adjustments.
- First, the least squares trend equation is determined, then the trend values for future periods are projected, and finally these values are adjusted for seasonal effects.

10.5 TIME SERIES SMOOTHING AND FORECASTING

- When the overall long-term trend movements in a time series is obscured by the amount of variation from year to year, it becomes difficult to judge whether any long-term upward or downward trend effect really exists in the series.
- Method of *moving-average* and method of *exponential smoothing* may be used to smooth a series and provide us with an overall impression of the pattern of movement in the data over time.

10.5.1 Moving-Average Method

- Moving-average method for smoothing a time series is highly subjective and dependent on the length of the period selected for constructing the averages.
- The period should be an integer that corresponds to the estimated average length of a cycle.
- For example, a three–month moving average would be calculated by taking the most recent three months data, averaging them and using it as the forecast for the next month.

10.5.2 Exponential Smoothing Method

- A very popular approach for smoothing a time series.
- This method allocates exponentially decreasing weights to the observations. The weights are determined by smoothing parameters (at least one).

- Exponential smoothing makes use of a *smoothing constant,* α. This is the percentage of the forecast affected by the most recent data point.
- The *basic equation* of exponential smoothing is

  New forecast = α(latest data point) + (1 – α)(previous forecast).
- The *basic equation* for exponential smoothing model:
  D_t = sales (demand) in period t, F_t = forecast for period t, α = smoothing constant,
  F_{t+1} = forecast for period t+1 (knowing sales in period t);
  $F_{t+1} = \alpha D_t + (1 - \alpha)F_t$
- Exponential Smoothing with Linear Trend:
  - Basic exponential smoothing will always lag behind any systematic increase or decrease in data.
  - G_t = one-period trend estimate, then the basic exponential smoothing equation is modified to include the trend estimate as $S_t = \alpha D_t + (1 - \alpha)(S_{t-1} + G_{t-1})$.
  - S_t is like an updated baseline forecast except that it doesn't project forward the trend for the following period.
  - A more convenient way to update the trend factor is to use exponential smoothing again, $G_t = \beta(S_t - S_{t-1}) + (1 - \beta)G_{t-1}$.
  - G_{t-1} plays a role similar to the previous value of the forecast in the basic equation; the latest value is represented by the difference between the last two baseline forecasts, $S_t - S_{t-1}$; a different smoothing constant β is used for some flexibility.
  - Since there are two equations, this model is called a *two-equation model with linear trend*, or a *two-equation model*.
  - The forecast for one period ahead is calculated as $F_{t+1} = S_t + G_t$
- Exponential Smoothing with Seasonal Factors:
  - In many situations, demand may follow a seasonal pattern.
  - It is desirable to take advantage of this information in adjusting forecasts obtained by basic exponential smoothing or exponential smoothing with trend.
  - Seasonality is included by defining individual *seasonal factors* for each season. For monthly data, there are 12 seasonal factors C_1 to C_{12}.
  - The seasonal factor for a season equals the average values for that period divided by the overall average.
  - Seasonal factors C_i are included in S_t equation: $S_t = \alpha(D_t/C_{t-N}) + (1 - \alpha)(S_{t-1} + G_{t-1})$
  - The seasonal factors are updated as new data becomes available.
  - For trend term, we again use exponential smoothing: $C_t = \gamma(D_t/S_t) + (1 - \gamma)C_{t-N}$
  - γ is a different smoothing constant for updating the seasonal factors.
  - F_{t+1} equation becomes $F_{t+1} = (S_t + G_t)C_{t+1-N}$

10.5.3 Model Initialization and Smoothing Constant Values

- A variety of ways to take past data and initialize an exponential smoothing model.
- One suggested procedure is to simply run through the data twice, use averages over the entire set of data to obtain average per period, average trend, and average seasonal factors, and go back to the first data point and use the equations to bootstrap forward just as you would if the data became available.
- To do this, you would need at least two and preferably three seasonal cycles to get good estimates of the seasonal factors.
- α and β generally are set between 0.10 and 0.30; the larger the value of the smoothing constant, the more responsive the forecast to recent changes in the data.
- However, sometimes the respond will simply be a response to random changes from month to month; for this reason, the α and β are kept ≤ 0.30.

- Seasonal smoothing factor γ often have a higher value, perhaps 0.20 to as high as 0.60, since each seasonal factor is updated only once in a complete seasonal cycle.

10.5.4 Forecast Error Measurement

- Since forecasts are 'always wrong', we need to analyze forecast error in order to make optimal decisions.
- Forecast error is defined as: *Error = Forecast – Actual*, or $e_t = F_t - D_t$
- Mean Squared Error (MSE) involves taking each error, squaring it, and taking the average.
- Mean Absolute Deviation (MAD) takes the absolute value of each error and average them.
- Mean Absolute Percentage Error (MAPE) takes the absolute percentage of each error and averages them.

10.5.5 New Product Forecasting

- For new products, there is generally no past data. In such cases, companies often analyze data from similar products.

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 10: TIME SERIES ANALYSIS AND FORECASTING
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 10.1 INTRODUCTION
- **Slide 5** – 10.1 INTRODUCTION (cont.)
- **Slide 6** – 10.2 ELEMENTS OF TIME SERIES
- **Slide 7** – 10.2 ELEMENTS OF TIME SERIES (cont.)
- **Slide 8** – 10.2.1 Trend
- **Slide 9** – 10.2.2 Cyclical Variation
- **Slide 10** – 10.2.2 Cyclical Variation (cont.)
- **Slide 11** – 10.2.3 Seasonal Variation
- **Slide 12** – 10.2.3 Seasonal Variation (cont.)
- **Slide 13** – 10.2.4 Irregular Variation
- **Slide 14** – 10.2.4 Irregular Variation (cont.)
- **Slide 15** – 10.3 TREND ANALYSIS
- **Slide 16** – 10.3.1 Linear Trend
- **Slide 17** – 10.3.2 Estimation of Trend Analysis by Least Squares Method
- **Slide 18** – 10.3.3 Estimation of Trend Analysis by Moving-Average Method
- **Slide 19** – 10.4 SEASONAL VARIATION
- **Slide 20** – 10.4 SEASONAL VARIATION (cont.)
- **Slide 21** – 10.4.1 Determining a Seasonal Index
- **Slide 22** – 10.4.1 Determining a Seasonal Index (cont.)

- **Slide 23 –** 10.4.2 Deseasonalizing Data
- **Slide 24 –** 10.4.3 Using Deseasonalized Data to Forecast
- **Slide 25 –** 10.5 TIME SERIES SMOOTHING AND FORECASTING
- **Slide 26 –** 10.5.1 Moving-Average Method
- **Slide 27 –** 10.5.2 Exponential Smoothing Method
- **Slide 28 –** 10.5.2 Exponential Smoothing Method (cont.)
- **Slide 29 –** 10.5.3 Model Initialization and Smoothing Constant Values
- **Slide 30 –** 10.5.3 Model Initialization and Smoothing Constant Values (cont.)
- **Slide 31 –** 10.5.4 Forecast Error Measurement
- **Slide 32–33 –** 10.5.4 Forecast Error Measurement (cont.)
- **Slide 34 –** 10.5.5 New Product Forecasting

INSTRUCTOR'S MANUAL

CHAPTER

11 Index Numbers

Learning Objectives

The study of this chapter should enable you to:

- ❖ Define uses and various types of index numbers
- ❖ Describe and calculate simple index number
- ❖ Describe and calculate unweighted indexes
- ❖ Describe and calculate weighted indexes

Key Teaching Points

11.1 INTRODUCTION

- Index numbers—the most widely used indicators in statistics.
- Commonly used as 'barometers' to indicate the states of the economic activities.
- Consumer Price Index (CPI) and other business indexes are published on a regular basis.
- CPI measures the changes in the price level of consumer goods and services.
- Why do we need an index? To reflect changes in a group of items (price or quantity).
- Easier to evaluate the trend in a time series composed of large numbers using index.
- 'Index number' is a measure aimed to illustrate changes in a variable or a group of variables.
- Three types of principal indexes; price index, quantity index and value index.
- A price index compares changes in prices, a quantity index determines the changes in quantities, and a value index measures the combined changes of both.
- Two popular approaches; aggregates method and average of relative method.
- The index computed in either method could be unweighted or weighted index.
- An unweighted index considers equal weights, and a weighted index assigns weights according to the values of items.

11.2 CHARACTERISTICS AND USES OF AN INDEX NUMBER

- The characteristics of an index number:
  - ¤ A percentage computed as a ratio of the current value to a base value (100).
  - ¤ Specialized averages for comparison of items that are stated in different units.
  - ¤ Measures changes that cannot be computed directly.
- The uses of index numbers:
  - ¤ Establish trends to reveal a general movement of the event.
  - ¤ Guide policy making. The CPI, for instance, is widely used as the basis to decide the wages of the employees from time to time.
  - ¤ Conclude the purchasing power of a currency
  - ¤ Deflate time series data to reflect reality.

11.3 BASE PERIOD AND BASE NUMBER

- Each index has a base. The base period for most indexes was 1967 (1967 = 100).
- Now, indexes have various base periods. The Malaysia CPI is currently 1994 = 100.
- Base number of most indexes is 100; no reason why other numbers cannot be used.

11.4 SIMPLE INDEX NUMBER AND VALUE INDEX

- Simple index number (the ratio of two values of the variable converted to a percentage) is used to evaluate the relative change in one variable.
- The most important use of an index in business and economic is to demonstrate the change in percentage of one or more items from one period to another.
- A value index is the ratio of the value of all items in a given period to the value of all items in the base period; it measures the combined changes of price and quantity.

11.5 UNWEIGHTED INDEX

- An index where equal weights are implicitly assigned to all items in a group.
- Three methods; relative method, average of relative method and aggregate method.

11.5.1 Unweighted Price Index

- Doesn't reflect the reality since the price changes are not linked to any usage levels.
- The base-period price is p_0, and a price other than the base period is p_1.
- Relative price index measures the change in price for a specific item.

$$\textit{Relative Price Index} = \frac{p_1}{p_0} \times 100$$

- Average of relative price index measures the overall performance of a certain price change.

$$\textit{Average of Relative Price Index} = \frac{\sum \frac{p_1}{p_0} \times 100}{k}; \text{ where } k \text{ is the number of items.}$$

- Aggregate price index determines the price change of a *group* of similar items; items taken into consideration have to be in the same unit.

$$\textit{Aggregate Price Index} = \frac{\sum p_1}{\sum p_0} \times 100$$

11.5.2 Unweighted Quantity Index

- Relative quantity index measures changes in terms of quantity for a specific item.

$$\textit{Relative Quantity Index} = \frac{q_1}{q_0} \times 100$$

- Average of relative quantity index measures the overall performance of a certain quantity change.

$$\textit{Average of Relative Quantity Index} = \frac{\sum \frac{q_1}{q_0} \times 100}{k}; \text{ where } k \text{ is the number of items.}$$

- Aggregate quantity index measures the quantity change of a group of similar items; the items have to be in the same unit.

$$\textit{Aggregate Quantity index} = \frac{\sum q_1}{\sum q_0} \times 100$$

11.6 WEIGHTED INDEX

- A major disadvantage of unweighted is all items are assumed of equal importance.
- A substantial price change for slow moving items can completely distort an index.
- There are three methods of computing a weighted index; Laspeyres, Paasche and fixed-weight aggregate methods.
- The methods differ only with respect to the period used for weighting.
- Laspeyres uses base-year weights, Paasche uses current-year weights, and fixed-weight aggregate method chooses one period other than base and current years.

11.6.1 Weighted Price Index

- The base-period price is p_0, and the selected period price is p_1.
- Laspeyres price index determines a weighted price index using base-period quantities as weights.

$$Laspeyres\ Price\ Index = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

- Paasche price index determines a weighted price index using current-period quantities as weights.

$$Paasche\ Price\ Index = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

- Paasche price index is preferable because it takes into account the change in price and consumption patterns.
- Fixed-weight aggregate price index does not use quantities consumed in the current or base period; it uses weights from a representative period—fixed weights from a single or several years.

$$Fixed\text{-}Weight\ Aggregate\ Price\ Index = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

11.6.2 Weighted Quantity Index

- The base-period quantity is q_0, and the selected period quantity is q_1.
- Laspeyres quantity index determines a weighted quantity index using base-period prices as weights.

$$Laspeyres\ Quantity\ Index = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

- Paasche quantity index determines a weighted quantity index using current-year prices as weights.

$$Paasche\ Quantity\ Index = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

- Fixed-weight aggregate quantity index does not use prices in the current or base period; it uses fixed weights from a representative period—a single or several years.

$$Fixed\text{-}Weight\ Aggregate\ Quantity\ Index = \frac{\sum q_1 p}{\sum q_0 p} \times 100$$

Teaching Notes

Reference to PowerPoint Slides

- **Slide 2** – CHAPTER 11: INDEX NUMBERS
- **Slide 3** – LEARNING OBJECTIVES
- **Slide 4** – 11.1 INTRODUCTION
- **Slide 5 – 7** – 11.1 INTRODUCTION (cont.)
- **Slide 8** – 11.2 CHARACTERISTICS AND USES OF AN INDEX NUMBER
- **Slide 9** – 11.2 CHARACTERISTICS AND USES OF AN INDEX NUMBER (cont.)
- **Slide 10** – 11.3 BASE PERIOD AND BASE NUMBER
- **Slide 11** – 11.4 SIMPLE INDEX NUMBER AND VALUE INDEX
- **Slide 12** – 11.4 SIMPLE INDEX NUMBER AND VALUE INDEX (cont.)
- **Slide 13** – 11.5 UNWEIGHTED INDEX
- **Slide 14** – 11.5.1 Unweighted Price Index
- **Slide 15–17** – 11.5.1 Unweighted Price Index (cont.)
- **Slide 18** – 11.5.2 Unweighted Quantity Index
- **Slide 19–20** – 11.5.2 Unweighted Quantity Index (cont.)
- **Slide 21** – 11.6 WEIGHTED INDEX
- **Slide 22** – 11.6 WEIGHTED INDEX (cont.)
- **Slide 23** – 11.6.1 Weighted Price Index
- **Slide 24–25** – 11.6.1 Weighted Price Index (cont.)
- **Slide 26** – 11.6.2 Weighted Quantity Index
- **Slide 27–28** – 11.6.2 Weighted Quantity Index (cont.)

SOLUTION MANUAL

CHAPTER 1 Introduction to Statistics

1 Differentiate between descriptive and inferential statistics.
Descriptive statistics just explains the sample data, whereas inferential statistics tries to reach conclusions that go beyond the existing data.

2 Explain the differences between primary and secondary data.
Primary data is the specific information collected by the person who is doing the research, whereas secondary data is any material that has been collected from published records (newspapers, journals, research papers, etc).

3 Define the following terms:
(a) Secondary data
Data that have been already collected by and readily available from other sources.
(b) Census
The procedure of systematically acquiring and recording information about the members of a given population.
(c) Inferential statistics
To apply the conclusions obtained from one experimental study to more general populations.
(d) Quantitative data
Data measured or identified on a numerical scale.

4 For each of the following, identify whether the descriptive or inferential statistics have been used.
(a) In general, men die earlier than women.
Inferential statistics
(b) A researcher has concluded that the property values will increase.
Inferential statistics
(c) In Malaysia, it is found that 45% of school children are obese in which 60% are males.
Descriptive statistics
(d) A study based on a random sample has revealed that the school children are obese because they always preferred fast foods.
Inferential statistics

5 Determine whether each of the following statements is TRUE or FALSE.
(a) If a researcher uses descriptive statistics, the researcher will be able to conclude about the population based on a sample. FALSE

(b) Probability is the basis of the inferential statistics. TRUE
(c) Marital status is an example of a qualitative data. TRUE
(d) The highest level of measurement is the ratio level. TRUE
(e) The examination grades (A to F) are an example of ordinal scale measurement. FALSE
(f) Phone survey is the most expensive method of data collection. FALSE

6 Identify the type of measurements (nominal, ordinal, interval and ratio) for each of the following:

(a) Test grades. Interval
(b) Size of shoe. Ordinal
(c) Type of blood. Nominal
(d) Weight of chicken in kg. Ratio
(e) The top five supermodels. Ordinal
(f) Rating given to the cleanliness of restaurants. Interval
(g) The times recorded by the runners in a 100-metres sprint. Ratio
(h) The ranking of the top 10 Malaysia's richest people for 2010. Ordinal
(i) The positions in a soccer team such as striker and goalkeeper. Ordinal
(j) The average day temperature recorded at 14 cities in Malaysia. Interval
(k) The number of accidents on a highway during the New Year festival. Ratio

SOLUTION MANUAL

CHAPTER 2 Concepts of Probability

1 A coin is tossed twice. Find the

(a) Sample space

S = {HH, HT, TH, TT}

(b) Probability of getting both heads

P(both heads) = 1/4

2 In a dice tossing experiment, two events are defined as follows:

A = {an odd number}; **B = {a number less than 4}.**

List the elements of

(a) A or B

$A \cup B = \{1, 2, 3, 5\}$

(b) A and B

$A \cap B = \{1, 3\}$

(c) A' or B'.

$A' \cup B' = \{2, 4, 5, 6\}$

3 Ten chairs of different colours are to be arranged in a circle. Determine the number of possible different arrangements.

Circular permutation: $(n-1)!$ The number of different arrangements (in a circle) of ten chairs of different colours = (10–1)! = 362 880

4 Using nine numbers 0, 1, 2, 3, 4, 5, 6, 7 and 8, answer the following:

(a) How many 5-digit numbers can be formed if repetition is not allowed?

$${}^{9}P_{5} = \frac{9!}{(9-5)!} = 9!/4! = 362\,880/24 = 15\,120$$

(b) How many 5-digit numbers greater than 5 000 can be formed if repetition is allowed?

The number of 5-digit numbers with repetition = 9^5 = 59 049

5-digit numbers greater than 5 000 = 59 049 – 5 001 = 54 048

5 Given that $P(A) = 0.65$, $P(B) = 0.36$ and $P(A \cap B) = 0.234$. Show whether

(a) events A and B are mutually exclusive
Since $P(A \cap B) \neq 0$, A and B are not mutually exclusive.

(b) events A and B are independent
$P(A) \times P(B) = 0.65 \times 0.36 = 0.234 = P(A \cap B)$, hence A and B are independent.

6 Consider two events G and B with the following probabilities:
$P(G) = 0.64$, $P(B) = 0.33$ and $P(G \cup B) = 0.84$

(a) Determine the probability of 'not G and not B'.
$P(G' \cap B') = 1 - P(G \cup B) = 1 - 0.84 = 0.16$

(b) Describe whether the two events are independent.
$P(G \cap B) = P(G) + P(B) - P(G \cup B) = 0.64 + 0.33 - 0.84 = 0.13$
$P(G) \times P(B) = 0.64 \times 0.33 = 0.2112 \neq P(G \cap B)$, hence the events are not independent.

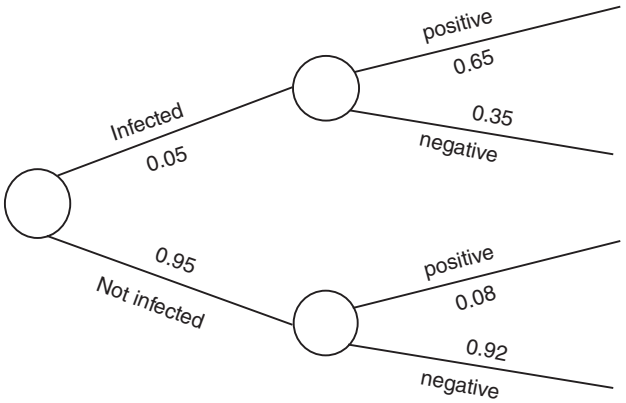
7 A restaurant offers a lunch set at $9.50 that consists of one main course, one dessert and one drink which can be selected from the menu below:
Main course: Fried Rice, Fried Noodles and Porridge
Dessert: Cake, Pudding, Pie and Ice cream
Drink: Coffee, Tea, Juices, Soft drinks and Mineral water

(a) How many different sets of lunches are possible?
No. of possible different sets of lunches = $3 \times 4 \times 5 = 60$

(b) What is the probability if a customer chooses tea as the drink?
P(a customer chooses tea) = $1/5 = 0.2$

8 If a person is infected with dengue, a test will show a positive result with a probability of 0.65. However, if the person is not infected, the probability of a positive result is 0.08. It is estimated that 5% of the population is infected with dengue.

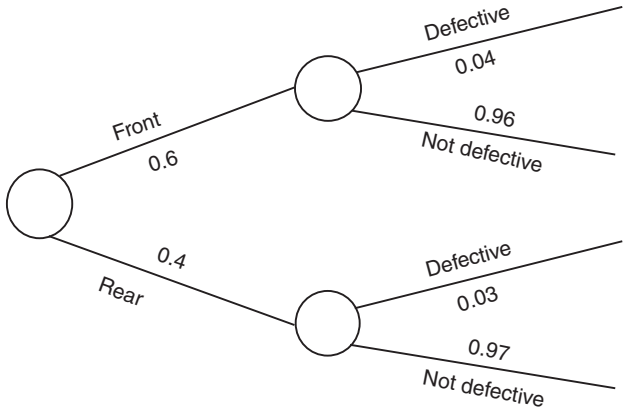
(a) Construct a tree diagram to illustrate the above problem.



(b) What is the probability that a person infected with dengue is tested negative?
P(negative given infected) = 0.35

9 An auto parts manufacturing company produces two types of car shock absorbers, 60% for front wheels and 40% for rear wheels. The finished shock absorbers are stored in one area. A shock absorber is randomly selected, and it is known that 4% of the front absorbers and 3% of the rear absorbers are defective.

(a) Construct a tree diagram to represent the problem.



(b) What is the probability of getting a defective shock absorber?

$P(\text{defective}) = P(\text{defective/front}) \times P(\text{front}) + P(\text{defective/rear}) \times P(\text{rear})$

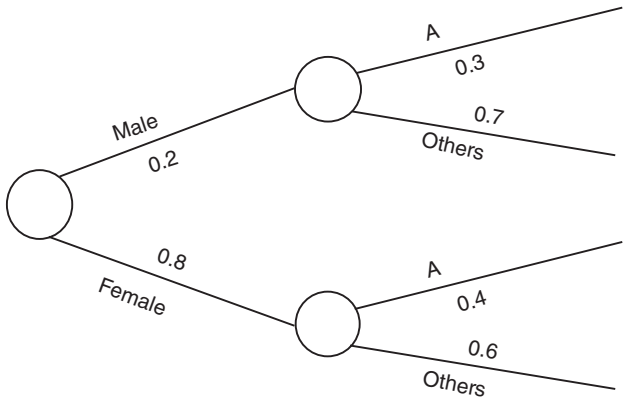
$= (0.04 \times 0.6) + (0.03 \times 0.4) = 0.24 + 0.12 = 0.36$

(c) If a shock absorber was inspected and found defective, what is the probability that it is a front wheel absorber?

$P(\text{front given defective}) = P(\text{front and defective})/P(\text{defective}) = 0.24/0.36 = 2/3$

10 A lecturer noticed that 20% of his students are males, while the rest are females. The probability that a male student obtained grade A in a final exam is 30%, while the probability that a female student obtained grade A in the exam is 40%.

(a) Illustrate the above problem using a tree diagram.



(b) Compute the probability that a particular student obtained grade A.

$P(\text{grade A}) = P(\text{A/male}) \times P(\text{male}) + P(\text{A/female}) \times P(\text{female})$

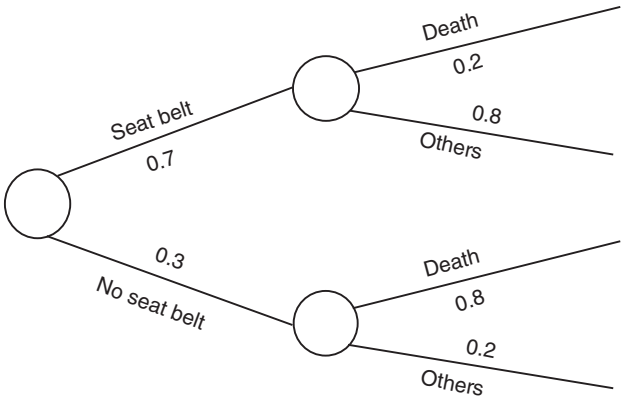
$= (0.3 \times 0.2) + (0.4 \times 0.8) = 0.06 + 0.32 = 0.38$

(c) If a student failed to obtain grade A, what is the probability that the student was a female?

$P(\text{female/other grades}) = P(\text{female \& other grades})/P(\text{other grades}) = 0.6 \times 0.8/0.62 = 0.77$

11 A study conducted by the Department of Road Safety has revealed that 70% of car drivers wear seat belts. The study also showed that 80% of driver deaths in serious road accidents were caused by not wearing the seat belts, while for those who wear seat belts the percentage of deaths is only 20%.

(a) Draw a tree diagram to represent the given situation.



(b) What is the probability that a car driver died in a serious road accident?

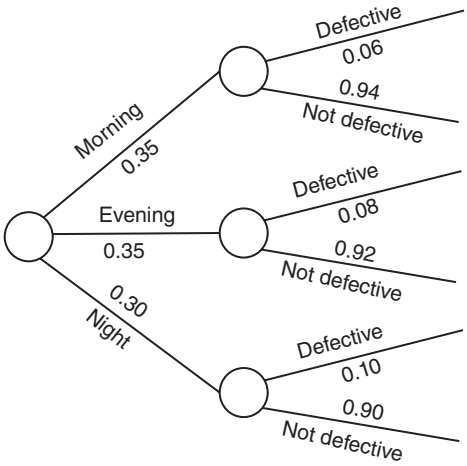
$P(\text{death}) = P(\text{death/seat belt}) \times P(\text{seat belt}) + P(\text{death/no seat belt}) \times P(\text{no seat belt})$
$= (0.2 \times 0.7) + (0.8 \times 0.3) = 0.14 + 0.24 = 0.38$

(c) If a car driver had died in a serious road accident, calculate the probability that the driver was not wearing the seat belt.

$P(\text{no seat belt/death}) = P(\text{no seat belt \& death})/P(\text{death}) = 0.24/0.38 = 0.63$

12 The manager of an electronic component manufacturer is concerned about the number of defects during each production shift. There are three shifts; morning, evening and night. Based on the past data, it was found that 35% of the components were produced during the morning shift, 35% during the evening shift and 30% during the night shift. It was also established that the percentages of the number of defective components produced during the three shifts are respectively 6%, 8% and 10%.

(a) Construct a tree diagram showing all the events in the problem.



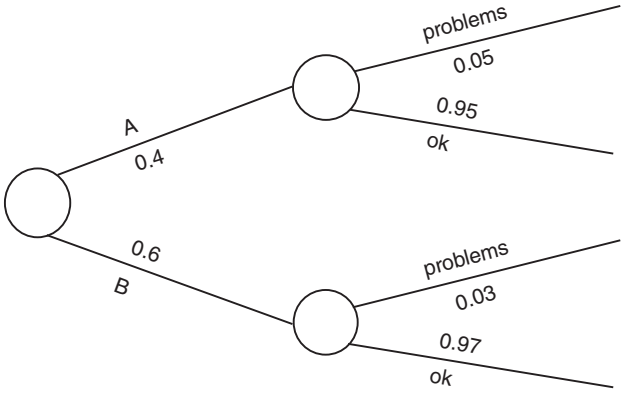
(b) Suppose a component randomly selected by the manager was defective. What is the probability that the component was produced during the evening shift?

$P(\text{evening given defective}) = P(\text{evening and defective})/P(\text{defective})$
$P(\text{evening and defective}) = P(\text{defective/evening}) \times P(\text{evening}) = 0.08 \times 0.35 = 0.028$

$P(\text{defective}) = P(\text{morning and defective}) + P(\text{evening and defective}) + P(\text{night and defective})$
$= (0.06 \times 0.35) + (0.028) + (0.10 \times 0.30) = 0.021 + 0.028 + 0.03 = 0.079$
$P(\text{evening given defective}) = 0.028/0.079 = 0.354$

13 A logistics firm requires additional trucks to meet increasing demand. The operations manager has decided to hire 40% of the trucks from company A and 60% from company B. It is known that 5% of the trucks from company A and 3% from company B are having engine problems.

(a) Draw an appropriate tree diagram to illustrate the above problem.



(b) Calculate the probability that the firm will get a truck with a bad engine.
$P(\text{bad engine}) = P(\text{problems \& A}) + P(\text{problems \& B})$
$= (0.05 \times 0.4) + (0.03 \times 0.6) = 0.02 + 0.018 = 0.038$

(c) If a truck hired by the firm has a good engine, calculate the probability that it came from company B.
$P(\text{B given good engine}) = P(\text{B and OK})/P(\text{OK})$
$P(\text{B and OK}) = P(\text{OK/B}) \times P(\text{B}) = 0.97 \times 0.6 = 0.582$
$P(\text{OK}) = P(\text{A and OK}) + P(\text{B and OK}) = (0.95 \times 0.4) + 0.582 = 0.38 + 0.582 = 0.962$
$P(\text{B given good engine}) = 0.582/0.962 = 0.605$

14 A faculty has three departments; Statistics, Actuarial Science and Operations Research. Each department consists of tutors, lecturers and senior lecturers, as shown in the table below.

	Statistics	Actuarial Science	Operations Research
Tutor	10	5	7
Lecturer	7	3	5
Senior Lecturer	3	2	3

Six persons must be selected at random for a short course.

(a) Find the number of ways that at least two from Operations Research were chosen.
Consider a binomial random variable with $p = P(\text{OR}) = 15/45 = 1/3$
$P(X \geq 2) = 1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1)$
$= 1 - (1/3)^0(2/3)^6 + 6(1/3)^1(2/3)^5 = 1 - 0.088 - 0.263 = 0.649$

(b) Find the probability that two senior lecturers were chosen.
Consider a binomial random variable with $p = P(\text{Senior}) = 8/45 = 0.1777778$
$P(X = 2) = 60(8/45)^2(37/45)^4 = 0.867$

15 The employees of an accounting firm were classified according to their levels of education and gender, as summarized in the table below.

Gender	Level of Education			
	College Certificate	Diploma	First Degree	Master Degree
Male	5	7	2	1
Female	8	10	5	2

(a) Find the probability that a randomly selected employee is a female with diploma.
P(female with diploma) = 10/40 = 1/4 = 0.25

(b) It is known that the employee is a male, find the probability that he has only a college certificate.
P(college certificate/male) = P(college certificate and male)/P(male)
= (5/40)/(15/40) = 5/15 = 1/3

SOLUTION MANUAL

CHAPTER 3 Sampling Methods and Sampling Distribution

1 Determine the sampling technique used in each of the following:

(a) A survey on dengue is to be conducted in Shah Alam. The population is divided into medium-income and high-income areas. A random sample that is selected will represent 10% of the population.
Stratified Random Sampling

(b) A random sample of a particular product is obtained by selecting every 50^{th} item from an assembly line.
Systematic Random Sampling

2 Determine whether each of the following statements is TRUE or FALSE:

(a) In cluster sampling, the characteristic of the population units in the same group is heterogeneous. TRUE
(b) Simple random sampling is based on non-probability concept. FALSE
(c) Snowball sampling requires a sampling frame. FALSE
(d) In stratified sampling, one of the demerits is that this sampling technique is not totally random. TRUE
(e) The variable religion is an example of a qualitative variable. TRUE
(f) The weight is considered as a continuous variable. TRUE
(g) One advantage of the median is that it is unique. TRUE
(h) An outlier is an extremely high or an extremely low data value when compared with the rest of the data. TRUE

3 A lecturer wishes to study the examination results of students from the Faculty of Computer Science which consists of 30 classes. He intends to choose only eight classes and all the students from these eight classes will be selected.

(a) State the population for this study.
Examination results of all students in 30 classes from the Faculty of Computer Science

(b) State the variable and its type for this study.
Examination result – discrete variable

(c) State the sampling technique used for this study.
Cluster random sampling

4 A researcher is interested in the cumulative grade point averages (CGPA) of the students who enrolled in three different Master programs; namely M701, M702 and M703.

(a) State the population and variable of interest.
Population: CGPAs of all students in three different programs
Variable: CGPA (continuous variable)

(b) Determine the sampling frame.
A random sample will be selected from the population (CGPAs of students in three different Master programs).

(c) Describe an appropriate sampling technique that should be used for this study and give two reasons for using this technique.
The appropriate sampling technique is stratified random sampling — consider the three programs as three homogeneous subgroups, and then take a simple random sample from each subgroup.
Reasons: 1) The samples taken are not only representing the whole population but also the three subgroups, and
2) The stratified random sampling generally has more statistical precisions compared to simple random sampling.

(d) Suppose the researcher has determined the CGPA of each student of those three programs. Would this represent a census or sample? Give your reason.
A census since the researcher has all the CGPAs of the population (three programs).

5 Star Cruises, the Leading Cruise Line in Asia-Pacific, offers special rates to Malaysian senior citizens. The agency wishes to conduct a survey on the ages of this group of customers. A sample 150 senior citizens who took the offer is selected at random.

(a) Explain the population for this survey.
All senior citizens who took the offer

(b) State the sampling frame.
A random sample of 150 senior citizens is selected from the population of senior citizens who took the offer.

(c) Describe the variable of interest and state its type.
Age of senior citizen – discrete variable

(d) Identify the most appropriate sampling technique and describe how it should be carried out.
Use systematic random sampling method, randomly select 150 senior citizens who took the offer from the list; a random starting point is selected, and then every *k*th name in the list is selected for the sample.

(e) Give the best data collection technique for this survey and state the advantage(s) and disadvantage(s).
Run through the list and check the ages of the selected customers.
Advantage – faster and convenient
Disadvantage – no information on age available or wrongly stated by customers

6 A researcher is conducting a simple survey to study the demographic characteristics of students at a public university. There are about 10 000 students at the university and a sample of 2 000 will be chosen as respondents.

(a) Describe the sampling frame for the survey.
A random sample of 2 000 will be selected from a population of 10 000 students at a public university.

(b) Define the type of variable (qualitative or quantitative) for each of the following demographic characteristics:
(i) Race. Qualitative
(ii) Academic program. Qualitative
(iii) Cumulative grade point average. Quantitative
(iv) Number of academic awards received. Quantitative

(c) State one possible probability sampling technique to choose the 2 000 students if the researcher manages to obtain all the students' names from the university. Give one advantage of this technique.
Use systematic random sampling – it is faster and convenient

(d) Explain how to select the sample using the sampling technique stated in (c).
The university has the names (and records) of all 10 000 students from which the researcher may randomly select 2 000 students; for instance, sort the student names, then starting at number 5 of the sorted list, select every 5th student to gather 2 000 student records.

7 A general manager of a company selling air conditioners wishes to investigate the level of satisfaction among customers. He had asked his assistant to gather the information of 300 customers who had recently bought air conditioners of various brands from the company, as shown in the table below. A random sample of 90 customers will be selected.

Brand of Air Conditioner	Number of Customers
LG	30
Mitsubishi	60
Panasonic	90
Sharp	20
Toshiba	70
York	30

(a) State the population of interest.
The 300 customers who had recently bought air conditioners from the company

(b) Identify the variable of interest in the study.
Brand of air conditioner

(c) Describe the sampling method that should be used in the study.
Stratified random sampling – select 30% of customers for each brand

(d) Obtain the number of customers selected as samples for each brand.

Brand of Air Conditioner	Number of Customers as Sample
LG	9
Mitsubishi	18
Panasonic	27

(contd.)

Brand of Air Conditioner	Number of Customers as Sample
Sharp	6
Toshiba	21
York	9
Total	90

(e) Explain how to select customers for the Panasonic brand using systematic sampling.
Systematic random sampling on Panasonic brand – using the list of customers who had recently bought Panasonic air conditioners, starting at customer number 3 (for instance), select every 3rd customer to gather 27 customers as a sample.

8 In the IT industry, customer service is a crucial factor affecting computer sales. The management of a computer company is interested to determine the level of customer satisfaction with the services provided by their service centres. The company has 50 service centres in the Klang Valley area. A sample of 10 centres was selected at random. All customers who had purchased the computers at these 10 centres were selected for the study. A questionnaire was posted to each of these customers.

(a) Explain the objective of the study.
To determine the level of customer satisfaction with the services provided by the centres

(b) State the population for this study.
All customers who had purchased computers at 50 centres

(c) Describe the sampling technique used for this study and state one reason for using this technique.
Cluster random sampling – only 10 centres from 50 are randomly selected
Reason – to reduce the cost of sampling for a large geographic area

(d) Give two techniques that can be used to select the 10 centres.
Simple random sampling and systematic random sampling

(e) Questionnaires were posted to the customers. One disadvantage of this approach is the poor response rate. Suggest how the company could increase the response rate.
The company could offer the customers a special discount or free service charges next time they come for services to increase the response rate.

9 A large firm is considering implementing a new salary scheme and wishes to determine the proportion of employees that agree with the new policy. The firm has 20 branches located throughout Malaysia. A sample of five branches was selected and the opinions of all employees regarding the new scheme were obtained.

(a) Describe the population and the sample for the study.
All employees working with the firm at 20 branches

(b) Explain the sampling frame for the study.
Five branches randomly selected from 20 and all employees were sampled.

(c) What type of statistics was used in the study?
Inferential statistics – qualitative data

(d) Identify the type of variable and the scale of measurement used in the study.
Opinion – qualitative variable – ordinal scale of measurement

(e) Describe the sampling technique used in the study.
Cluster random sampling – five branches randomly selected from 20 and all employees were sampled

(f) If systematic random sampling was employed to select five branches from 20, explain how it will be conducted.
Systematic random sampling – list (number) the 20 branches, starting from number 4, for instance, select every 4^{th} branch (4^{th}, 8^{th}, 12^{th}, 16^{th}, 20^{th})

(g) Determine the most appropriate data collection method to be used in the study, and give one advantage and one disadvantage of using this method.
Stratified random sampling (select all 20 branches), and take a simple random sample from each branch.
Advantage – all branches are represented in the sample
Disadvantage – cost and time consuming to sample all branches

10 A cable TV company has established a new online customer support service to help the customers on any matter related to the company's products. The company wants to investigate the effectiveness of the new customer support service by selecting a sample of 1000 customers using the information available from the database. The categories of products and the percentage of customers subscribing each category are summarized as follows:

Category of Products	Percentage of Customers (%)
All Products	50
Movies/Sports/News	30
Movies/News	10
Sports/News	10

(a) State the population of interest for this study.
All customers available in the database.

(b) Describe the sampling frame.
Sample 1 000 customers using the database based on the percentage of customers subscribing each category of products.

(c) State the variable to be measured.
The effectiveness of the new customer support service – qualitative variable

(d) Recommend an appropriate sampling technique to be used in this study, and describe how it should be carried out.
Proportional stratified random sampling – randomly select a number of customers based on the specified percentage from each category to gather 1 000 (500, 300, 100, 100)

(e) State a suitable data collection method for this study, and give one advantage and one disadvantage.
Call all selected 1 000 customers – convenient but time consuming
Send questionnaire by email – cost saving but low response rate

SOLUTION MANUAL

CHAPTER 4 Random Variables and Probability Distributions

1 Which of the following random variables are discrete and which are continuous?

(a) The wages of workers in a manufacturing firm.
Continuous

(b) The time taken to complete a task.
Continuous

(c) The prices of mobile phones displayed at a phone shop.
Discrete

(d) The number of pumps at a petrol station.
Discrete

2 Data was collected over 100 days tabulating the number of room service calls in a budget hotel.

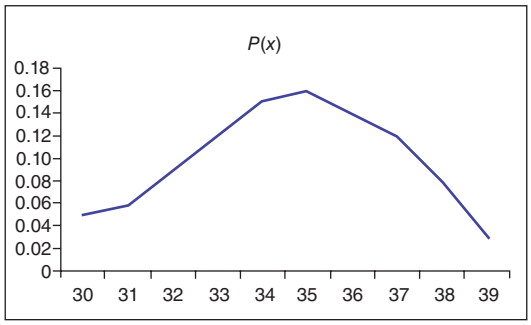
x	30	31	32	33	34	35	36	37	38	39
Frequency	5	6	9	12	15	16	14	12	8	3
P(x)										

(a) Use the relative frequency to calculate $P(x)$. What does each $P(x)$ represent?

x	*Frequency*	*P(x)*
30	5	0.05
31	6	0.06
32	9	0.09
33	12	0.12
34	15	0.15
35	16	0.16
36	14	0.14
37	12	0.12
38	8	0.08
39	3	0.03
Total	100	1.00

P(x) is the probability distribution for a random variable X (the number of room service calls in a budget hotel per day)

(b) Graph the probability distribution.



(c) Find the mean and variance of the number of calls.

x	$P(x)$	$x \cdot P(x)$	$(x-\mu)^2 \cdot P(x)$
30	0.05	1.50	1.053405
31	0.06	1.86	0.773286
32	0.09	2.88	0.603729
33	0.12	3.96	0.303372
34	0.15	5.10	0.052215
35	0.16	5.60	0.026896
36	0.14	5.04	0.278334
37	0.12	4.44	0.696972
38	0.08	3.04	0.930248
39	0.03	1.17	0.583443
Total	1.00	34.59	5.3019

Mean = $E(X) = \sum_x x \cdot P(x) = 34.59$ calls per day

Variance = $V(X) = E[(X-\mu)^2] = \sum_x (x-\mu)^2 P(x) = 5.3019$

3 Let X be a random variable with the following probability distribution. Determine $E(X)$, $E(X^2)$ and $V(X)$.

x	$P(x)$
1	0.1
2	0.2
3	0.3
4	0.3
5	0.1

x	$P(x)$	$x \cdot P(x)$	$x^2 \cdot P(x)$	$(x-\mu)^2 \cdot P(x)$
1	0.1	0.1	0.1	0.441
2	0.2	0.4	0.8	0.242
3	0.3	0.9	2.7	0.003
4	0.3	1.2	4.8	0.243
5	0.1	0.5	2.5	0.361
Total	1.0	3.1	10.9	1.29

$E(X) = \sum_x x \cdot P(x) = 3.1$

$E(X)^2 = \sum_x x^2 \cdot P(x) = 10.9$

$V(X) = \sum_x (x - \mu)^2 P(x) = 1.29$

4 Three coins are flipped at once, and let *X* be the number of tails. Find the expected value and the variance of *X*.

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
0	$(0.5)^3 = 0.125$	0	0.28125
1	$3(0.5)^3 = 0.375$	0.375	0.09375
2	$3(0.5)^3 = 0.375$	0.75	0.09375
3	$(0.5)^3 = 0.125$	0.375	0.28125
Total	1.00	1.5	0.75

Mean = $E(X) = \sum_x x \cdot P(x) = 1.5$ tails

Variance = $V(X) = \sum_x (x - \mu)^2 P(x) = 0.75$ tails

5 A particular product from an assembly line is known to have 5% defects. For a quality control process, 10 units are randomly selected from each production run. Let X denote the number of defectives among the 10 selected.

(a) Find the expected number of defectives and the standard deviation.

X is a binomial random variable with $p = 0.05$ and $n = 10$

Expected number of defectives = $\mu = E(X) = np = (10)(0.05) = 0.5$

Standard deviation = $\sigma = \sqrt{np(1-p)} = \sqrt{0.5(0.95)} = 0.475$

(b) If it costs \$15 to repair one defective product, calculate the expected repair cost every time a sample of 10 units is selected.

The expected repair cost every time a sample of 10 units is selected = 0.5(15) = \$7.50

6 A warehouse manager has established the following probability distribution for the daily shortage of a particular item.

y	0	1	2	3	4	5
$P(y)$	0.10	0.30	0.35	0.10	0.10	0.05

The penalty cost incurred per unit shortage is fixed at \$25. Calculate the mean and variance of the daily shortage cost of the item.

y	$P(y)$	$y \cdot P(y)$	$(y - \mu)^2 \cdot P(y)$
0	0.10	0	0.38025
1	0.30	0.3	0.27075
2	0.35	0.7	0.000875
3	0.10	0.3	0.11025
4	0.10	0.4	0.42025
5	0.05	0.25	0.465125
Total	1.00	1.95	1.6475

$E(Y) = \sum_y y \cdot P(y) = 1.95$ units. The mean of the shortage cost $=1.95 \times 25 = \$48.75$

$V(Y) = \sum_y (y - \mu)^2 P(y) = 1.6475$ units.

The variance of the shortage cost $=1.6475 \times 25 = \$41.1875$

7 The proportion of time, *X*, that a manufacturing machine is in operation during an 8-hour shift is a random variable with the following density function.

$$f(x) = \begin{cases} 3x^2, 0 \le x \le 1 \\ 0, \text{elsewhere} \end{cases}$$

(a) Calculate the expected value $E(X)$ and the variance $V(X)$.

$$E(X) = \int_0^1 x \cdot f(x)dx = \int_0^1 3x^3 dx = \frac{3}{4}x^4 \Big|_0^1 = 3/4 = 0.75$$

$$V(X) = E(X^2) - \mu^2; E(X^2) = \int_0^1 x^2 \cdot f(x)dx = \int_0^1 3x^4 dx = \frac{3}{5}x^5 \Big|_0^1 = 3/5$$

Hence, $V(X) = 3/5 - (3/4)^2 = 0.0375$

(b) For the machine under study, the profit per shift is given by $Y = 100X - 8$. Calculate $E(Y)$ and $V(Y)$.

$E(Y) = E(100X - 8) = 100E(X) - 8 = 100(0.75) - 8 = 67$

$V(Y) = V(100X - 8) = 100^2 V(X) = 100^2(0.0375) = 375$

8 The proportion of time per day that a doctor at a clinic is busy is a random variable *X* with density function

$$f(x) = \begin{cases} x(ax^2 - 1), 0 \le x \le 1 \\ 0, \text{elsewhere} \end{cases}$$

(a) Find the value of a that makes $f(x)$ a probability density function.

$$\int_0^1 x(ax - 1)dx = \int_0^1 (ax^2 - x)dx = \left(\frac{1}{3}ax^3 - \frac{1}{2}x^2\right) \Big|_0^1 = a/3 - 0.5$$

For a probability density function, $a/3 - 0.5 = 1$, hence $a = 4.5$

(b) Calculate $E(X)$ and $V(X)$.

$$E(X) = \int_0^1 x \cdot x(4.5x - 1)dx = \int_0^1 (4.5x^3 - x^2)dx = \left(\frac{4.5}{4}x^4 - \frac{1}{3}x^3\right) \Big|_0^1 = 4.5/4 - 1/3 = 0.7917$$

$$V(X) = E(X^2) - \mu^2; E(X^2) = \int_0^1 x^2 \cdot x(4.5x - 1)dx = \int_0^1 (4.5x^4 - x^3)dx = \left(\frac{4.5}{5}x^5 - \frac{1}{4}x^4\right) \Big|_0^1$$

$$= 4.5/5 - 1/4 = 0.65$$

Hence, $V(X) = 0.65 - (0.7917)^2 = 0.0232$

(c) Find the probability that, on any particular day, the doctor is busy at least 65% of the time.

$$P(X \ge 0.65) = \int_{0.65}^1 (4.5x^2 - x)dx = \left(1.5x^3 - \frac{1}{2}x^2\right) \Big|_{0.65}^1$$

$$= (1.5 - 0.5) - (0.41194 - 0.21125) = 1 - 0.20069 = 0.799$$

9 For the coming semester, 50 new students will register at a college. The registrar of the college noted that each semester the rate of withdrawal has been 25%. Consider this problem as a binomial experiment.

(a) What are the two outcomes S and F?
S – a student withdraws from the college, F - a student remains at the college

(b) What are the values of n, p and q?
$n = 50$, $p = 0.25$, $q = 0.75$

(c) Compute the probability that in the coming semester 15 new students will withdraw from the college.
P(15 new students withdraw from the college) = $b(15; 50, 0.25)$
Using Poisson Approximation to Binomial, $\mu = (50)(0.25) = 12.5$
$P(X = 15) = P(X \leq 15) - P(X \leq 14) = 0.806 - 0.725 = 0.081$ (from Poisson tables)

10 A group of students consisting of 5 males and 10 females wish to select three representatives to attend a student conference. They do so by placing their names in a box and drawing three names.

(a) What is the probability that all three selected students are males?
Let X = no. of males; $P(X = 3) = b(3; 15, 1/3) = {}^{15}C_3\,(1/3)^3(2/3)^{12} = 0.13$

(b) What is the probability that at least two females were selected?
Let Y = no. of females; $P(Y \geq 2) = 1 - P(0) - P(1) = 1 - b(0; 15, 2/3) - b(1; 15, 2/3)$
$= 1 - (2/3)^0(1/3)^{15} - 15(2/3)^1(1/3)^{14}$
$= 1 - 0.000002 = 0.999998 = 1$

11 In an exam of 20 multiple-choice questions, each question is provided with five possible answers of which only one is correct. Suppose that you have no time to prepare for the exam, and you have no choice but to answer all questions by guessing.

(a) What is the probability that you will answer all questions correctly?
P(all 20 answers are correct) = $P(X = 20) = b(20; 20, 0.2) = (0.2)^{20}(0.8)^0 = 0.00$

(b) What is the probability that you will get zero mark?
P(all 20 answers are incorrect) = $P(X = 0) = b(0; 20, 0.2) = (0.2)^0(0.8)^{20} = 0.01153$

(c) What is the probability that you will pass the exam if the passing mark is 50%?
P(at least 10 answers are correct) = $P(X \geq 10) = 1 - P(X \leq 9)$;
$P(X \leq 9)$ for $b(x; 20, 0.2)$ from Binomial tables is 0.9974;
hence $P(X \geq 10) = 1 - 0.9974 = 0.0026$

12 Suppose an electric car requires at least three battery cells for its power. The probability that any one of these cells will fail is 0.10, and the cells operate and fail independently.

(a) Determine the minimum number of battery cells required so that the electric car can operate without failure.
P(the electric car operates without failure) = P(at least 3 battery cells operate)
$= P(X \geq 3) = 1 - P(X \leq 2)$; we need to determine n such that $P(X \geq 3) = 1$ or $P(X \leq 2) = 0$
From Binomial tables, with $p = 0.9$, the first $P(X \leq 2) = 0.0000$ occurs when $n = 8$.
Hence, the minimum number of battery cells required so that the electric car can operate without failure is 8.

(b) Find the minimum number of battery cells the car must have so that there is a 95% probability that it will be operational.

P(the electric car operates without failure) = P(at least 3 battery cells operate) = 0.95
We need to determine n such that $P(X \geq 3) = 0.95$ or $P(X \leq 2) = 0.05$

From Binomial tables, with $p = 0.9$, $P(X \leq 2) = 0.0523$ when $n = 4$ and $P(X \leq 2) = 0.0086$ when $n = 5$.

Hence, the minimum number of battery cells the car must have so that there is a 95% probability that it will be operational is 5.

13 The average number of cars sold by the Luxury Auto is three cars per day.

(a) What is the probability that exactly four cars will be sold tomorrow?
No. of cars sold per day is a Poisson r.v. with $\mu = 3$
P(exactly four cars will be sold tomorrow) = $e^{-\mu} \cdot \mu^x / x! = e^{-3} \cdot 3^4 / 4! = 0.168$

(b) What is the probability that at least three cars will be sold tomorrow?
P(at least 3 cars will be sold tomorrow) = $P(X \geq 3) = 1 - P(X \leq 2)$
From Poisson tables, with $\mu = 3$, $P(X \leq 2) = 0.423$, hence $P(X \geq 3) = 1 - 0.423 = 0.577$

14 The number of road accidents recorded on a freeway possesses a Poisson distribution with an average of three accidents per week.

(a) What is the probability that there will be no accident in a particular week?
$\mu = 3$; P(no accident per week) = $P(X = 0) = e^{-3} \cdot 3^0 / 0! = 0.0498$

(b) What is the probability that there will be at least three accidents in a particular week?
P(at least 3 accidents per week) = $P(X \geq 3) = 1 - P(X \leq 2)$
Same as in 13(b), $P(X \leq 2) = 0.423$, hence $P(X \geq 3) = 1 - 0.423 = 0.577$

(c) What is the probability that there will be exactly five accidents in a particular week?
P(exactly 5 accidents per week) = $P(X = 5) = e^{-3} \cdot 3^5 / 5! = 0.100819$

(d) Find the expected number of road accidents on the freeway per year if the weekly numbers of recorded accidents are independent.
Consider 52 weeks per year, let $Y = 52X$;
Since $E(X) = 3$, then $E(Y) = 52 \cdot E(X) = 52(3) = 156$ per year

15 Each month the Immigration Department had arrested an average of 2 500 illegal immigrants. Assuming that the numbers of monthly arrests are independent, determine the following:

(a) The probability that less than 2 000 illegal immigrants will be arrested in a particular month.
Poisson with $\mu = 2\,500$; using normal approx. $P(X < 2\,000) = P(Z < (2\,000 - 2\,500)/2\,500)$
$= P(Z < -0.2) = P(Z > 0.2) = 0.4207$

(b) The probability that at least 4 500 illegal immigrants will be arrested in a two month period.
$P(X \geq 4\,500) = P(Z \geq (4\,500 - 2\,500)/2\,500) = P(Z \geq 0.8) = 0.2119$

(c) The probability that exactly 3 000 arrests are made in a particular month.
$P(X = 3\,000) = P(X \leq 3000.5) - P(X \leq 2999.5) = P(Z \leq 0.2002) - P(Z \leq 0.1998)$
(using MS Excel) = 0.57934 – 0.57918 = 0.00616

16 Use the standard normal probability table to find the area under the standard normal curve between the following values.

(a) $z = 0$ and $z = 2.3$
$P(0 < Z < 2.3) = P(Z > 0) - P(Z > 2.3) = 0.5 - 0.0107 = 0.4893$

(b) $z = 0$ and $z = 1.68$
$P(0 < Z < 1.68) = P(Z > 0) - P(Z > 1.68) = 0.5 - 0.0465 = 0.4535$

(c) $z = 0.24$ and $z = 0.33$
$P(0.24 < Z < 0.33) = P(Z > 0.24) - P(Z > 0.33) = 0.4052 - 0.3707 = 0.0345$

(d) $z = -2.575$ and $z = 0$
$P(-2.575 < Z < 0) = P(0 < Z < 2.575) = P(Z > 0) - P(Z > 2.575) = 0.5 - 0.005 = 0.495$

(e) $z = -2.81$ and $z = -1.35$
$P(-2.81 < Z < -1.35) = P(1.35 < Z < 2.81) = P(Z > 1.35) - P(Z > 2.81) = 0.0885 - 0.0025 = 0.086$

(f) $z = -1.73$ and $z = 0.49$
$P(-1.73 < Z < 0.49) = P(0 < Z < 1.73) + P(0 < Z < 0.49) = 0.5 - P(Z > 1.73) + 0.5 - P(Z > 0.49)$
$= 1 - 0.0418 - 0.3121 = 0.6461$

17 Use the standard normal probability table to find the value of z for each of the following.

(a) $P(0 \le Z \le z) = 0.41$ or $P(Z > z) = 0.09$, hence $z = 2.365$

(b) $P(Z \ge z) = 0.25$ hence $z = 0.6745$

(c) $P(Z \le z) = 0.95$ or $P(Z > z) = 0.05$, hence $z = 1.645$

(d) $P(-z \le Z \le z) = 0.88$ or $P(Z > z) = 0.06$, hence $z = 1.555$

18 If X is a normal random variable with a mean of 50 and a standard deviation of 8, how many standard deviations away from the mean is each of the following values of X?

(a) $x = 52$ $|52 - 50|/8 = 0.25$ standard deviation

(b) $x = 35$ $|35 - 50|/8 = 1.875$ standard deviation

(c) $x = 64$ $|64 - 50|/8 = 1.75$ standard deviation

(d) $x = 37$ $|37 - 50|/8 = 1.625$ standard deviation

19 An average tire used by a transportation company lasts 20 000 km with a standard deviation of 100 km. Assuming that the distance of the tire is normally distributed, what is the probability that a tire used by the company will last at most 30 000 km?

$P(X \le 30\,000) = P(Z \le [30\,000 - 20\,000]/100) = P(Z \le 100) = 1.00$

20 Suppose scores on an examination are normally distributed. If the examination has a mean of 60% and a standard deviation of 15%, what is the probability that a student who takes the examination will score between 80% and 90%?

$P(80 \le X \le 90) = P([80 - 60]/15 \le Z \le [90 - 60]/15) = P(1.333 \le Z \le 2)$
$= P(Z > 1.333) - P(Z > 2) = 0.0913 - 0.0228 = 0.0685$

21 A medical expert wants to study the effectiveness of a pain killer that has been used for many years. He found that the percentage of effectiveness of the drug is normally distributed with a standard deviation of 5%. Suppose he takes a sample of 15 patients who had taken the drug, find the probability that the sample mean percentage will be within 3% of the population mean.

P(sample mean percentage within 3% of population mean)

$$= P(|\bar{x} - \mu| \le 3) = P\left(|z| \le \frac{3}{5/\sqrt{15}}\right) = P(|z| \le 2.324) = P(-2.324 \le z \le 2.324)$$
$$= 1 - 2P(z > 2.324) = 1 - 2(0.0101) = 0.9798$$

SOLUTION MANUAL

CHAPTER 5 Descriptive Statistics: Describing, Exploring and Comparing Data

1 Construct a stem and leaf chart for the following data:

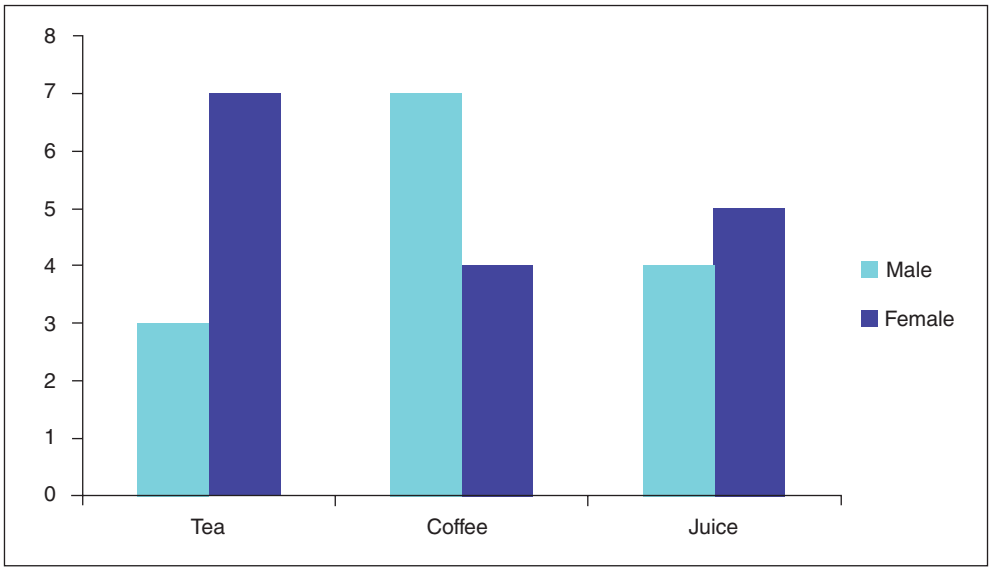
74	94	25	45	65	12	86	23	83	10
72	26	62	87	18	49	47	64	93	59

1	0	2	8
2	3	5	6
4	5	7	9
5	9		
6	2	4	5
7	2	4	
8	3	6	7
9	3	4	

2 Thirty students attending a seminar were asked to choose one of three choices of beverages they preferred, and the results were summarized as follows:

Beverage	Male	Female
Tea	3	7
Coffee	7	4
Juice	4	5

Illustrate the above data using a multiple bar chart and briefly comment on the chart.

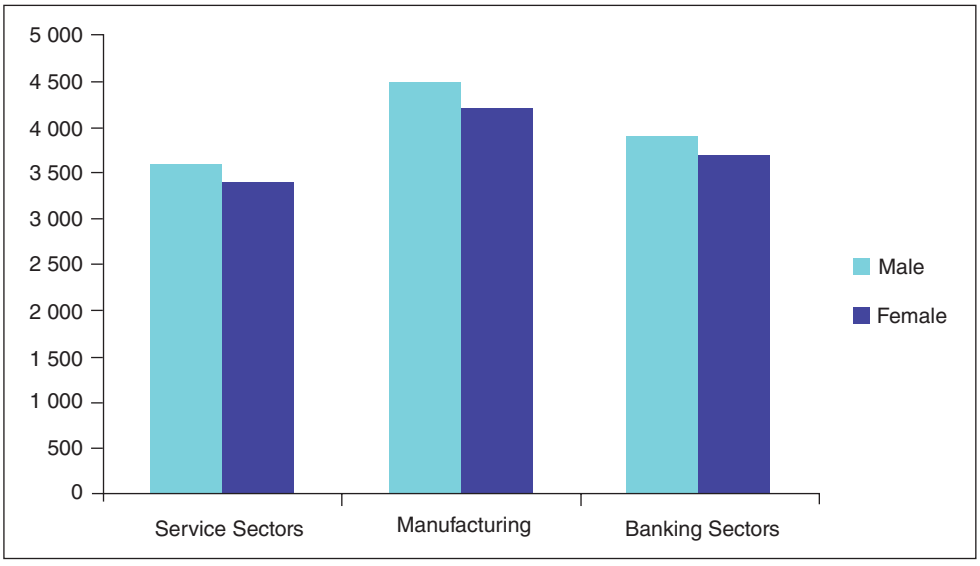


The number of females who preferred tea is twice that of males.
The number of males who preferred coffee is twice that of females.
The numbers of males and females who preferred juice are about the same.

3 Draw an appropriate bar chart to present the following information:

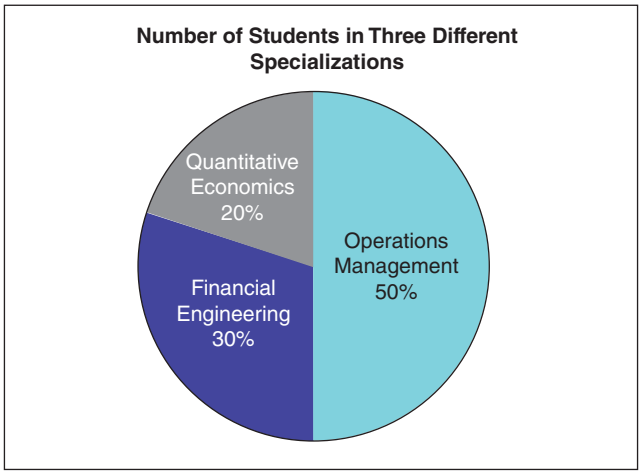
Average Monthly Salary of Operations Manager ($)

Sex	Service Sectors	Manufacturing	Banking Sectors
Male	3 600	4 500	3 900
Female	3 400	4 200	3 700



4 The following table shows the number of students registered for MSc Quantitative Science in three different specializations. Construct a pie chart showing the percentage of each specialization.

Specialization	Number of Students
Operations Management	25
Financial Engineering	15
Quantitative Economics	10



5 The average number of hours spent on a computer per day by nine computer programmers are given below:

8 7 9 6 7 10 11 9 12

Calculate the mean, median and mode.

Mean = (8 + 7 + 9 + 6 + 7 + 10 + 11 + 9 + 12)/9 = 8.778
Median = 9
Mode = 7 and 9

6 For ten consecutive weeks, a bank manager recorded the number of clients at the bank as follows:

212 210 208 216 205 210 213 207 209 220

(a) Compute the measures of central tendency (mean, median and mode).
Mean = (212 + 210 + 208 + 216 + 205 + 210 + 213 + 207 + 209 + 220)/10 = 211
Median = 210
Mode = 210

(b) From (a), describe the shape of the distribution.
Since the three measures are almost the same, the shape of distribution is nearly symmetrical.

7 1Malaysia Store has 15 outlets throughout Malaysia. The monthly sales ($'000) generated by the outlets during a festive season were as follows:

14.3 51.1 21.2 32.7 22.8 21.1 51.4 61.3
32.3 71.3 41.5 14.5 55.6 47.8 63.5

Find the mean, median and mode for the above data, and hence describe the distribution.

Mean = (14.3 + 51.1 + 21.2 + 32.7 + 22.8 + 21.1 + 51.4 + 61.3 + 32.3 + 71.3 + 41.5 + 14.5 + 55.6 + 47.8 + 63.5)/15
= 40.16 ($'000)
Median = 41.5 ($'000)
Mode = not available

Since mean is less than median, the distribution is negatively skewed (skewed to the left).

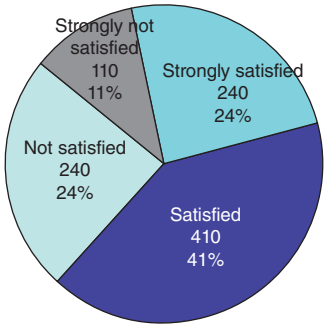
8 A random sample of 1 000 customers was asked about their level of satisfaction of a new laundry detergent. The responses were summarized as in the table below.

Level of Satisfaction	Number of Customers
Strongly satisfied	240
Satisfied	410
Not satisfied	240
Strongly not satisfied	110

(a) Define the variable of interest and its type.
Level of satisfaction – Qualitative variable

(b) Determine the type of measurement scale used.
Ordinal level measurement scale

(c) State the most appropriate chart to represent this data.
Pie chart

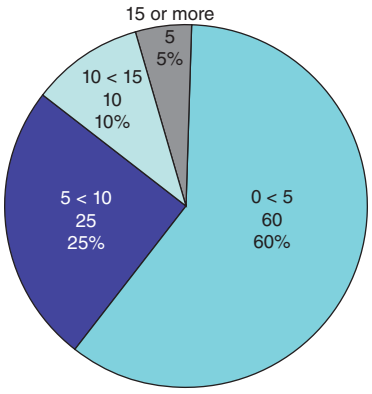


(d) State the most appropriate measure of central tendency for this data.
Mode (since ordinal level and categorical)

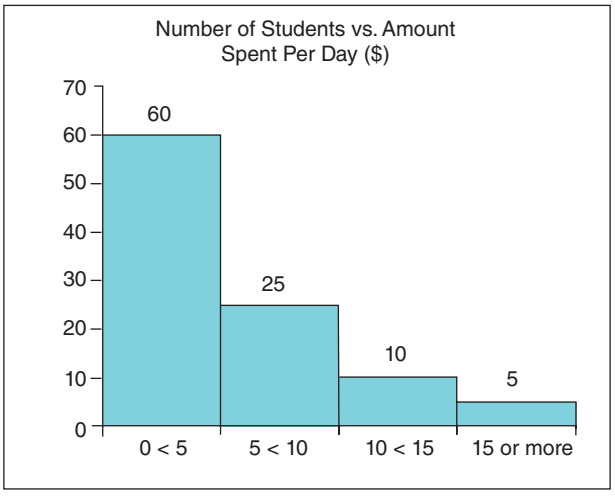
9 The table below shows the amount of money spent at school per day (in $) by a sample of 100 secondary school students.

Amount Spent Per Day ($)	Number of Students
0 < 5	60
5 < 10	25
10 < 15	10
15 or more	5

(a) Construct a pie chart for the given data.



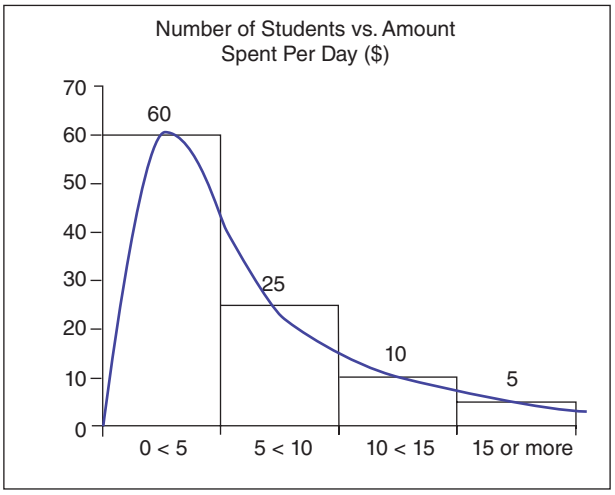
(b) Construct a histogram for the given data.



(c) Calculate the mean, median and mode.
Mean and median cannot be determined since there is an open-ended class.

$$\text{Mode} = \hat{x} = 0 + \left[\frac{60}{60 + 35}\right]5 = 3.158$$

(d) Sketch the distribution and describe its shape.
The distribution can be sketched using a histogram and the peak is at mode = 3.158 (skewed to the right or positively skewed).



10 A sample of 40 workers at a manufacturing company was randomly selected, and their monthly salaries were summarized as in the following table.

Monthly Salary ($)	Number of Workers
500 and less than 1 000	7
1 000 and less than 1 500	12
1 500 and less than 2 000	8
2 000 and less than 2 500	5
2 500 and less than 3 000	4
3 000 and less than 3 500	3
3 500 and less than 4 000	1

(a) Compute the mean and interpret.

Midpoint ($) ($x$)	Frequency (f)	$f \cdot x$
750	7	5 250
1 250	12	15 000
1 750	8	14 000
2 250	5	11 250
2 750	4	11 000
3 250	3	9 750
3 750	1	3 750
Total	40	70 000

Mean $= \bar{x} = \sum_{i=1}^{k} f_i x_i / n =$ 70 000/40 = 1 750

The mean or average monthly salary of 40 workers is $1 750.

(b) Compute the variance of the monthly salary.

Midpoint ($) ($x$)	Frequency (f)	$f \cdot x$	$f \cdot x^2$
750	7	5 250	3 937 500
1 250	12	15 000	18 750 000
1 750	8	14 000	24 500 000
2 250	5	11 250	25 312 500
2 750	4	11 000	30 250 000
3 250	3	9 750	31 687 500
3 750	1	3 750	14 062 500
Total	40	70 000	148 500 000

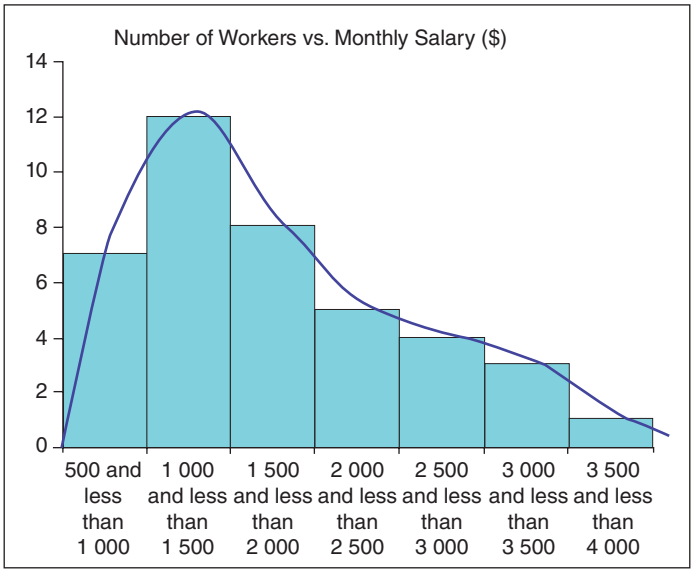
$$\text{Variance} = s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i x_i\right)^2\right] = (1/39)[148\,500\,000 - (1/40)(70\,000)^2]$$
$$= 666\,524.966$$

(c) Determine the modal value and interpret.
The modal class is '1 000 and less than 1 500'

$$\text{Mode} = \hat{x} = 1\,000 + \left[\frac{5}{5+4}\right]500 = 1\,277.78$$

Majority of the workers have monthly salary about $1 277.78.

(d) Draw a histogram to illustrate the data, and describe the shape of the distribution.



The shape of the distribution is skewed to the right (positively skewed).

(e) The mean and standard deviation of monthly salary of another manufacturing company are $1 500 and $750, respectively. Determine which company offers a more stable monthly salary.
Calculate the coefficient of variation.

$$CV = \frac{s}{\bar{x}}(100) = (\sqrt{666\,524.966/1\,750}) \times 100 = 46.65\%$$

Another company, $CV = (750/1\,500) \times 100 = 50\%$
Hence, the current company has a more stable monthly salary.

11 The following table represents the monthly food expenditures for forty families located in a small town.

Food Expenditure ($ thousands)	Number of Families
1 up to 2	18
2 up to 3	10
3 up to 4	7
4 up to 5	3
5 up to 6	2

(a) Calculate the mean and standard deviation of the food expenditures.

Midpoint (x)	Number of Families (f)	$f \cdot x$	$f \cdot x^2$
1.5	18	27	40.5
2.5	10	25	62.5
3.5	7	24.5	85.75
4.5	3	13.5	60.75
5.5	2	11	60.5
Total	40	101	310

$$\text{Mean} = \bar{x} = \sum_{i=1}^{k} f_i x_i / n = 101/40 = 2.525 \text{ (\$'000)}$$

$$\text{Std. Dev.} = s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i x_i\right)^2\right]} = \sqrt{(1/39)[310 - (1/40)(101)^2]}$$

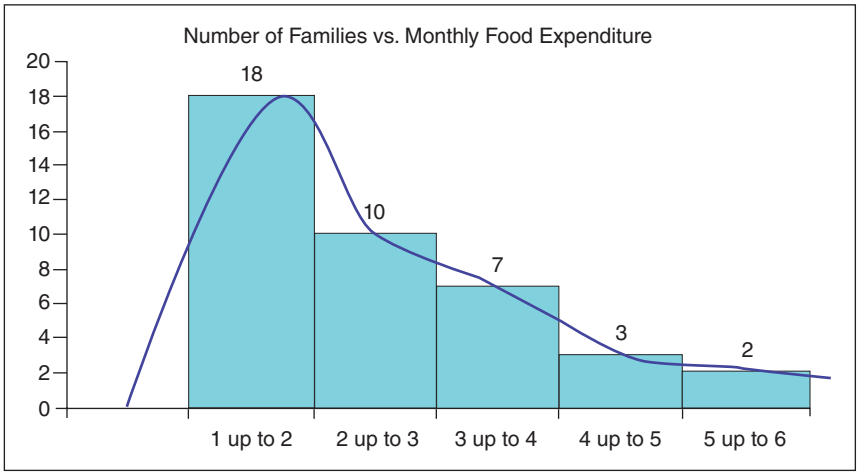
$$= 1.1873 \text{ (\$'000)}$$

(b) Determine the mode of the expenditures and explain its meaning.
The modal class is '1 up to 2'.

$$\text{Mode} = \hat{x} = 1 + \left[\frac{18}{18+8}\right]1 = 1.69231 \text{ (\$'000)}$$

Majority of the families spent around \$1 692.31 per month on food.

(c) Draw a histogram to illustrate the data, and describe the shape of the distribution.

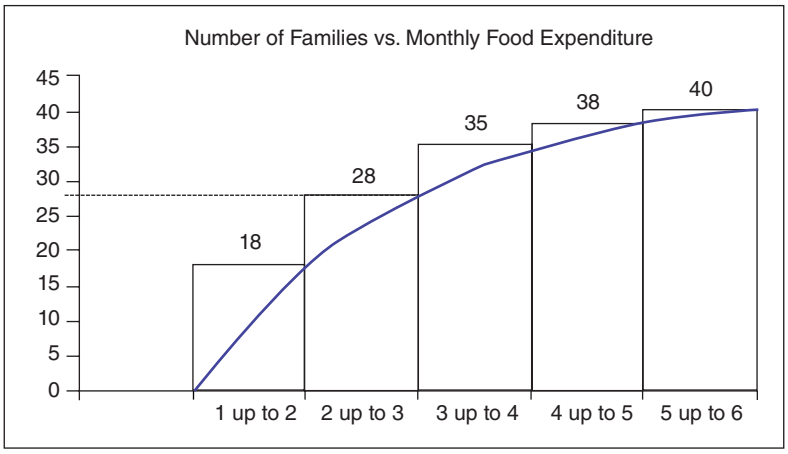


The shape of the distribution is skewed to the right (positively skewed).

(d) Draw a less-than ogive and estimate the percentage of families that spent more than \$3 000 on foods.

Food Expenditure (\$ thousands)	Number of Families	Less-than Cumulative Frequency
1 up to 2	18	18
2 up to 3	10	28
3 up to 4	7	35
4 up to 5	3	38
5 up to 6	2	40

Less-than ogive



The percentage of families that spent more than \$3 000 on foods is (40–28)/40 = 0.3 or 30%

(e) It was found that the mean and standard deviation of the monthly food expenditures of families in another town were \$3 600 and \$650, respectively. Using an appropriate measure, compare the dispersions of these two towns.
Calculate the coefficient of variation.

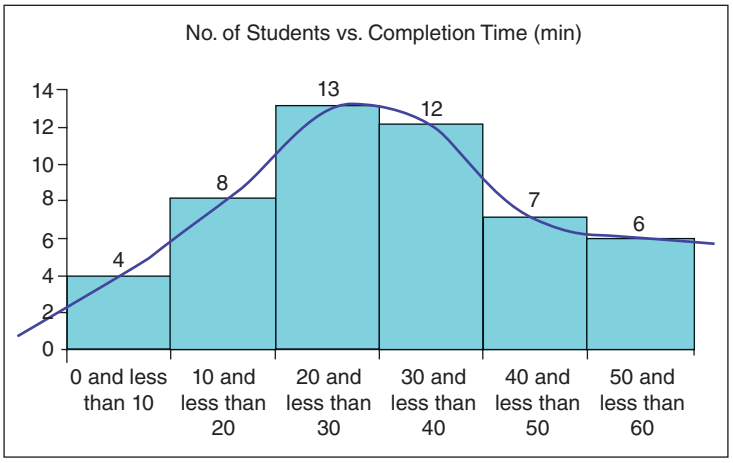
$CV = \frac{s}{\bar{x}}(100) = (1.1873 / 2.525) \times 100 = 47\%$

Another town, $CV = (650/3600) \times 100 = 18\%$
Hence, the monthly food expenditures of families in another town are less variable or more stable.

12 A lecturer is interested in determining the time taken by his students to complete a quiz. A random sample of 50 students is selected, and their completion times (in minutes) were summarized in the table below.

Completion Time (minutes)	Frequency
0 and less than 10	4
10 and less than 20	8
20 and less than 30	13
30 and less than 40	12
40 and less than 50	7
50 and less than 60	6

(a) Draw a histogram and describe the shape of the distribution.

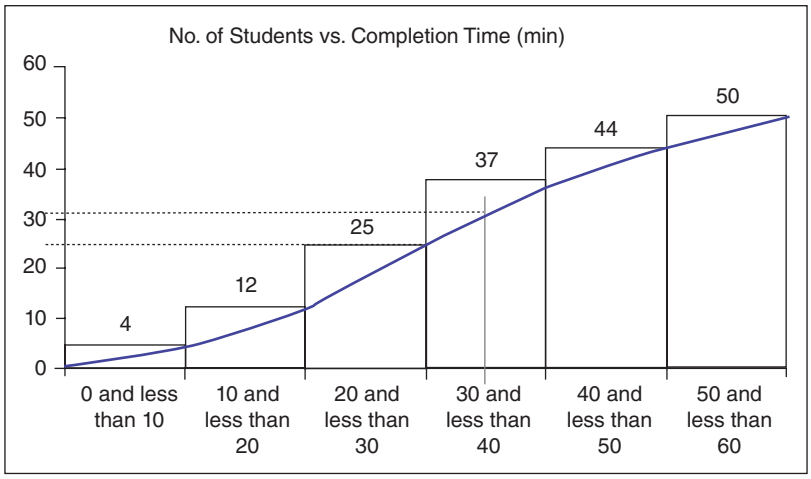


The shape of the distribution is slightly skewed to the left and almost bell-shaped.

(b) Construct a 'less-than' ogive and then estimate the median.

Completion Time (minutes)	Frequency	Less-than Cum. Freq.
0 and less than 10	4	4
10 and less than 20	8	12
20 and less than 30	13	25
30 and less than 40	12	37
40 and less than 50	7	44
50 and less than 60	6	50

Less-than ogive



The median is at 50% of the data (25 students); from ogive, it is estimated at 30 min.

(c) Using the ogive obtained in (b), determine the number of students who took more than 35 minutes to complete the quiz.
The number of students who took more than 35 minutes to complete the quiz is 50 – 32 = 18

(d) Compute the mean completion time and the standard deviation.

Midpoint (x) (minutes)	f	$f \cdot x$	$f \cdot x^2$
5	4	20	100
15	8	120	1 800
25	13	325	8 125
35	12	420	14 700
45	7	315	14 175
55	6	330	18 150
Total	50	1 530	57 050

$$\text{Mean} = \bar{x} = \sum_{i=1}^{k} f_i x_i / n = 1\,530/50 = 30.6 \text{ minutes}$$

$$\text{Std. dev.} = s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i x_i\right)^2\right]} = \sqrt{(1/49)[57\,050 - (1/50)(1\,530)^2]}$$
$$= 14.45 \text{ minutes}$$

(e) Compute the Pearson's coefficient of skewness and describe the shape of the distribution.

Pearson's coefficient of skewness, $r = \dfrac{3(\bar{x} - \tilde{x})}{s} = 3(30.6 - 30)/14.45 = 0.1246$

The shape of the distribution is slightly skewed to right and almost symmetrical.

13 A researcher had conducted a survey on three different groups of consumers (A, B and C) to determine the level of satisfaction for a particular product. The results (in percent) were summarized as follows:

Statistical Measures	A	B	C
Mean	85	65	65
Median	75	65	75
Mode	65	65	85
Standard deviation	12	8	10

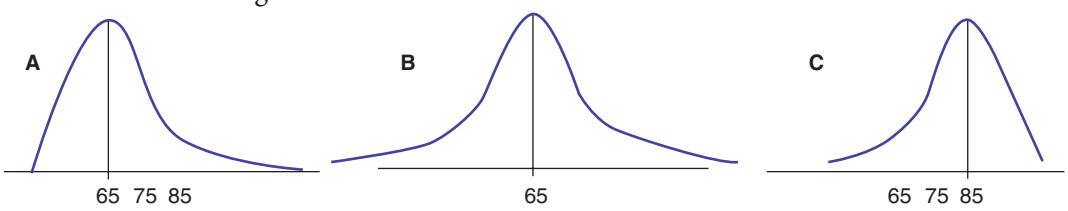
(a) Using an appropriate measure, determine which group has the most consistent level of satisfaction.

Calculate the coefficient of variation.

$CV_A = (12/85) \times 100 = 14.12\%$, $CV_B = (8/65) \times 100 = 12.31\%$, $CV_C = (10/65) \times 100 = 15.38\%$

Group B has the most consistent level of satisfaction.

(b) Sketch the distribution for each group using the mean, median and mode. Which distribution is skewed to the right?



Distribution A is skewed to the right.

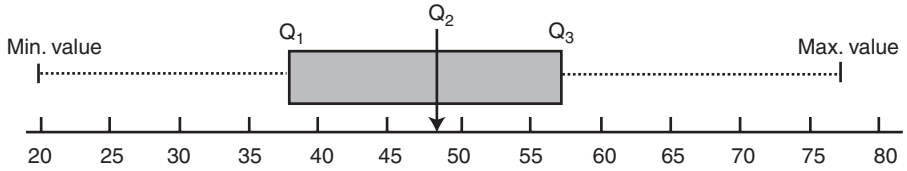
14 The following stem-and-leaf display describes the weekly entertainment expenses (in $′00) of a consultant firm.

2	0 1 4
3	0 2 4 7 8
4	0 1 3 5 7 7 8 9
5	1 3 4 5 5 6 8
6	2 3 4 6
7	1 5 7

(a) Determine the first, second and third quartiles.
First quartile (25% of 30 or 7.5) is between seventh and eighth values;
(37 + 38)/2 = 37.5 or $3 750
Second quartile (50% of 30 or 15) is the 15th value; 48 or $4 800
Third quartile (75% of 30 or 22.5) is between 22nd and 23rd values;
(56 + 58)/2 = 57 or $5 700

(b) Compute the interquartile range.
Interquartile range = (third quartile – first quartile) = 57 – 37.5 = 19.5 or $1 950

(c) Draw a box-plot representing the data. Comment on the skewness.



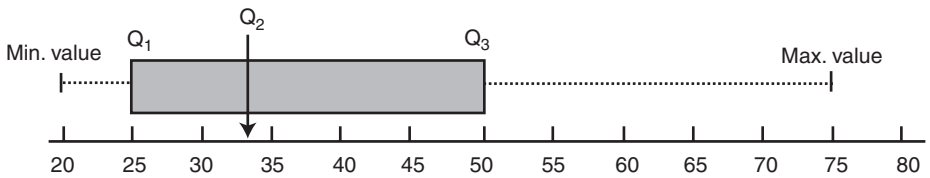
The skewness is close to zero (slightly positive).

15 The new management of Tower Hotel decided to analyze the number of calls per day at the receptionist counter during a two-week school holiday. From the collected data, the following were obtained.

Minimum Number of Calls	Quartile 1	Quartile 2	Quartile 3	Maximum Number of Calls
20	25	33	50	75

(a) What is the average number of calls per day did the hotel receive?
Since the distribution is highly skewed, the suitable average is the median.
Hence, the average number of calls per day is 33.

(b) Draw an appropriate diagram to illustrate the distribution of the number of calls received per day and comment on the skewness.



The distribution of the number of calls received per day is highly skewed to the right.

SOLUTION MANUAL

CHAPTER 6 Inferential Statistics: Estimation and Hypothesis Testing

1 Let X_1, X_2 and X_3 be independent, identically distributed random variables with mean μ and variance σ^2, and $A = (X_1 + X_2 + X_3)/3$. Is A an unbiased estimator of μ?

A is an un biased estimator of μ if $E(A) = \mu$.

$E(A) = E[(X_1 + X_2 + X_3)/3] = [E(X_1) + E(X_2) + E(X_3)]/3 = (m + m + m)/3 = \mu$.

Hence, A is an unbiased estimator of μ.

2 Suppose that X is a binomial random variable with parameters n and p. Is $\hat{p} = X/n$ an unbiased estimator of p?

$\hat{p} = X/n$ is an unbiased estimator of p if $E(\hat{p}) = p$.

$E(\hat{p}) = E(X/n) = E(X)/n = \mu/n$; substitute $\mu = np$ we have $E(\hat{p}) = np/n = p$.

Hence, $\hat{p} = X/n$ is an unbiased estimator of p.

3 From a random sample of 50 graduating students at a private college, the mean CGPA is 2.95 with a standard deviation of 3.25. Construct a 90% confidence interval for the mean CGPA for all graduating students at this college.

$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$; since $n = 50$ is large, replace σ with s;
$z_{0.10/2} = z_{0.05} = < 1.645$;

Thus, the 90% confidence interval for the mean CGPA is

$2.95 - 1.645(3.25)/\sqrt{50} < \mu < 2.95 + 1.645(3.25)/\sqrt{50}$
$= 2.194 < \mu < 3.706$ or $[2.194, 3.706]$.

4 Simple random samples of students are selected from two different faculties; 12 students from Faculty of Science and 16 students from Faculty of Business. The sample from Faculty of Science has an average CGPA of 2.75 with a standard deviation of 2.85, and the sample from Faculty of Business has an average CGPA of 2.95 with a standard deviation of 4.25. Determine the 95% confidence interval for the difference in CGPAs at the two faculties assuming that CGPAs in both faculties came from normal distributions.

[Confidence Interval for $\mu_1 - \mu_2$; $\sigma_1 \neq \sigma_2$ and unknown]

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

$$\upsilon = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]} = \frac{(0.6769 + 1.1289)^2}{[0.041654 + 0.08496]} = 25.75 \approx 26;$$

Hence, $t_{0.025}$ with d.f. = 26 is 2.056.

The 95% confidence interval for the difference in CGPAs is

$(2.75 - 2.95) - 2.056\sqrt{1.8058} < \mu_1 - \mu_2 < (2.75 - 2.95) + 2.056\sqrt{1.8058}$
$= -0.2 - 2.763 < \mu_1 - \mu_2 < -0.2 + 2.763 = -2.963 < \mu_1 - \mu_2 < 2.563$ or [−2.963, 2.563].

5 In a study on household income in a small town, 64 families were randomly selected. It is found that 24 families have monthly household incomes of less than $2 500. Find a 90% confidence interval for the percentage of all families in the town with household incomes less than $2 500 per month.

[Confidence Interval for proportion or percentage; large sample]

$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$; sample proportion = 0.375 or 37.5%, $z_{0.05} = 1.645$.

The 90% confidence interval for the percentage of families with < $2 500 per month is

$0.375 - 1.645\sqrt{\frac{0.375(0.625)}{64}} < p < 0.375 + 1.645\sqrt{\frac{0.375(0.625)}{64}} = 0.27545 < p < 0.47455,$
or in percentage [27.545%, 47.455%].

6 In a primary school, there are two standard six classes with 35 and 40 students, respectively. Within the first class, 18 students are female, whereas within the second class, 22 are female. Assuming that the data follows the normal distribution, find the 95% confidence interval of the difference between the female proportions of the two classes.

[Large-sample confidence interval for $p_1 - p_2$]

$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$; $z_{0.025} = 1.96$.

The 95% confidence interval of the difference between two female proportions is

$(0.5143 - 0.55) - 1.96\sqrt{\frac{0.5143(0.4857)}{35} + \frac{0.55(0.45)}{40}} < p_1 - p_2 < (0.5143 - 0.55) + 1.96\sqrt{\frac{0.5143(0.4857)}{35} + \frac{0.55(0.45)}{40}}$
$= -0.262 < p_1 - p_2 < 0.1906$, or [−0.262, 0.1906].

7 The mean kilometres travelled for a random sample of 20 cars in a car park is 54 275 km with a variance of 13 282. Construct a 97% confidence interval for the variance.

$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$; $\chi^2_{\alpha/2,n-1} = \chi^2_{0.015,19} = 34.742$; $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.985,19} = 8.159$;

The 97% confidence interval for the variance is

$\frac{(20-1)13\,282}{34.742} < \sigma^2 < \frac{(20-1)13\,282}{8.159} = 7263.77 < \sigma^2 < 30930.02$, or [7263.77, 30930.02].

8 Suppose two groups of adults were selected randomly and their heights were measured. The first group of 35 adults has mean height of 167 cm with a sample variance of 28.5 cm, whereas the second group of 30 has mean height of 164 cm with a sample variance of 32.4 cm. Find a 90% confidence interval for the true ratio of the two variances.

$$\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(\nu_1, \nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(\nu_1, \nu_2);$$

$f_{\alpha/2}(\nu_1, \nu_2) = f_{0.05}(34.29) = 1.8317, f_{\alpha/2}(\nu_2, \nu_1) = f_{0.05}(29.34) = 1.8020.$

The 90% confidence interval for the true ratio of the two variances is

$$\frac{28.5}{32.4}\frac{1}{(1.8317)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{28.5}{32.4}(1.8020) = 0.4802 < \frac{\sigma_1^2}{\sigma_2^2} < 1.5851, \text{ or } [0.4802, 1.5851].$$

9 A car manufacturer claimed that the average gas mileage of its new brand of hybrid car is 25 miles per liter. The gas mileages for ten randomly selected hybrid cars of the new brand are recorded as 24.4, 25.1, 22.6, 26.2, 25.3, 23.2, 21.9, 23.8, 24.5 and 25.0. Assume that the gas mileage is normally distributed. Does the sample data support the manufacturer's claim at the 0.01 significance level?

[Testing for the Population Mean: Small Sample, Population Variance Unknown]

1. Null and alternate hypotheses: H_0: $\mu = 25$ m/l versus H_1: $\mu \neq 25$ m/l.
2. Significance level: $\alpha = 0.01$.
3. Sample standard deviation, $s = 1.325$, sample mean = 24.2.

   Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{24.2 - 25}{1.325/\sqrt{10}} = -1.909.$
4. Decision rule: Critical region: $|t| > t_{0.005} = 3.25$ ($d.f. = 9$), i.e. reject H_0 if $|t| > 3.25$.
5. Decision: Since $|t| = 1.909 < 3.25$, do not reject H_0 and conclude that the sample data support the manufacturer's claim (25 m/l) at the 0.01 significance level.

10 A town uses thousands of street light bulbs each year. The brand of bulb the town currently uses has a mean life of 1 500 hours. A supplier claims that its new brand of street light bulbs (with the same cost) has a mean life of more than 1 500 hours. The mayor has decided to purchase the new brand if the test evidence supports the supplier's claim at the 0.025 level of significance. Suppose 100 bulbs of the new brand were tested with the following results: $\bar{x}$ = 1 620 hours and s = 162 hours. Will the mayor of the town purchase the new brand of street light bulbs?

[Testing for the Population Mean: Large Sample, Population Variance Unknown]

1. Null and alternate hypotheses: H_0: $\mu = 1\,500$ hours versus H_1: $\mu > 1\,500$ hours.
2. Significance level: $\alpha = 0.025$.
3. Test statistic: $z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{1\,620 - 1\,500}{162/\sqrt{100}} = 7.407.$
4. Decision rule: Critical region: $z > z_{0.025} = 1.96$, i.e. reject H_0 if $z > 1.96$.
5. Decision: Since $z = 7.407 > 1.96$, reject H_0 and agree that the new brand has a mean life of more than 1 500 hours—the mayor should purchase the new brand.

11 The average amount of time male and female college students spend playing computer games each day is believed to be the same. A random sample of 15 male students yield a mean of 5 hours with a standard deviation of 2.5 hours, whereas a random sample of 10 female students give a mean of 4 hours with a standard deviation of 3.5 hours. Is there a difference in the average amount of time male and female students play computer games each day? Test at the 5% level of significance.

[Testing for the Two Population Means: Small Samples, Population Variances Unknown]

1. Null and alternate hypotheses: H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$.
2. Significance level: $\alpha = 0.05$.
3. Test statistic: $s_p = \sqrt{\dfrac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{n_1 + n_2 - 2}} = \sqrt{\dfrac{(14)(6.25) + (9)(12.25)}{15 + 10 - 2}} = 2.9322,$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{((1/n_1) + (1/n_2))}} = \frac{(5 - 4) - 0}{2.9322\sqrt{((1/15) + (1/10))}} = 0.8354.$$

4. Decision rule: Critical region: $|t| > t_{0.025,23} = 2.069$, i.e. reject H_0 if $|t| > 2.069$.
5. Decision: Since $t = 0.8354 < 2.069$, do not reject H_0 and we are unable to conclude that there is a difference in the average amount of time male and female students play computer games each day.

12 A kitchen cabinet manufacturer is interested in comparing assembly times for two assembly processes. A random sample of five kitchen cabinets is selected and the assembly time (in minutes) on each cabinet of each process is recorded, as shown in the table below.

Cabinet	Process 1	Process 2	d_i
1	120	145	25
2	165	160	-5
3	95	90	-5
4	110	115	5
5	85	80	-5

Based on the data, can the manufacturer conclude, at the 10% level of significance, that the mean assembly times for the two processes differ?

[Testing for the Two Population Means: Paired Observations]

1. Null and alternate hypotheses: H_0: $\mu_D = 0$ versus H_1: $\mu_D \neq 0$.
2. Significance level: $\alpha = 0.10$.
3. Test statistic: $\bar{d} = 3$ and $s_d = 13.038$; $t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}} = \dfrac{3 - 0}{13.038/\sqrt{5}} = 0.5145.$
4. Decision rule: Critical region: $|t| > t_{0.05,4} = 2.132$, i.e. reject H_0 if $|t| > 2.132$.
5. Decision: Since $t = 0.5145 < 2.132$, do not reject H_0 and we are unable to conclude that the mean assembly times for the two processes differ at 10% level.

SOLUTION MANUAL

CHAPTER 7 *F*-Test and Analysis of Variance (ANOVA)

1 Using a two-tailed test and the 0.05 level of significance, what is the critical *F* value for a sample of nine observations in the numerator and ten in the denominator?

$F_{0.05/2,\ 9\text{-}1,\ 10\text{-}1} = F_{0.025,\ 8,\ 9} = 4.10$ (from *F*-distribution table).

2 The following hypotheses are given:

$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$.

The first random sample of ten observations gives a standard deviation of 15, while the second random sample of eleven observations gives a standard deviation of 12. At the 0.01 level of significance, test whether there is a difference in the variation of the two populations.

The appropriate test statistic is the *F*-distribution.

Since this is a two-tailed test, the significance level is 0.005, found by 0.01/2 = 0.005. There are $n_1 - 1 = 10 - 1 = 9$ degrees of freedom in the numerator, and $n_2 - 1 = 11 - 1 = 10$ degrees of freedom in the denominator. The critical value, from *F*-distribution table, is $F_{0.005,\ 9,\ 10} = 5.97$.

The ratio of the sample variances, $s_1^2 / s_2^2 = 15^2/12^2 = 1.5625$. Since this test statistic is less than the critical value, the null hypothesis is not rejected, and hence the variation in the two populations is the same.

3 A lecturer wishes to compare the test scores of his students from two different classes. The mean test score of 25 students in the first class is 75% with a standard deviation of 30%. The mean test score of 26 students in the second class is 77% with a standard deviation of 25%. At the 0.01 significance level, can he conclude that there is more variation in the first class?

$H_0 : \sigma_1^2 = \sigma_2^2$ (same variation) $\qquad H_1 : \sigma_1^2 > \sigma_2^2$ (more variation in the first class).

The appropriate test statistic is the *F*-distribution. This is a one-tailed test with $n_1 - 1 = 25 - 1 = 24$ degrees of freedom in the numerator, and $n_2 - 1 = 26 - 1 = 25$ degrees of freedom in the denominator. The critical value, from *F*-table, is $F_{0.01,\ 24,\ 25} = 2.62$.

The ratio of the sample variances is $30^2/25^2 = 1.44$. Since this test statistic is less than the critical value 2.62, the null hypothesis is not rejected, and hence the variation in the two classes is the same.

4 Five students were selected for a special program based on three different tests. The scores (in %) of the tests are summarized as follows:

Test 1	Test 2	Test 3
86	93	83
76	82	94
100	74	95
98	93	84

At the 0.05 significance level, test the hypothesis that the means of the three tests are equal.

(a) State the null and alternate hypotheses, and the decision rule?

$H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_1 : at least one mean is different.

There is k-1 = 3-1 = 2 degrees of freedom in the numerator, and there are 12 observations (three samples of 4 each). Therefore, there are N-k=12-3 = 9 degrees of freedom in the denominator. From F-table, the critical value is $F_{0.05,\,2,\,9} = 4.26$.

The decision rule is not to reject the null hypothesis H_0 if the F value is less than or equal to 4.26, or reject H_0 and accept H_1 if the F value is greater than 4.26.

(b) Compute SST, SSE, SS total, and complete an ANOVA table.

	Test 1		Test 2		Test 3		Total
	Score (%) x_1	Score squared x_1^2	Score (%) x_2	Score squared x_2^2	Score (%) x_3	Score squared x_3^2	
	86	7 396	93	8 649	83	6 889	
	76	5 776	82	6 724	94	8 836	
	100	10 000	74	5 476	95	9 025	
	98	9 604	93	8 649	84	7 056	
Col. total: T_c	360		342		356		1 058
Sample size: n_c	4		4		4		12
Sum of squares: x^2		32 776		29 498		31 806	94 080

$$\text{SST} = \sum\left[\frac{T_c^2}{n_c}\right] - \frac{(\sum x)^2}{N} = \left[\frac{360^2}{4} + \frac{342^2}{4} + \frac{356^2}{4}\right] - \frac{1\,058^2}{12} = 93\,325 - 93\,280.33 = 44.67.$$

$$\text{SSE} = \sum x^2 - \sum\left[\frac{T_c^2}{n_c}\right] = 94\,080 - 93\,325 = 755.$$

SS total = SST + SSE = 44.67 + 755 = 799.67.

ANOVA table:

Source of variation	(1) Sum of squares	(2) Degrees of freedom	(3) Mean square (1)/(2)
Between treatments	SST = 44.67	k – 1 = 3 - 1 = 2	SST/2 = 44.67/2 = 22.333
Error (within treatments)	SSE = 755	N – k=12 – 3 = 9	SSE/9 = 755/9 = 83.889
SS total	799.67		

(c) State your decision regarding the null hypothesis.

The test statistic, $F = \dfrac{SST/2}{SSE/(N-k)} = \dfrac{MSTR}{MSE} = \dfrac{22.333}{83.889} = 0.266$.

Since the test statistic $F = 0.266$ is less than the critical value of 4.26, H_0 is not rejected at the 0.05 level. Hence, the means of the three tests are equal.

5 An operations manager of a large firm is studying the monthly sales (in $'000) of three subsidiary companies over a six-month period as given in the table below.

Company A	Company B	Company C
152	148	150
135	158	148
130	136	126
142	138	128
125	140	140
128	127	135

At the 0.01 level of significance, can we conclude that there is a difference in the means of monthly sales of the three subsidiary companies over a six-month period?

(a) State the null and alternate hypotheses, and the decision rule?

$H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_1: there is a difference in the means of monthly sales.

There is k-1 = 3-1 = 2 degrees of freedom in the numerator, and there are 18 observations (three samples of 6 each). Therefore, there are N-k=18-3= 15 degrees of freedom in the denominator. From F-table, the critical value is $F_{0.01,\,2,\,15} = 6.36$.

The decision rule is not to reject H_0 if the test statistic (F value) is less than or equal to 6.36, or reject H_0 and accept H_1 if the F value is greater than 6.36.

(b) Compute SST, SSE, SS total and complete an ANOVA table.

	Company A		Company B		Company C		Total
	Sales ($'000) x_1	Sales squared x_1^2	Sales ($'000) x_2	Sales squared x_2^2	Sales ($'000) x_3	Sales squared x_3^2	
	152	23 104	148	21 904	150	22 500	
	135	18 225	158	24 964	148	21 904	
	130	16 900	136	18 496	126	15 876	
	142	20 164	138	19 044	128	16 384	
	125	15 625	140	19 600	140	19 600	
	128	16 384	127	16 129	135	18 225	
Col. total: T_c	812		847		827		2 486
Sample size: n_c	6		6		6		18
Sum of squares: x^2		110 402		120 137		114 489	345 028

$$\text{SST} = \sum\left[\frac{T_c^2}{n_c}\right] - \frac{(\sum x)^2}{N} = \left[\frac{812^2}{6} + \frac{847^2}{6} + \frac{827^2}{6}\right] - \frac{2\,486^2}{18} = 343\,447 - 343\,344 = 103.$$

$$\text{SSE} = \sum x^2 - \sum\left[\frac{T_c^2}{n_c}\right] = 345\,028 - 343\,447 = 1\,581.$$

SS total = SST + SSE = 103 + 1 581 = 1 684.

ANOVA table:

Source of variation	(1) Sum of squares	(2) Degrees of freedom	(3) Mean square (1)/(2)
Between treatments	SST=103	$k - 1 = 3 - 1 = 2$	SST/2 = 103/2 = 51.5
Error (within treatments)	SSE = 1 581	$N - k = 18 - 3 = 15$	SSE/15 = 1581/15 = 105.4
SS total	1 684		

(c) State your decision regarding the null hypothesis.

The test statistic, $F = \dfrac{SST/2}{SSE/15} = \dfrac{51.5}{105.4} = 0.489.$

Since the test statistic $F = 0.489$ is less than the critical value of 6.36, H_0 is not rejected. Hence, the means of monthly sales of the three companies are equal.

6 A manager of a company wishes to study the number of children belonging to each of his employees. The employees are divided according to four different departments and each department has different number of employees. The data obtained are as follows:

Department 1	Department 2	Department 3	Department 4
3	2	3	2
2	0	4	6
4	1	5	3
1	3	4	
5	2		
3	4		
2			

Test the hypothesis that the means number of children of employees at the four departments are equal at the 0.10 significance level.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. H_1 : the means number of children are not equal.

The degrees of freedom in the numerator is $k - 1 = 4 - 1 = 3$. Since there are 20 observations, the degrees of freedom in the denominator is $N - k = 20 - 4 = 16$. From F-table, the critical value is $F_{0.10,\,3,\,16} = 2.46$.

	Depart. 1		Depart. 2		Depart. 3		Depart. 4		Total
	No. of children x_1	No. children squared x_1^2	No. of children x_2	No. children squared x_2^2	No. of children x_3	No. children squared x_3^2	No. of children x_4	No. children squared x_4^2	
	3 2 4 1 5 3 2	9 4 16 1 25 9 4	2 0 1 3 2 4	4 0 1 9 4 16	3 4 5 4	9 16 25 16	2 6 3	4 36 9	
Col. total: T_c	20		12		16		11		59
Sample size: n_c	7		6		4		3		20
Sum of squares: x^2		68		34		66		49	217

$$\text{SST} = \sum\left[\frac{T_c^2}{n_c}\right] - \frac{(\sum x)^2}{N} = \left[\frac{20^2}{7} + \frac{12^2}{6} + \frac{16^2}{4} + \frac{11^2}{3}\right] - \frac{59^2}{20}$$

$$= (57.143 + 24 + 64 + 40.333) - 174.05 = 11.426.$$

$$\text{SSE} = \sum x^2 - \sum\left[\frac{T_c^2}{n_c}\right] = 217 - 185.476 = 31.524.$$

SS total = SST + SSE = 11.426 + 31.524 = 42.95.

ANOVA table:

Source of variation	(1) Sum of squares	(2) Degrees of freedom	(3) Mean square (1)/(2)
Between treatments	SST = 11.426	k – 1 = 4 – 1 = 3	SST/3 = 11.426/3 = 3.809
Error (within treatments)	SSE = 31.524	$N - k$ = 20 – 4 = 16	SSE/16 = 31.524/16 = 1.97025
SS total	42.95		

The test statistic, F = 3.809/1.97025 = 1.933. Since the test statistic is less than the critical value of 2.46, H_0 is not rejected. Hence, the means number of children of employees at the four departments are equal.

7 A fresh PhD graduate has job offers from four private universities. To decide which offer to take, he asked a sample of lecturers currently teaching at these universities about their monthly salaries. The information is summarized in the following table.

Monthly Salaries of Senior Lecturers (in $'000)			
University A	University B	University C	University D
2.5	4.5	2.8	3.0
2.8	3.3	4.2	2.5
3.5	2.2	3.6	5.0
4.0	2.5		

At the 0.10 level of significance, is there a difference in the mean salaries of lecturers among the four universities?

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. H_1 : the mean salaries of lecturers are not equal.

The degrees of freedom in the numerator is $k - 1 = 4 - 1 = 3$. The degrees of freedom in the denominator $N - k = 14 - 4 = 10$. From F-table, the critical value is $F_{0.10,\,3,\,10} = 2.73$.

	University A		University B		University C		University D		Total
	Salary ($'000) x_1	Salary squared x_1^2	Salary ($'000) x_2	Salary squared x_2^2	Salary ($'000) x_3	Salary squared x_3^2	Salary ($'000) x_4	Salary squared x_4^2	
	2.5	6.25	4.5	20.25	2.8	7.84	3.0	9	
	2.8	7.84	3.3	10.89	4.2	17.64	2.5	6.25	
	3.5	12.25	2.2	4.84	3.6	12.96	5.0	25	
	4.0	16	2.5	6.25					
Col. total: T_c	12.8		12.5		10.6		10.5		46.4
Sample size: n_c	4		4		3		3		14
Sum of squares: x^2		42.34		42.23		38.44		40.25	163.26

$$\text{SST} = \sum\left[\frac{T_c^2}{n_c}\right] - \frac{(\sum x)^2}{N} = \left[\frac{12.8^2}{4} + \frac{12.5^2}{4} + \frac{10.6^2}{3} + \frac{10.5^2}{3}\right] - \frac{46.4^2}{14}$$

$$= (40.96 + 39.0625 + 37.4533 + 36.75) - 153.783 = 0.4428.$$

$$\text{SSE} = \sum x^2 - \sum\left[\frac{T_c^2}{n_c}\right] = 163.26 - 154.2258 = 9.0342.$$

SS total = SST + SSE = 0.4428 + 9.0342 = 9.477.

ANOVA table:

Source of variation	(1) Sum of squares	(2) Degrees of freedom	(3) Mean square (1)/(2)
Between treatments	SST = 0.4428	$k - 1 = 4 - 1 = 3$	SST/3 = 0.4428/3 = 0.1476
Error (within treatments)	SSE = 9.0342	$N - k = 4 - 4 = 10$	SSE/10=9.0342/10= 0.90342
SS total	9.477		

The test statistic, $F = 0.1476/0.90342 = 0.163$. Since the test statistic is less than the critical value of 2.73, H_0 is not rejected. Hence, the mean salaries of lecturers among the four universities are equal.

SOLUTION MANUAL

CHAPTER 8 Chi-square Applications

1 Given the following contingency table, for the test of homogeneity, calculate the expected count (*E*) for all cells.

Treatment				Total
1	30	30	15	75
2	60	90	75	225
Total	90	120	90	300

Expected count E_{ij} = *Row Total i * Column Total j/Grand Total.*

E_{11}	75 × 90/300	22.5
E_{12}	75 × 120/300	30
E_{13}	75 × 90/300	22.5
E_{21}	225 × 90/300	67.5
E_{22}	225 × 120/300	90
E_{23}	225 × 90/300	67.5

2 The following table shows the number of students for three groups of a degree program that agreed or disagreed with a new curriculum.

	Student Group		
	1	2	3
Agreed	18	22	12
Disagreed	24	16	28

Can we conclude that the opinions of the three student groups regarding the new curriculum are homogeneous?

Test for homogeneity; H_0: three groups are homogeneous

	1	2	3	Row Total
1	18	22	12	52
2	24	16	28	68
Column Total	42	38	40	120

O	E	$(O-E)^2/E$
18	$52 \times 42/120 = 18.2$	0.002198
22	$52 \times 38/120 = 16.47$	1.856764
12	$52 \times 40/120 = 17.33$	1.639290
24	$68 \times 42/120 = 23.8$	0.001858
16	$68 \times 38/120 = 21.53$	1.420385
28	$68 \times 40/120 = 22.67$	1.253149
		6.173644

The critical value at $\alpha = 5\%$ level of significance with degrees of freedom $(2 - 1)(3 - 1) = 2$ is $\chi^2_2(0.05) = 5.99$. Since the test statistic 6.17 is greater than this critical value, reject H_0, and hence the three groups are not homogeneous.

3 Two dice are tossed 100 times with the following results:

x	2	3	4	5	6	7	8	9	10	11	12
f	3	5	11	10	8	23	15	10	9	4	2

At the 0.05 level of significance, can we conclude that the two dice are balanced?

Test of Goodness-of-fit;

H_0: the two dice are balanced (each side has equal chance of occurring, 1/6)

x	O	E	$(O-E)^2/E$
2	3	$1/36 \times 100 = 2.778$	0.0178
3	5	$2/36 \times 100 = 5.555$	0.0556
4	11	$3/36 \times 100 = 8.333$	0.8533
5	10	$4/36 \times 100 = 11.111$	0.1111
6	8	$5/36 \times 100 = 13.889$	2.4969
7	23	$6/36 \times 100 = 16.667$	2.4067
8	15	$5/36 \times 100 = 13.889$	0.0889
9	10	$4/36 \times 100 = 11.111$	0.1111
10	9	$3/36 \times 100 = 8.333$	0.0533
11	4	$2/36 \times 100 = 5.555$	0.4356
12	2	$1/36 \times 100 = 2.778$	0.2178
			6.848

The critical value at 5% level of significance with degrees of freedom = 11 – 1 = 10 is $\chi^2_{10}(0.05) = 18.307$. Since the test statistic 6.848 is less than this critical value, do not reject H_0, and hence the two dice are balanced.

4 A study claimed that the satisfaction level of residents of an apartment block regarding services provided by the management are as follows: 30% are highly satisfied; 25% are satisfied; 20% are not sure, 15% are dissatisfied; and 10% are highly dissatisfied. One year later, after improving the quality of all services, the management wants to know whether these percentages have changed or not. A random sample of 50 residents of the same apartment block were asked and found that 18 are highly satisfied; 15 are satisfied; 7 are not sure, 5 are dissatisfied and 5 are highly dissatisfied. At the 0.01 level of significance, conduct the Chi-square test for goodness-of-fit and state your conclusions.

Test of Goodness-of-fit; H_0: the percentages remain the same

Satisfaction level	O	E	$(O-E)^2/E$
Highly satisfied	18	15	0.6
Satisfied	15	12.5	0.5
Not sure	7	10	0.9
Dissatisfied	5	7.5	0.833
Highly dissatisfied	5	5	0
			2.833

The critical value at 1% level of significance with degrees of freedom = 5 – 1 = 4 is $\chi^2_4\,(0.01)$ = 13.2767. Since the test statistic 2.8333 is less than the critical value, do not reject H_0, and hence the percentages remain the same.

5 In a survey, 200 respondents (120 in Kuala Lumpur and 80 in Johor Bahru) were asked to select their favourite public transports for long-distance travels within Malaysia. Of the Kuala Lumpur respondents, 60 selected airplane, 42 selected express bus and 18 selected train. Of the Johor Bahru respondents, 25 selected airplane, 35 selected express bus and 20 selected train. At the 0.10 significance level, test whether the Kuala Lumpur and Johor Bahru respondents are independent of the public transports preference for long-distance travels?

Public Transport	KL	JB	Row Total
Airplane	60	25	85
Express bus	42	35	77
Train	18	20	38
Column Total	120	80	200

Test of Independence;

H_0: KL and JB respondents are independent of the public transports preference.

O	E	$(O-E)^2/E$
60	$85 \times 120/200 = 51$	1.588
25	$85 \times 80/200 = 34$	2.382
42	$77 \times 120/200 = 46.2$	0.382
35	$77 \times 80/200 = 30.8$	0.573
18	$38 \times 120/200 = 22.8$	1.011
20	$38 \times 80/200 = 15.2$	1.516
		7.451

The critical value at 10% level of significance with degrees of freedom = (3 – 1)(2 – 1) = 2 is χ^2_2 (0.10) = 4.60517. Since the test statistic 7.451 is greater than the critical value, reject H_0 and hence the KL and JB respondents are not independent of the public transports preference for long-distance travels.

6 A researcher wishes to know the opinions of the people in the three east coast states of Malaysia (Kelantan, Terengganu and Pahang) on three types of rice that are popular among them. A random sample of 300 respondents (100 in each state) is selected, and each of them was asked to choose one type of rice they most prefer. The results are given below:

	Nasi Dagang	*Nasi Berlauk*	*Nasi Kerabu*
Kelantan	23	46	31
Terengganu	48	27	25
Pahang	33	38	29

Conduct the suitable test and state your conclusion.

State	*Nasi Dagang*	*Nasi Berlauk*	*Nasi Kerabu*	Row Total
Kelantan	23	46	31	100
Terengganu	48	27	25	100
Pahang	33	38	29	100
Column Total	104	111	85	300

Test of Independence; test at 5% level of significance.

H_0: Respondents of the three states are independent of the type of rice they most prefer.

O	*E*	$(O-E)^2/E$
23	100 × 104/300 = 34.667	3.926
46	100 × 111/300 = 37	2.189
31	100 × 85/300 = 28.333	0.251
48	100 × 104/300 = 34.667	5.128
27	100 × 111/300 = 37	2.703
25	100 × 85/300 = 28.333	0.392
33	100 × 104/300 = 34.667	0.080
38	100 × 111/300 = 37	0.027
29	100 × 85/300 = 28.333	0.016
		14.712

The critical value at 5% level of significance with degrees of freedom = (3 – 1)(3 – 1) = 4 is χ^2_4 (0.05) = 9.48773. Since the test statistic 14.712 is greater than the critical value, reject H_0 and conclude that the respondents of the three states are not independent of the type of rice they most prefer.

7 A random sample of 70 families in a rural area was classified according to the monthly household income and the number of family members:

Monthly Household Income	Number of Family Members			
	2–5	6–9	10–12	More than 12
Less than $1 000	10	15	5	0
$1 001 to $2 000	8	7	3	2
$2 001 to $3 000	7	5	2	1
More than $3 000	3	2	0	0

Test the hypothesis, at the 0.01 significance level, that the size of a family is independent of the monthly household income.

Monthly Household Income	Number of Family Members				Row Total
	2–5	6–9	10–12	> 12	
Less than $1 000	10	15	5	0	30
$1 001 to $2 000	8	7	3	2	20
$2 001 to $3 000	7	5	2	1	15
More than $3 000	3	2	0	0	5
Column Total	28	29	10	3	70

Test of Independence;

H_0: The size of a family is independent of the monthly household income.

O	E	$(O-E)^2/E$
10	$30 \times 28/70 = 12.0000$	0.3333
15	$30 \times 29/70 = 12.4286$	0.5320
5	$30 \times 10/70 = 4.2857$	0.1190
0	$30 \times 3/70 = 1.2857$	1.2857
8	$20 \times 28/70 = 8.0000$	0.0000
7	$20 \times 29/70 = 8.2857$	0.1995
3	$20 \times 10/70 = 2.8571$	0.0071
2	$20 \times 3/70 = 0.8571$	1.5238
7	$15 \times 28/70 = 6.0000$	0.1667
5	$15 \times 29/70 = 6.2143$	0.2373
2	$15 \times 10/70 = 2.1429$	0.0095
1	$15 \times 3/70 = 0.6429$	0.1984
3	$5 \times 28/70 = 2.0000$	0.5000
2	$5 \times 29/70 = 2.0714$	0.0025
0	$5 \times 10/70 = 0.7143$	0.7143
0	$5 \times 3/70 = 0.2143$	0.2143
		6.0435

The critical value at 1% level of significance with degrees of freedom = $(4 - 1)(4 - 1) = 9$ is $\chi^2_9(0.01) = 21.666$. Since the test statistic 6.0435 is less than the critical value, do not reject H_0 and conclude that the size of a family is independent of the monthly household income.

SOLUTION MANUAL

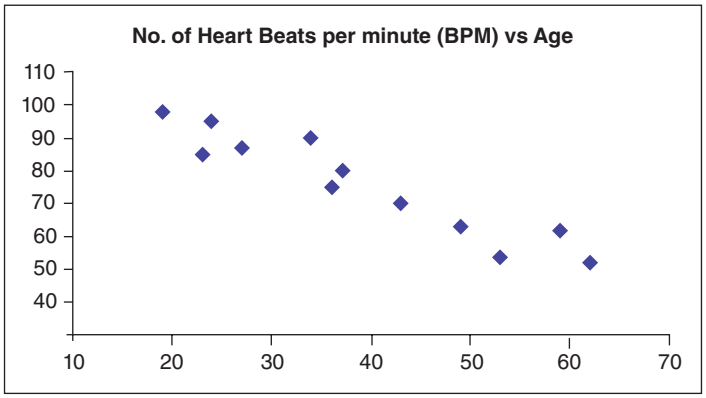
CHAPTER 9 Simple Linear Regression and Correlation

1 From a random sample of 12 outpatients, the following data were obtained:

Patient	1	2	3	4	5	6	7	8	9	10	11	12
Age (years)	36	34	49	43	23	37	24	53	27	19	62	59
No. of Heart Beats per minute (BPM)	75	90	63	70	85	80	95	54	87	98	52	62

Decide whether there is a correlation between the age and the BPM. Justify your answer.

Draw a scatter diagram, as obtained below:



There is a strong correlation between the Age and the BPM; as the Age increases, the BPM decreases—negative correlation.

2 A research was carried out to study the relationship between inflation rate and unemployment rate. The data were summarized as in the table below.

Inflation Rate	Unemployment Rate
0.7	4.5
1.5	9.0
2.4	9.5
2.8	10.5
3.5	6.5
3.5	6.8
4.5	4.0
6.5	5.5

(a) State the independent variable.
Inflation rate

(b) Compute the product moment coefficient of correlation for the data and comment on its value.

Use the formula $r = \dfrac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$;

Inflation Rate (X)	Unemployment Rate (Y)	X^2	XY	Y^2
0.7	4.5	0.49	3.15	20.25
1.5	9.0	2.25	13.5	81
2.4	9.5	5.76	22.8	90.25
2.8	10.5	7.84	29.4	110.25
3.5	6.5	12.25	22.75	42.25
3.5	6.8	12.25	23.8	46.24
4.5	4.0	20.25	18	16
6.5	5.5	42.25	35.75	30.25
25.4	56.3	103.34	169.15	436.49

$$r = \frac{8(169.15) - (25.4)(56.3)}{\sqrt{[8(103.34) - (25.4)^2][8(436.49) - (56.3)^2]}}$$

$= -76.82/241.8762 = -0.3176$; a weak negative correlation.

(c) Construct a least-squares regression equation of the two variables.
Regression equation: $Y' = a + bX$;

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{8(169.15) - (25.4)(56.3)}{8(103.34) - (25.4)^2} = -76.82/181.56 = -0.423;$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{56.3}{8} + 0.423\left(\frac{25.4}{8}\right) = 8.381.$$

The least-squares regression equation is $Y' = 8.381 - 0.423X$.

(d) Compute the coefficient of determination and explain its meaning.
The coefficient of determination is $r^2 = (-0.3176)^2 = 0.1009$; we can say that only 10.09% of the variation in the unemployment rate is explained by the variation in the inflation rate.

(e) Estimate the inflation rate for an unemployment rate of 7.0.
$Y' = 8.381 - 0.423X$ or $X = (Y' - 8.381)/-0.423$; if $Y' = 7.0$, then the inflation rate is estimated as $X = (7.0 - 8.381)/-0.423 = 3.265$.

3 A health researcher wishes to show that running on a treadmill can help reduce weight. Ten volunteers were asked to record the total number of minutes they ran on a treadmill for one week and the amount of weight they had lost during the week. The results are shown below.

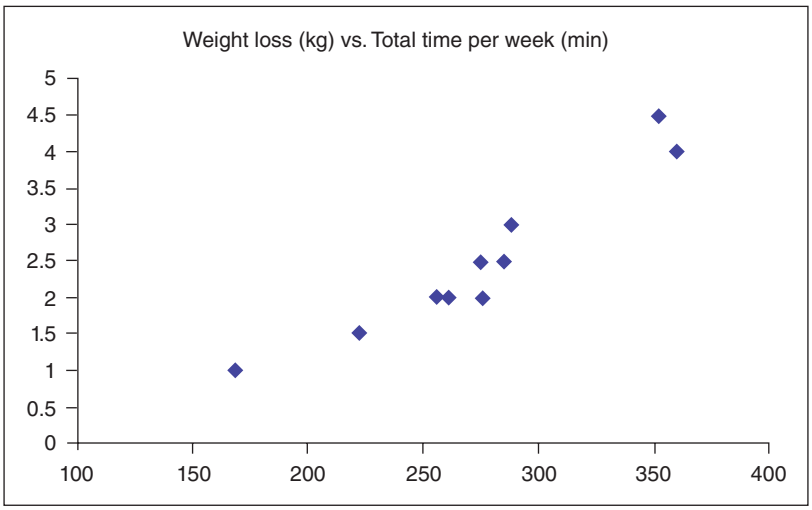
Weight loss (kg)	Total time per week (minutes)
1.0	168
2.5	285
2.0	256

(contd.)

Weight loss (kg)	Total time per week (minutes)
1.5	222
3.0	288
4.0	360
4.5	352
2.0	261
2.5	275
2.0	276

(a) State the dependent and independent variables.
Dependent variable: Weight loss
Independent variable: Total time per week treadmill.

(b) Sketch a scatter diagram of the data and describe the relationship of the two variables.
Scatter diagram:



There is a strong positive relationship between weight loss and total time per week on treadmill.

(c) Determine the product moment correlation coefficient between the total number of minutes on a treadmill per week and the amount of weight loss. Interpret your answer.

Use the formula $r = \dfrac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$;

Weight loss (kg) Y	Total time per week (min) X	X^2	XY	Y^2
1.0	168	28 224	168	1
2.5	285	81 225	712.5	6.25
2.0	256	65 536	512	4
1.5	222	49 284	333	2.25
3.0	288	82 944	864	9
4.0	360	129 600	1 440	16
4.5	352	123 904	1 584	20.25

(contd.)

Weight loss (kg) Y	Total time per week (min) X	X^2	XY	Y^2
2.0	261	68 121	522	4
2.5	275	75 625	687.5	6.25
2.0	276	76 176	552	4
25	2 743	780 639	7 375	73

$$r = \frac{10(7\,375) - (2\,743)(25)}{\sqrt{[10(780\,639) - (2\,743)^2][10(73) - (25)^2]}}$$

$= 5\,175/5444.796 = 0.95045$; a very strong positive correlation.

(d) Construct the linear regression equation for the data.
Regression equation: $Y' = a + bX$;

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{10(7\,375) - (2\,743)(25)}{10(780\,639) - (2\,743)^2} = 5\,175/282\,341 = 0.018;$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{25}{10} - 0.018\left(\frac{2\,743}{10}\right) = -2.4374.$$

The linear regression equation is $Y' = -2.4374 + 0.018X$.

(e) Estimate the amount of weight loss if a person ran on a treadmill for four hours in one week.

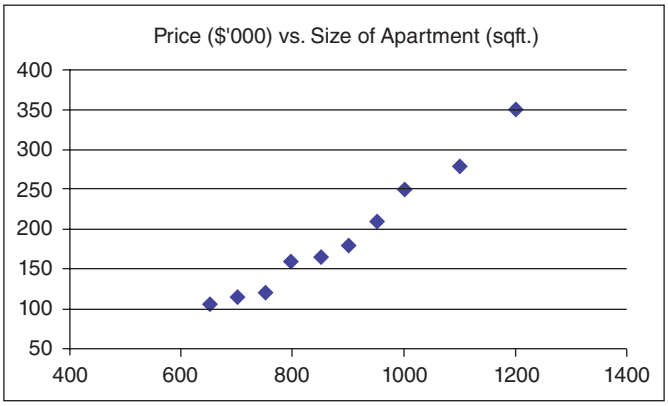
If time on a treadmill is 4 hours/week, Weight loss = $-2.4374 + 0.018(240) = 1.8826$kg.

4 A real estate agent wants to know the relationship between the size of apartment (in square feet) and the selling price (in $'000) in a city. From a survey, he has summarized the following information:

Size of Apartment	Price ($'000)
650	105
700	115
750	120
800	160
850	165
900	180
950	210
1 000	250
1 100	280
1 200	350

(a) Determine the dependent variable and independent variable.
Dependent variable: Price ($'000).
Independent variable: Size of apartment (sqft.).

(b) Sketch a scatter diagram showing the relationship between the two variables.



There is a very strong positive relationship between price ($′000) and size of apartment (sqft).

(c) Calculate the product moment correlation coefficient and interpret.

Use the formula $r = \dfrac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$;

Size of Apartment X	Price ($′000) Y	X^2	XY	Y^2
650	105	422 500	68 250	11 025
700	115	490 000	80 500	13 225
750	120	562 500	90 000	14 400
800	160	640 000	128 000	25 600
850	165	722 500	140 250	27 225
900	180	810 000	162 000	32 400
950	210	902 500	199 500	44 100
1000	250	1 000 000	250 000	62 500
1100	280	1 210 000	308 000	78 400
1200	350	1 440 000	420 000	122 500
8900	1935	8 200 000	1 846 500	431 375

$$r = \frac{10(1\,846\,500) - (8\,900)(1\,939)}{\sqrt{[10(8\,200\,000) - (8\,900)^2][10(431\,375) - (1\,935)^2]}}$$

$= 1\,243\,500/1\,260\,545.418 = 0.986$; a very strong positive correlation.

(d) Construct the regression equation using the least square method.

Regression equation: $Y' = a + bX$;

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{10(1\,846\,500) - (8\,900)(1\,935)}{10(8\,200\,000) - (8\,900)^2} = 1\,243\,500/2\,790\,000 = 0.4457;$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{1\,935}{10} - 0.4457\left(\frac{8\,900}{10}\right) = -203.173.$$

The regression equation is $Y' = -203.173 + 0.4457X$.

(e) Estimate the price of an apartment with the size of 836 square feet.

If the size is 836 sqft, the price = $-203.173 + 0.4457(836) = 169.4322$ ($′000).

(f) Obtain the coefficient of determination and give your comment.

The coefficient of determination is $r^2 = (0.986)^2 = 0.9722$; we can say that 97.22% of the price is explained by the variation in the size of apartment.

5 Suppose a police station wants to study the relationship between the time taken (in minutes) to reach the scene of an emergency and the distance from the station (in km). Eight recent emergency calls give the following data.

Time Taken (minutes)	10	7	12	9	11	15	5	6
Distance (km)	4.4	2.8	5.6	3.3	4.1	6.5	1.7	2.0

(a) Compute the Pearson product moment correlation coefficient. Interpret your result.

Distance (km), X	Time Taken (minutes), Y	X^2	XY	Y^2
4.4	10	19.36	44	100
2.8	7	7.84	19.6	49
5.6	12	31.36	67.2	144
3.3	9	10.89	29.7	81
4.1	11	16.81	45.1	121
6.5	15	42.25	97.5	225
1.7	5	2.89	8.5	25
2	6	4	12	36
30.4	75	135.4	323.6	781

$$r = \frac{8(323.6) - (30.4)(75)}{\sqrt{[8(135.4) - (30.4)^2][8(781) - (75)^2]}}$$

$= 308.8/314.7728 = 0.981$; a very strong positive correlation.

(b) Use the least squares method to construct the regression equation of the time taken to reach the scene against the distance from the police station.

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{8(323.6) - (30.4)(75)}{8(135.4) - (30.4)^2} = 308.8/159.04 = 1.942;$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{75}{8} - 1.942\left(\frac{30.4}{8}\right) = 1.9954.$$

The regression equation is $Y' = 1.9954 + 1.942X$.

(c) Describe the value of b (the slope of the regression equation).

The value of b, 1.942 indicates that for each km distance from the police station, the time taken to reach the scene is about 2 minutes.

(d) Compute the coefficient of determination. Interpret your result.

The coefficient of determination is $r^2 = (0.981)^2 = 0.9624$; we can say that 96.24% of the time taken to reach the scene is explained by the variation in the distance from the police station.

(e) Estimate the time taken to reach the scene of an emergency call if the distance from the station is 2.5km.

If the distance is 2.5km, then:

Time Taken = 1.9954 + 1.942(2.5) = 6.85 minutes.

6 The manager of a catering company is studying the relationship between the number of students at boarding schools and the winning bids of catering prices per student.

School	Number of Students	Winning Bid ($)
1	300	12.00
2	350	11.80
3	400	11.50
4	450	11.50
5	500	11.00
6	550	10.80
7	600	10.50
8	650	10.50
9	700	10.30
10	750	10.20
11	800	9.80
12	850	9.60

(a) State the independent variable and the dependent variable.

Dependent variable: Winning Bid.
Independent variable: Number of Students.

(b) Calculate the Pearson's product moment correlation coefficient to determine whether there is a linear relationship between the number of students and the winning bid.

Number of Students, X	Winning Bid ($), Y	X^2	XY	Y^2
300	12.00	90 000	3 600	144
350	11.80	122 500	4 130	139.24
400	11.50	160 000	4 600	132.25
450	11.50	202 500	5 175	132.25
500	11.00	250 000	5 500	121
550	10.80	302 500	5 940	116.64
600	10.50	360 000	6 300	110.25
650	10.50	422 500	6 825	110.25
700	10.30	490 000	7 210	106.09
750	10.20	562 500	7 650	104.04
800	9.80	640 000	7 840	96.04
850	9.60	722 500	8 160	92.16
6 900	129.5	4 325 000	72 930	1404.21

$$r = \frac{12(72\,930) - (6\,900)(129.5)}{\sqrt{[12(4\,325\,000) - (6\,900)^2][12(1\,404.21) - (129.5)^2]}}$$

$= -18\,390/18\,556.8936 = -0.991$; a very strong negative linear relationship.

(c) Construct the equation of the regression line of the winning bid against the number of students using the least square method.

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{8(323.6) - (30.4)(75)}{8(135.4) - (30.4)^2} = -18\,390/4\,290\,000 = -0.0043;$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{129.5}{12} + 0.0043\left(\frac{6\,900}{12}\right) = 13.2645.$$

The regression equation is $Y' = 13.2645 - 0.0043X$.

(d) Determine the coefficient of determination and comment on its value.

The coefficient of determination is $r^2 = (-0.991)^2 = 0.9821$; we can say that 98.21% of the winning bid is explained by the variation in the number of students.

(e) Estimate the winning bid if the number of students is 780.

If the number of students is 780, the Winning Bid = 13.2645 – 0.0043(780) = $9.91.

7 Two judges of a reality TV program were evaluating nine participants. Each of the judges has ranked the participants from 1 to 9 in order of their performances as shown in the following table.

Participant	1	2	3	4	5	6	7	8	9
Judge 1	6	3	9	1	7	5	2	8	4
Judge 2	5	6	4	3	7	9	2	8	1

Compute the correlation between the rankings of the two judges and interpret.

Use the Spearman's coefficient of rank correlation; $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$.

Participant	Rank		Difference between ranks, d	Difference squared, d^2
	Judge 1	Judge 2		
1	6	5	1	1
2	3	6	–3	9
3	9	4	5	25
4	1	3	–2	4
5	7	7	0	0
6	5	9	–4	16
7	2	2	0	0
8	8	8	0	0
9	4	1	3	9
			0	64

$r_s = 1 - 6(64)/9(80) = 0.4667$.
A value of 0.4467 indicates a rather weak relationship between two judges.

8 The following table shows the grades of ten students in Advanced Calculus and Introduction to Computing examinations.

Student	1	2	3	4	5	6	7	8	9	10
Adv. Calculus	E	D	D	D	C	C	C	B	B	A
Intro. to Computing	D	D	E	C	B	C	D	A	C	B

Determine the Spearman's rank correlation coefficient and interpret.

Student	Adv. Calculus	Intro. to Computing	Rank		Difference between ranks, d	Difference squared, d^2
			Adv. Calculus	Intro. to Computing		
1	E	D	10	8	2	4
2	D	D	8	8	0	0
3	D	E	8	10	–2	4
4	D	C	8	5	3	9
5	C	B	5	2.5	2.5	6.25
6	C	C	5	5	0	0
7	C	D	5	8	–3	9
8	B	A	2.5	1	1.5	2.25
9	B	C	2.5	5	–2.5	6.25
10	A	B	1	2.5	–1.5	2.25
					0	43

$r_s = 1 - 6(43)/10(99) = 0.7394$.
A value of 0.7394 indicates a rather strong relationship between two examinations.

9 The dean of a faculty decided to set up two panels to evaluate the ten students who were nominated for excellence awards. The results are given below.

Student	Ranking by Panel 1	Ranking by Panel 2
1	10	9
2	5	6
3	2	3
4	8	8
5	7	5
6	3	2
7	4	4
8	6	7
9	1	1
10	9	10

Calculate Spearman's rank correlation coefficient. Interpret the result.

Student	Rank		Difference between ranks, d	Difference squared, d^2
	Panel 1	Panel 2		
1	10	9	1	1
2	5	6	–1	1
3	2	3	–1	1
4	8	8	0	0
5	7	5	2	4
6	3	2	1	1
7	4	4	0	0

(contd.)

Student	Rank		Difference between ranks, d	Difference squared, d^2
	Panel 1	Panel 2		
8	6	7	−1	1
9	1	1	0	0
10	9	10	−1	1
			0	10

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - 6(10)/10(99) = 0.9394.$$

A value of 0.9394 indicates a very strong relationship between two panels.

10 Nine CEOs were nominated by a society for the best CEO of the year. The President and the Deputy President of the society were asked to rank the CEOs according to their performances. The results are shown in the following table.

CEO	C1	C2	C3	C4	C5	C6	C7	C8	C9
Ranking by President	9	7	8	5	6	4	3	1	2
Ranking by Deputy President	8	9	5	7	6	4	1	2	3

Do the President and the Deputy President have similar opinions? Support your answer.

CEO	Rank		Difference between ranks, d	Difference squared, d^2
	President	Deputy Pres.		
C1	9	8	1	1
C2	7	9	−2	4
C3	8	5	3	9
C4	5	7	−2	4
C5	6	6	0	0
C6	4	4	0	0
C7	3	1	2	4
C8	1	2	−1	1
C9	2	3	−1	1
			0	24

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - 6(24)/9(80) = 0.8.$$

A value of 0.8 indicates a fairly strong relationship between two panels. Hence, the President and the Deputy President have quite similar opinions.

SOLUTION MANUAL

CHAPTER 10 Time Series Analysis and Forecasting

1 Describe the following terms.

(a) Secular trend.

Secular trend is a movement or trend in a series over very long periods of time (long-term trend).

(b) Irregular variation.

Irregular variation is a fluctuation in time series, short in duration, erratic in nature and follows no regularity in the occurrence pattern.

(c) Seasonal factors.

Seasonal factors are the factors that reflect the seasonal variations by repeating every year to the same extent.

2 Identify whether the following events are trend, cyclical, seasonal or irregular component of time series.

(a) The increase in the number of customers in supermarkets during a festive season. Seasonal

(b) The decrease in the price of mobile phones over the years as more people use them. Trend

(c) The high demand for foods when the economy is doing well. Irregular

(d) The increases in the price of fuels due to shortages. Irregular

(e) The increase in patients in an outpatient unit due to H1N1 outbreak. Irregular

(f) The increase in unemployment rate due to rising inflation. Irregular

3 The following time series plot shows quarterly sales (in millions) of a steel manufacturer from 2007 to 2010.

Describe the trend and the seasonal effects of the sales.

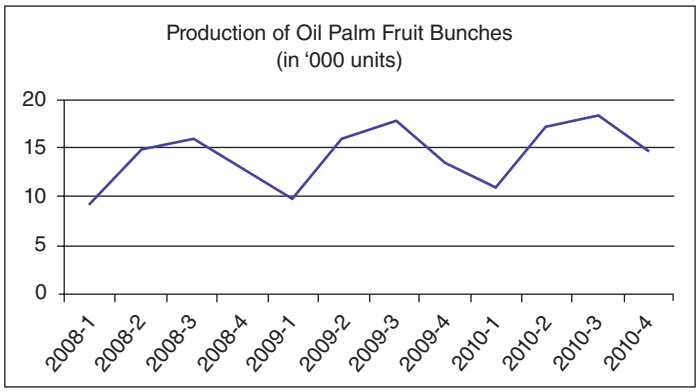
Trend: The quarterly sales of the steel manufacturer increase at a constant rate for the period 2007–2010.

Seasonal effects: The plot shows seasonal quarterly sales with low sales in the first quarters and high sales in the third quarters.

4 The following table shows the quarterly production of oil palm fruit bunches (in '000 units) at a private plantation for the years 2008 to 2010.

Year	Quarter			
	1	2	3	4
2008	9.3	14.8	16.1	12.7
2009	10.0	16.0	17.7	13.7
2010	10.9	17.3	18.4	14.7

(a) Sketch a time series plot for the data.



(b) Compute the trend values using the moving-average method.
The plot shows that the cycle is repeated every four quarters or year (seasonal fluctuations). Hence, four-quarter moving-average would completely average out the seasonal fluctuations. The trend values can be computed as follows:

Year-Quarter	Production (in '000 units)	4-quarter moving total	4-quarter moving-average (Trend)
2008-1	9.3		
2008-2	14.8	52.9	**13.225**
2008-3	16.1	53.6	**13.4**
2008-4	12.7	54.8	**13.7**
2009-1	10	56.4	**14.1**
2009-2	16	57.4	**14.35**
2009-3	17.7	58.3	**14.575**
2009-4	13.7	59.6	**14.9**
2010-1	10.9	60.3	**15.075**
2010-2	17.3	61.3	**15.325**
2010-3	18.4		
2010-4	14.7		

(c) Calculate the quarterly seasonal indexes using the multiplicative model.

The quarterly seasonal indexes can be determined using the values produced in (b), then calculate the centred moving average and the specific seasonal as follows:

Year-Quarter	Production (in '000) (1)	4-quarter moving total	4-quarter moving-average	Centred moving-average (2)	Specific Seasonal (1)/(2)
2008-1	9.3				
2008-2	14.8	52.9	13.225	13.3125	1.2094
2008-3	16.1	53.6	13.4	13.55	0.9373
2008-4	12.7	54.8	13.7	13.9	0.7194
2009-1	10	56.4	14.1	14.225	1.1248
2009-2	16	57.4	14.35	14.4625	1.2238
2009-3	17.7	58.3	14.575	14.7375	0.9296
2009-4	13.7	59.6	14.9	14.9875	0.7273
2010-1	10.9	60.3	15.075	15.2	1.1382
2010-2	17.3	61.3	15.325		
2010-3	18.4				
2010-4	14.7				

Next, reorganize the specific seasonal indexes as the following, and then compute the mean index for each quarter:

Year	Quarter				
	1	2	3	4	
2008			1.2094	0.9373	
2009	0.7194	1.1248	1.2238	0.9296	
2010	0.7273	1.1382			
Mean	0.7233	1.1315	1.2166	0.9335	4.0049
Typical index	72.24	113.01	121.51	93.24	400.00

(d) Describe the seasonal indexes for the 2nd and 4th quarters.

Seasonal index for the 2nd quarter = 113.01; i.e. the production in the 2nd quarter is 13.01% above the typical quarter.

Seasonal index for the 4th quarter = 93.24; i.e. the production in the 4th quarter is 6.76% below the typical quarter.

(e) Forecast the production of oil palm fruit bunches for the 3rd quarter of 2011.

To prepare the data for the forecast, first the original quarterly production must be deseasonalized by dividing each value with the corresponding seasonal index, as shown below:

Year-Quarter	Production (in '000)	Seasonal index	Deseasonalized production
2008-1	9.3	0.7224	12.874
2008-2	14.8	1.1301	13.096
2008-3	16.1	1.2151	13.250

(contd.)

Year-Quarter	Production (in '000)	Seasonal index	Deseasonalized production
2008-4	12.7	0.9324	13.621
2009-1	10	0.7224	13.843
2009-2	16	1.1301	14.158
2009-3	17.7	1.2151	14.567
2009-4	13.7	0.9324	14.693
2010-1	10.9	0.7224	15.089
2010-2	17.3	1.1301	15.308
2010-3	18.4	1.2151	15.143
2010-4	14.7	0.9324	15.766

The next step is to construct a trend equation for the deseasonalized data, as obtained below:

$$b = \frac{\sum tY - (\sum Y)(\sum t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{1151.615 - (171.407)(78)/12}{650 - (78)^2/12} = \frac{37.4695}{143} = 0.262$$

$$a = \frac{\sum Y}{n} - b\frac{\sum t}{n} = \frac{171.407}{12} - 0.262\left(\frac{78}{12}\right) = 14.284 - 1.703 = 12.581$$

Therefore, $Y' = 12.581 + 0.262t$.
For the 3rd quarter of 2011 ($t = 15$), $Y' = 12.581 + 0.262(15) = 16.511$.
Thus, the 3rd quarter of 2011 forecast = 16.511(1.2151) = 20.0625 ('000 units).

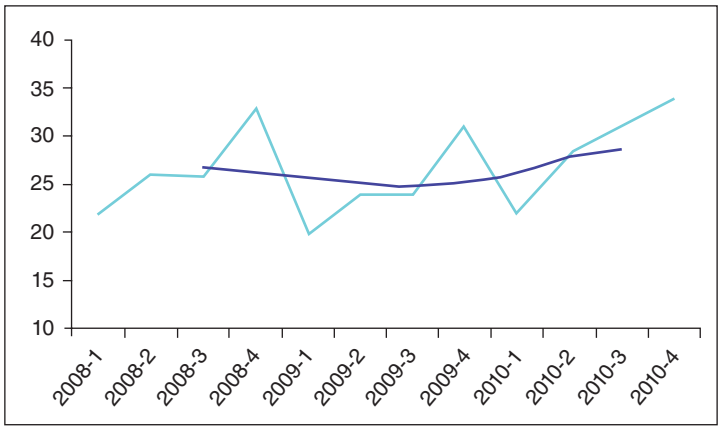
5 The following table shows the quarterly averages of the number of rooms occupied per day at a budget hotel in Malacca for the period 2008–2010.

Year/Quarter	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2008	22	26	26	33
2009	20	24	24	31
2010	22	28	31	34

(a) Calculate the trend values using the moving-average method.

Year-Quarter	Ave. occupied rooms per day	4-quarter moving total	4-quarter moving-average (Trend)
2008-1	22		
2008-2	26		
2008-3	26	107	26.75
2008-4	33	105	26.25
2009-1	20	103	25.75
2009-2	24	101	25.25
2009-3	24	99	24.75
2009-4	31	101	25.25
2010-1	22	105	26.25
2010-2	28	112	28.00
2010-3	31	115	28.75
2010-4	34		

(b) Plot the time series and the trend on the same diagram.



(c) Compute the seasonal indexes using the multiplicative model.

From (b), the centred moving average and the specific seasonal are as follows:

Year-Quarter	Ave. occupied rooms per day	4-quarter moving total	4-quarter moving-average	Centred moving-average	Specific Seasonal
2008-1	22				
2008-2	26				
2008-3	26	107	26.75	26.500	0.9811
2008-4	33	105	26.25	26.000	1.2692
2009-1	20	103	25.75	25.500	0.7843
2009-2	24	101	25.25	25.000	0.9600
2009-3	24	99	24.75	25.000	0.9600
2009-4	31	101	25.25	25.750	1.2039
2010-1	22	105	26.25	27.125	0.8111
2010-2	28	112	28.00	28.375	0.9868
2010-3	31	115	28.75		
2010-4	34				

Now reorganize the specific seasonal indexes as the following, and then compute the mean index for each quarter:

Year	Quarter				
	1	2	3	4	
2008			0.9811	1.2692	
2009	0.7843	0.9600	0.9600	1.2039	
2010	0.8111	0.9868			
Mean	0.7977	0.9734	0.9706	1.2365	3.9782
Typical index	80.21	97.87	97.59	124.33	400.00

(d) Describe the seasonal indexes for Quarter 1 and Quarter 3.

Seasonal index for Quarter 1 = 80.21; i.e. the average occupied rooms per day is 19.79% below the typical quarter.

Seasonal index for Quarter 3 = 97.59; i.e. the average occupied rooms per day is 2.41% below the typical quarter.

(e) Estimate the average number of rooms occupied per day for Quarter 4 of 2011.

Now the original quarterly data must be deseasonalized by dividing each with the corresponding seasonal index, as shown below:

Year-Quarter	Ave. occupied rooms per day	Seasonal index	Deseasonalized ave. occupied rooms
2008-1	22	0.8021	27.428
2008-2	26	0.9787	26.566
2008-3	26	0.9759	26.642
2008-4	33	1.2433	26.542
2009-1	20	0.8021	24.935
2009-2	24	0.9787	24.522
2009-3	24	0.9759	24.593
2009-4	31	1.2433	24.934
2010-1	22	0.8021	27.428
2010-2	28	0.9787	28.609
2010-3	31	0.9759	31.766
2010-4	34	1.2433	27.347

Now construct a trend equation for the deseasonalized data, as obtained below:

$$b = \frac{\sum tY - (\sum Y)(\sum t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{2120.616 - (321.312)(78)/12}{650 - (78)^2/12} = \frac{32.088}{143} = 0.2244$$

$$a = \frac{\sum Y}{n} - b\frac{\sum t}{n} = \frac{321.312}{12} - 0.2244\left(\frac{78}{12}\right) = 26.776 - 1.4586 = 25.3174$$

Therefore, $Y' = 25.3174 + 0.2244t$.
For Quarter 4 of 2011 ($t = 16$), $Y' = 25.3174 + 0.2244(16) = 28.9078$.
Thus, Quarter 4 of 2011 forecast = 28.9078(1.2433) = 35.94 occupied rooms per day.

6 The production of bottled mineral water at Healthy Water over the period 2008-2010 is given in the following table.

Year	Production ('000 bottles)			
	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2008	70	90	80	130
2009	110	140	100	150
2010	110	130	110	170

The trend line was computed as $T = 79.24 + 5.63t$, where $t = 1$ for the first Quarter of 2008.

(a) Determine the trend values.

Year-Quarter	Production ('000 bottles)	4-quarter moving total	4-quarter moving-average (Trend)
2008-1	70		
2008-2	90		
2008-3	80	370	92.50
2008-4	130	410	102.50
2009-1	110	460	115.00
2009-2	140	480	120.00
2009-3	100	500	125.00
2009-4	150	500	125.00
2010-1	110	490	122.50
2010-2	130	500	125.00
2010-3	110	520	130.00
2010-4	170		

(b) Compute the seasonal index for each quarter.

From (b), centred moving-average and specific seasonal are obtained as follows:

Year-Quarter	Production ('000 bottles)	4-quarter moving total	4-quarter moving-average	Centred moving-average	Specific Seasonal
2008-1	70				
2008-2	90				
2008-3	80	370	92.50	97.50	0.8205
2008-4	130	410	102.50	108.75	1.1954
2009-1	110	460	115.00	117.50	0.9362
2009-2	140	480	120.00	122.50	1.1429
2009-3	100	500	125.00	125.00	0.8000
2009-4	150	500	125.00	123.75	1.2121
2010-1	110	490	122.50	123.75	0.8889
2010-2	130	500	125.00	127.50	1.0196
2010-3	110	520	130.00		
2010-4	170				

Reorganize the specific seasonal indexes as the following, and compute the mean index for each quarter:

Year	Quarter				
	1	2	3	4	
2008			0.8205	1.1954	
2009	0.9362	1.1429	0.8000	1.2121	
2010	0.8889	1.0196			
Mean	0.9125	1.0813	0.8103	1.2037	4.0078
Typical index	91.07	107.92	80.87	120.14	400.00

(c) Interpret the seasonal index for the first and third quarters.

Seasonal index for Quarter 1 = 91.07; i.e. the production is 8.93% below the typical quarter.

Seasonal index for Quarter 3 = 80.87; i.e. the production is 19.13% below the typical quarter.

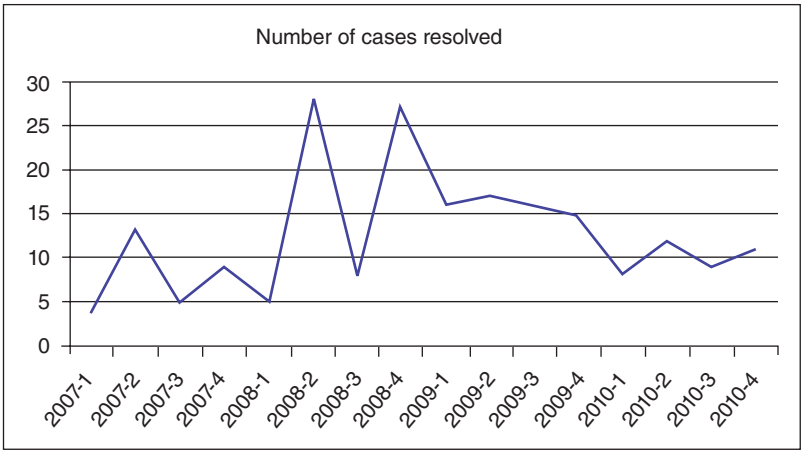
(d) Forecast the production for the first quarter of 2011.

$T = 79.24 + 5.63t$; for Quarter 1 of 2011, $t = 13$, there for $T = 152.43$.
Thus, production for Quarter 1 of 2011 is 152.43(0.9107) = 138.818 ('000 bottles).

7 The number cases resolved by a lawyer were recorded quarterly as in the table below for the period 2007–2010.

Year	Number of cases resolved			
	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2007	4	13	5	9
2008	5	28	8	27
2009	16	17	16	15
2010	8	12	9	11

(a) Illustrate the above data using a time series plot.



(b) Using the moving-average method, obtain the trend values.

Year-Quarter	No. of cases resolved	4-quarter moving total	4-quarter moving-average (Trend)
2007-1	4		
2007-2	13	31	7.75
2007-3	5	32	8.00
2007-4	9	47	11.75
2008-1	5	50	12.50
2008-2	28	68	17.00
2008-3	8	79	19.75
2008-4	27	68	17.00
2009-1	16	76	19.00
2009-2	17	64	16.00
2009-3	16	56	14.00
2009-4	15	51	12.75
2010-1	8	44	11.00
2010-2	12	40	10.00
2010-3	9		
2010-4	11		

(c) Compute the seasonal index for each quarter.

From (b), centred moving-average and specific seasonal are obtained as follows:

Year-Quarter	No. of cases resolved	4-quarter moving-total	4-quarter moving-average	Centred moving-average	Specific Seasonal
2007-1	4				
2007-2	13				
2007-3	5	31	7.75	7.88	0.6349
2007-4	9	32	8.00	9.88	0.9114
2008-1	5	47	11.75	12.13	0.4124
2008-2	28	50	12.50	14.75	1.8983
2008-3	8	68	17.00	18.38	0.4354
2008-4	27	79	19.75	18.38	1.4694
2009-1	16	68	17.00	18.00	0.8889
2009-2	17	76	19.00	17.50	0.9714
2009-3	16	64	16.00	15.00	1.0667
2009-4	15	56	14.00	13.38	1.1215
2010-1	8	51	12.75	11.88	0.6737
2010-2	12	44	11.00	10.50	1.1429
2010-3	9	40	10.00		
2010-4	11				

Reorganize the specific seasonal indexes and compute the mean index for each quarter:

Year	Quarter				
	1	2	3	4	
2007			0.6349	0.9114	
2008	0.4124	1.8983	0.4354	1.4694	
2009	0.8889	0.9714	1.0667	1.1215	
2010	0.6737	1.1429			
Mean	0.6583	1.3375	0.7123	1.1674	3.8755
Typical index	67.94	138.05	73.52	120.49	400.00

(d) Describe the seasonal index for Quarter 1 and Quarter 3.

Seasonal index for Q1 = 67.94; i.e. 32.06% below the typical quarter.

Seasonal index for Q3 = 73.52; i.e. 26.48% below the typical quarter.

(e) Forecast the number of cases resolved for the third quarter of 2011.

The original data must be deseasonalized by dividing each value with the corresponding seasonal index, as shown below:

Year-Quarter	No. of cases resolved	Seasonal index	Deseasonalized no. of cases resolved
2007-1	4	0.6794	5.888
2007-2	13	1.3805	9.417
2007-3	5	0.7352	6.801

(contd.)

Year-Quarter	No. of cases resolved	Seasonal index	Deseasonalized no. of cases resolved
2007-4	9	1.2049	7.469
2008-1	5	0.6794	7.359
2008-2	28	1.3805	20.283
2008-3	8	0.7352	10.881
2008-4	27	1.2049	22.408
2009-1	16	0.6794	23.550
2009-2	17	1.3805	12.314
2009-3	16	0.7352	21.763
2009-4	15	1.2049	12.449
2010-1	8	0.6794	11.775
2010-2	12	1.3805	8.693
2010-3	9	0.7352	12.242
2010-4	11	1.2049	9.129

Now construct a trend equation for the deseasonalized data, as obtained below:

$$b = \frac{\sum tY - (\sum Y)(\sum t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{1817.273 - (202.422)(136)/16}{1\,496 - (136)^2/16} = \frac{96.686}{340} = 0.2844$$

$$a = \frac{\sum Y}{n} - b\frac{\sum t}{n} = \frac{202.422}{16} - 0.2844\left(\frac{136}{16}\right) = 12.6514 - 2.4174 = 10.234$$

Therefore, $Y' = 10.234 + 0.2844t$.
For Quarter 3 of 2011 ($t = 19$), $Y' = 10.234 + 0.2844(19) = 15.6376$.
Thus, Quarter 3 of 2011 forecast = 15.6376(0.7352) = 11.497 cases resolved.

8 A private college offers diploma programs with three semesters per year. The numbers of students graduated each semester for the period 2007-2010 are given in the following table.

Year	Semester 1	Semester 2	Semester 3
2007	231	192	189
2008	264	222	230
2009	354	387	370
2010	350	349	331

(a) Compute the trend values using the moving-average method.

Year-Semester	No. of graduated students	3-semester moving total	3-semester moving average (Trend)
2007-1	231		
2007-2	192	612	204.00
2007-3	189	645	215.00
2008-1	264	675	225.00
2008-2	222	716	238.67
2008-3	230	806	268.67
2009-1	354	971	323.67
2009-2	387	1 111	370.33
2009-3	370	1 107	369.00
2010-1	350	1 069	356.33
2010-2	349	1 030	343.33
2010-3	331		

(b) Determine the seasonal index for each semester.

From (a), the specific seasonal indexes are obtained as follows:

Year-Semester	No. of graduated students (1)	3-semester moving total	3-semester moving average (2)	Specific Seasonal (1)/(2)
2007-1	231			
2007-2	192	612	204.00	0.9412
2007-3	189	645	215.00	0.8791
2008-1	264	675	225.00	1.1733
2008-2	222	716	238.67	0.9302
2008-3	230	806	268.67	0.8561
2009-1	354	971	323.67	1.0937
2009-2	387	1 111	370.33	1.0450
2009-3	370	1 107	369.00	1.0027
2010-1	350	1 069	356.33	0.9822
2010-2	349	1 030	343.33	1.0165
2010-3	331			

Reorganize the specific seasonal indexes and compute the mean index for each semester:

Year	Semester			
	1	2	3	
2007		0.9412	0.8791	
2008	1.1733	0.9302	0.8561	
2009	1.0937	1.0450	1.0027	
2010	0.9822	1.0165		
Mean	1.0831	0.9832	0.9126	2.9789
Typical index	109.07	99.02	91.91	300.00

(c) Describe the seasonal index for the third semester.

Seasonal index for Semester 3 = 91.91; i.e. the number of graduated students is 8.09% below the typical semester.

(d) Estimate the number of students that will graduate in the second semester of 2011.

Deseasonalize the original data by dividing each value with the corresponding seasonal index, as shown below:

Year-Semester	No. of graduated students	Seasonal index	Deseasonalized no. of graduated students
2007-1	231	1.0907	211.79
2007-2	192	0.9902	193.90
2007-3	189	0.9191	205.64
2008-1	264	1.0907	242.05
2008-2	222	0.9902	224.20
2008-3	230	0.9191	250.24
2009-1	354	1.0907	324.56
2009-2	387	0.9902	390.83
2009-3	370	0.9191	402.57
2010-1	350	1.0907	320.89
2010-2	349	0.9902	352.45
2010-3	331	0.9191	360.13

Construct a trend equation for the deseasonalized data:

$$b = \frac{\sum tY - (\sum Y)(\sum t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{25236.39 - (3479.26)(78)/12}{650 - (78)^2/12} = \frac{2621.2}{143} = 18.33$$

$$a = \frac{\sum Y}{n} - b\frac{\sum t}{n} = \frac{3479.26}{12} - 18.33\left(\frac{78}{12}\right) = 289.938 - 119.145 = 170.793$$

Therefore, $Y' = 170.793 + 18.33t$.
For Semester 2 of 2011 ($t = 14$), $Y' = 170.793 + 18.33(14) = 427.413$.
Thus, Semester 2 of 2011 forecast = 427.413(0.9902) = 423.224 graduated students.

9 The table below shows the monthly number of tourist arrivals to Malaysia from 2008 to 2010.

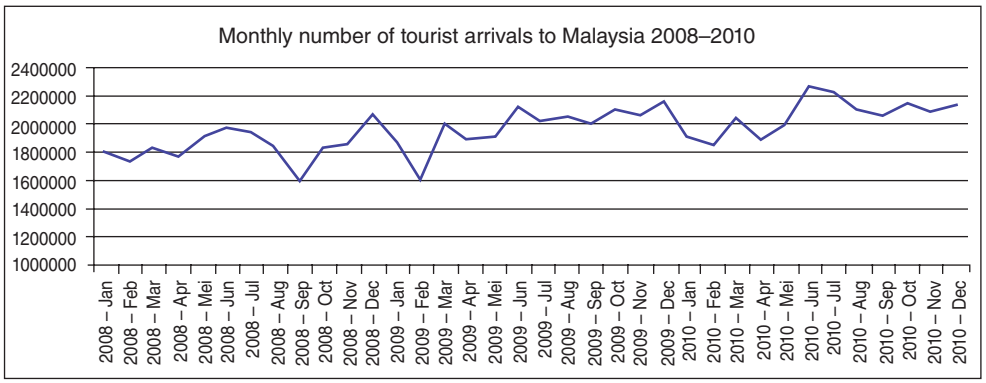
Month	2008	2009	2010
January	1 780 134	1 871 099	1 896 918
February	1 742 468	1 613 309	1 832 300
March	1 819 689	1 975 776	2 022 590
April	1 760 326	1 883 873	1 877 934
May	1 899 148	1 894 059	1 992 277
June	1 961 355	2 108 328	2 246 084
July	1 928 082	2 003 724	2 214 092
August	1 839 235	2 030 337	2 099 485

(contd.)

Month	2008	2009	2010
September	1 599 418	1 997 535	2 053 406
October	1 818 304	2 078 485	2 137 735
November	1 845 645	2 048 595	2 081 354
December	2 058 684	2 141 071	2 123 021

(Source: Tourism Malaysia, http://corporate.tourism.gov.my)

(a) Draw a time series plot for the data.



(b) Determine the trend values for the number of tourist arrivals to Malaysia using the moving-average method.

Year-Month	No. of tourist arrivals	12-month moving total	12-month moving-average (Trend)	Year-Month	No. of tourist arrivals	12-month moving-total	12-month moving-average (Trend)
2008-Jan	1780134			2009-Jul	2003724	23672010	**1972667.50**
2008-Feb	1742468			2009-Aug	2030337	23891001	**1990916.75**
2008-Mar	1819689			2009-Sep	1997535	23937815	**1994817.92**
2008-Apr	1760326			2009-Oct	2078485	23931876	**1994323.00**
2008-May	1899148			2009-Nov	2048595	24030094	**2002507.83**
2008-Jun	1961355	22052488	**1837707.33**	2009-Dec	2141071	24167850	**2013987.50**
2008-Jul	1928082	22143453	**1845287.75**	2010-Jan	1896918	24378218	**2031518.17**
2008-Aug	1839235	22014294	**1834524.50**	2010-Feb	1832300	24447366	**2037280.50**
2008-Sep	1599418	22170381	**1847531.75**	2010-Mar	2022590	24503237	**2041936.42**
2008-Oct	1818304	22293928	**1857827.33**	2010-Apr	1877934	24562487	**2046873.92**
2008-Nov	1845645	22288839	**1857403.25**	2010-May	1992277	24595246	**2049603.83**
2008-Dec	2058684	22435812	**1869651.00**	2010-Jun	2246084	24577196	**2048099.67**
2009-Jan	1871099	22511454	**1875954.50**	2010-Jul	2214092		
2009-Feb	1613309	22702556	**1891879.67**	2010-Aug	2099485		
2009-Mar	1975776	23100673	**1925056.08**	2010-Sep	2053406		
2009-Apr	1883873	23360854	**1946737.83**	2010-Oct	2137735		
2009-May	1894059	23563804	**1963650.33**	2010-Nov	2081354		
2009-Jun	2108328	23646191	**1970515.92**	2010-Dec	2123021		

(c) Determine the seasonal index for each month.

From (b), the specific seasonal indexes are obtained as follows:

Year-Month	No. of tourist arrivals (1)	12-month moving-total	12-month moving-average	Centred moving-average (2)	Specific Seasonal (1)/(2)
2008-Jan	1780134				
2008-Feb	1742468				
2008-Mar	1819689				
2008-Apr	1760326				
2008-May	1899148				
2008-Jun	1961355				
2008-Jul	1928082	22052488	1837707.33	1841497.542	1.0470
2008-Aug	1839235	22143453	1845287.75	1839906.125	0.9996
2008-Sep	1599418	22014294	1834524.50	1841028.125	0.8688
2008-Oct	1818304	22170381	1847531.75	1852679.542	0.9814
2008-Nov	1845645	22293928	1857827.33	1857615.292	0.9936
2008-Dec	2058684	22288839	1857403.25	1863527.125	1.1047
2009-Jan	1871099	22435812	1869651.00	1872802.75	0.9991
2009-Feb	1613309	22511454	1875954.50	1883917.083	0.8564
2009-Mar	1975776	22702556	1891879.67	1908467.875	1.0353
2009-Apr	1883873	23100673	1925056.08	1935896.958	0.9731
2009-May	1894059	23360854	1946737.83	1955194.083	0.9687
2009-Jun	2108328	23563804	1963650.33	1967083.125	1.0718
2009-Jul	2003724	23646191	1970515.92	1971591.708	1.0163
2009-Aug	2030337	23672010	1972667.50	1981792.125	1.0245
2009-Sep	1997535	23891001	1990916.75	1992867.333	1.0023
2009-Oct	2078485	23937815	1994817.92	1994570.458	1.0421
2009-Nov	2048595	23931876	1994323.00	1998415.417	1.0251
2009-Dec	2141071	24030094	2002507.83	2008247.667	1.0661
2010-Jan	1896918	24167850	2013987.50	2022752.833	0.9378
2010-Feb	1832300	24378218	2031518.17	2034399.333	0.9007
2010-Mar	2022590	24447366	2037280.50	2039608.458	0.9917
2010-Apr	1877934	24503237	2041936.42	2044405.167	0.9186
2010-May	1992277	24562487	2046873.92	2048238.875	0.9727
2010-Jun	2246084	24595246	2049603.83	2048851.75	1.0963
2010-Jul	2214092	24577196	2048099.67		
2010-Aug	2099485				
2010-Sep	2053406				
2010-Oct	2137735				
2010-Nov	2081354				
2010-Dec	2123021				

Reorganize the specific seasonal indexes and compute the mean index for each month:

Year	Month												
	1	2	3	4	5	6	7	8	9	10	11	12	
2008							1.0470	0.9996	0.8688	0.9814	0.9936	1.1047	
2009	0.9991	0.8564	1.0353	0.9731	0.9687	1.0718	1.0163	1.0245	1.0023	1.0421	1.0251	1.0661	
2010	0.9378	0.9007	0.9917	0.9186	0.9727	1.0963							

(contd.)

Year	Month												
	1	2	3	4	5	6	7	8	9	10	11	12	
Mean	0.9685	0.8786	1.0135	0.9459	0.9707	1.0841	1.0317	1.0121	0.9356	1.0118	1.0094	1.0854	11.9469
Typical index	**97.28**	**88.25**	**101.80**	**95.01**	**97.50**	**108.89**	**103.62**	**101.66**	**93.97**	**101.63**	**101.38**	**109.02**	1200

(d) Use exponential smoothing with smoothing factor of 0.20 to calculate the forecast for January 2011. Calculate the mean squared error (MSE), the mean absolute deviation (MAD) and the mean absolute percentage error (MAPE).

Exponential Smoothing Forecast = 0.2(latest data point) + 0.8(previous forecast)

Year-Month	No. of tourist arrivals	Exponential Smoothing Forecast	Error (F_t-D_t)	Mean Squared Error (MSE)	MAD	MAPE (%)
2008-Jan	1780134	1780134.00				
2008-Feb	1742468	1780134.00	37666.00	1418727556.00	37666.00	2.16%
2008-Mar	1819689	1772600.80	–47088.20	2217298579.24	47088.20	2.59%
2008-Apr	1760326	1782018.44	21692.44	470561953.15	21692.44	1.23%
2008-May	1899148	1777679.95	–121468.05	14754486684.93	121468.05	6.40%
2008-Jun	1961355	1801973.56	–159381.44	25402442906.45	159381.44	8.13%
2008-Jul	1928082	1833849.85	–94232.15	8879698229.32	94232.15	4.89%
2008-Aug	1839235	1852696.28	13461.28	181206043.73	13461.28	0.73%
2008-Sep	1599418	1850004.02	250586.02	62793355193.19	250586.02	15.67%
2008-Oct	1818304	1799886.82	–18417.18	339192562.20	18417.18	1.01%
2008-Nov	1845645	1803570.26	–42074.74	1770284161.34	42074.74	2.28%
2008-Dec	2058684	1811985.20	–246698.80	60860295922.16	246698.80	11.98%
2009-Jan	1871099	1861324.96	–9774.04	95531794.55	9774.04	0.52%
2009-Feb	1613309	1863279.77	249970.77	62485386151.02	249970.77	15.49%
2009-Mar	1975776	1813285.62	–162490.38	26403124738.21	162490.38	8.22%
2009-Apr	1883873	1845783.69	–38089.31	1450795294.05	38089.31	2.02%
2009-May	1894059	1853401.55	–40657.45	1653027871.03	40657.45	2.15%
2009-Jun	2108328	1861533.04	–246794.96	60907750487.19	246794.96	11.71%
2009-Jul	2003724	1910892.03	–92831.97	8617773742.84	92831.97	4.63%
2009-Aug	2030337	1929458.43	–100878.57	10176486303.61	100878.57	4.97%
2009-Sep	1997535	1949634.14	–47900.86	2294492164.45	47900.86	2.40%
2009-Oct	2078485	1959214.31	–119270.69	14225496569.23	119270.69	5.74%
2009-Nov	2048595	1983068.45	–65526.55	4293728610.96	65526.55	3.20%
2009-Dec	2141071	1996173.76	–144897.24	20995209904.99	144897.24	6.77%
2010-Jan	1896918	2025153.21	128235.21	16444268751.08	128235.21	6.76%
2010-Feb	1832300	1999506.17	167206.17	27957902270.24	167206.17	9.13%
2010-Mar	2022590	1966064.93	–56525.07	3195083134.93	56525.07	2.79%
2010-Apr	1877934	1977369.95	99435.95	9887507527.13	99435.95	5.29%
2010-May	1992277	1957482.76	–34794.24	1210639312.21	34794.24	1.75%
2010-Jun	2246084	1964441.61	–281642.39	79322438104.94	281642.39	12.54%
2010-Jul	2214092	2020770.08	–193321.92	37373362900.38	193321.92	8.73%
2010-Aug	2099485	2059434.47	–40050.53	1604045126.93	40050.53	1.91%
2010-Sep	2053406	2067444.57	14038.57	197081567.41	14038.57	0.68%
2010-Oct	2137735	2064636.86	–73098.14	5343338157.34	73098.14	3.42%
2010-Nov	2081354	2079256.49	–2097.51	4399558.56	2097.51	0.10%
2010-Dec	2123021	2079675.99	–43345.01	1878789889.81	43345.01	2.04%
2011-Jan Forecast		**2088344.99**		16488720277.85	100161.14	5.14%

10 Consider the time series data in Question 7, and answer the following:

(a) Estimate the four seasonal factors C_i using all the data.

Reorganize the time series data as follows:

Quarter	Number of cases resolved				Seasonal Factors = average quarter / 4-year average
	2007	2008	2009	2010	
1	4	5	16	8	8.25/12.6875 = **0.6502**
2	13	28	17	12	17.5/12.6875 = **1.3793**
3	5	8	16	9	9.5/12.6875 = **0.7488**
4	9	27	15	11	15.5/12.6875 = **1.2217**
4-year Average = 12.6875					

(b) Estimate the quarterly trend factor G using all the data.

Estimate the quarterly trend by taking the increase in number of cases solved from Quarter 1 of 2007 to Quarter 1 of 2010 (8 – 4 = 4) and dividing by 12 quarters, to obtain an estimate of $G = 0.333$ cases per quarter.

(c) Estimate the initial S value by taking the average quarterly number of cases solved in the first year and then subtracting two quarters' worth of trend.

To estimate the initial S value, we calculate average 2007 = 7.75 and then subtract two quarters of trend (7.75 – 2(0.333) = 7.083) and use 7.083 as an initial estimate of previous S before Quarter 1 of 2007.

(d) Perform exponential smoothing with trend and seasonality on this data. Use 0.20 for all smoothing constants. What is your forecast for the first quarter of 2011?

Use the following equations:

$S_t = \alpha\,(D_t/C_{t-N}) + (1-\alpha)(S_{t-1} + G_{t-1})$, $\quad G_t = \beta\,(S_t - S_{t-1}) + (1-\beta)G_{t-1}$,

$C_t = \gamma\,(D_t/S_t) + (1-\gamma)C_{t-N}$, $\quad F_{t+1} = (S_t + G_t)C_{t+1-N}$

Using 0.2 for all smoothing factors, we can calculate S, G, C and F as follows:

$S_1 = 0.2(4/0.6502) + 0.8(7.083 + 0.333) = 7.1632.$

$G_1 = 0.2(7.1632 - 7.083) + 0.8(0.333) = 0.28244.$

$C_1 = 0.2(4/7.1632) + 0.8(0.6502) = 0.63184.$

$F_2 = (7.1632 + 0.28244)1.3793 = 10.26977.$

In similar manner, we may calculate S, G, C and F for other values of t, as shown in the following table:

Year-Quarter	Number of cases resolved	S_t	G_t	C_t	F_t
		7.083	0.333		
2007-1	4	7.1632	0.2824	0.6318	
2007-2	13	7.8415	0.3616	1.4350	10.2698
2007-3	5	7.8980	0.3006	0.7257	6.1425
2007-4	9	8.0322	0.2673	1.2015	10.0162
2008-1	5	8.2223	0.2519	0.6271	5.2440
2008-2	28	10.6817	0.6934	1.6723	12.1605
2008-3	8	11.3050	0.6794	0.7221	8.2544
2008-4	27	14.0820	1.0989	1.3446	14.3987

(contd.)

Year-Quarter	Number of cases resolved	S_t	G_t	C_t	F_t
		7.083	**0.333**		
2009-1	16	17.2476	1.5122	0.6872	9.5199
2009-2	17	17.0411	1.1685	1.5373	31.3715
2009-3	16	18.9994	1.3265	0.7461	13.1483
2009-4	15	18.4918	0.9596	1.2379	27.3309
2010-1	8	17.8894	0.6472	0.6392	13.3672
2010-2	12	16.3905	0.2180	1.3763	28.4970
2010-3	9	15.6994	0.0362	0.7115	12.3911
2010-4	11	14.3656	-0.2378	1.1435	19.4797
				Forecast for Quarter 1 2011	**9.0306**

SOLUTION MANUAL

CHAPTER 11 Index Numbers

1 The prices of three brands of LED televisions and the quantities sold at Brothers Electrical for 2008 and 2010 are as follows.

Brand	2008		2010	
	Price ($)	Quantity	Price ($)	Quantity
Sony	8 900	80	5 500	130
Samsung	7 500	90	4 200	170
LG	6 600	110	3 900	220

Using 2008 as the base year, calculate:

(a) Average of relative price index for the three brands for 2010.

$$\textit{Average of Relative Price Index} = \frac{\Sigma \frac{p_{2010}}{p_{2008}} \times 100}{k}$$

$= [(5\,500/8\,900) \times 100 + (4\,200/7\,500) \times 100 + (3\,900/6\,600) \times 100]/3$
$= (61.798 + 56 + 59.091)/3 = 58.963.$

(b) Laspeyres quantity index for 2010 and interpret.

$$\textit{Laspeyres Quantity Index} = \frac{\Sigma q_{2010} p_{2008}}{\Sigma q_{2008} p_{2008}} \times 100$$

$q_{2008}p_{2008}$	$q_{2010}p_{2008}$
$80 \times 8\,900 = 712\,000$	$130 \times 8\,900 = 1\,157\,000$
$90 \times 7\,500 = 675\,000$	$170 \times 7\,500 = 1\,275\,000$
$110 \times 6\,600 = 726\,000$	$220 \times 6\,600 = 1\,452\,000$
$\Sigma = 2\,113\,000$	$\Sigma = 3\,884\,000$

Laspeyres Quantity Index = (3884000/2113000) × 100 = 183.81
Hence, the total quantity of LED TVs sold has increased by 83.81% from 2008 to 2010.

2 The prices of three LCD TV models and their quantities sold by a company in 2009 and 2010 are as follows.

LCD TV Model	2009		2010	
	Price ($)	Quantity	Price ($)	Quantity
A	3 000	160	2 700	170
B	2 200	165	1 900	160
C	1 700	300	1 500	330

Using 2009 as the base year, calculate:

(a) Average of relative price index for the three models for 2010.

$$\textit{Average of Relative Price Index} = \frac{\Sigma \frac{p_{2010}}{p_{2009}} \times 100}{k}$$

$= [(2\,700/3\,000) \times 100 + (1\,900/2\,200) \times 100 + (1\,500/1\,700) \times 100]/3$
$= (90 + 86.364 + 88.235)/3 = 88.20.$

(b) Laspeyres quantity index for 2010 and give your comment.

$$\textit{Laspeyres Quantity Index} = \frac{\Sigma\, q_{2010} p_{2009}}{\Sigma\, q_{2009} p_{2009}} \times 100$$

$q_{2009}p_{2009}$	$q_{2010}p_{2009}$
$160 \times 3\,000 = 480\,000$	$170 \times 3\,000 = 510\,000$
$165 \times 2\,200 = 363\,000$	$160 \times 2\,200 = 352\,000$
$300 \times 1\,700 = 510\,000$	$330 \times 1\,700 = 561\,000$
$\Sigma = 1\,353\,000$	$\Sigma = 1\,423\,000$

Laspeyres Quantity Index = $(1\,423\,000/1\,353\,000) \times 100 = 105.17$
Hence, the total quantity of three LCD TV models sold has increased by 5.17% from 2009 to 2010.

(c) If the change in price for model B in 2011 is a decrease of 15% relative to 2009, find the price of model B for 2011.

If the price for model B in 2011 is decreased by 15% relative to 2009, then

$$\begin{aligned} p_{\text{B, 2011}} &= p_{\text{B, 2009}} \times (1.00 - 0.15) \\ &= 2\,200 \times 0.85 = 1\,870. \end{aligned}$$

3 A researcher has gathered the following information on the prices and quantities for a number of imported fruits for the years 2009 and 2010.

Fruit	Quantity ('000 kg)		Price per kg ($)	
	2009	2010	2009	2010
Apple	46.8	48.1	5.37	5.55
Grape	35.5	34.9	13.55	14.28
Pear	34.4	38.2	7.99	6.99

(a) Using the year 2009 as the base period, calculate an unweighted average of relative quantity index for the year 2010 and interpret.

$$\textit{Average of Relative Quantity Index} = \frac{\Sigma \frac{q_{2010}}{q_{2009}} \times 100}{k}$$

$= [(48.1/46.8) \times 100 + (34.9/35.5) \times 100 + (38.2/34.4) \times 100]/3$
$= (102.78 + 98.31 + 111.05)/3 = 104.05.$

Hence, on the average, the quantity of fruits sold has increased by 4.05% from 2009 to 2010.

(b) Compute the Laspeyres quantity index for the year 2010.

$$\textit{Laspeyres Quantity Index} = \frac{\Sigma\, q_{2010} p_{2009}}{\Sigma\, q_{2009} p_{2009}} \times 100$$

$q_{2009}p_{2009}$	$q_{2010}p_{2009}$
$46.8 \times 5.37 = 251.316$	$48.1 \times 5.37 = 258.297$
$35.5 \times 13.55 = 481.025$	$34.9 \times 13.55 = 472.895$
$34.4 \times 7.99 = 274.856$	$38.2 \times 7.99 = 305.218$
$\Sigma = 1007.197$	$\Sigma = 1036.41$

Laspeyres Quantity Index = $(1036.41/1007.197) \times 100 = 102.90$.
Hence, the total quantity of fruits has increased by 2.9% from 2009 to 2010.

(c) Compute the Paasche price index for year 2010.

$$\textit{Paasche Price Index} = \frac{\Sigma\, p_{2010} q_{2010}}{\Sigma\, p_{2009} q_{2010}} \times 100$$

$p_{2009}q_{2010}$	$p_{2010}q_{2010}$
$5.37 \times 48.1 = 258.297$	$5.55 \times 48.1 = 266.955$
$13.55 \times 34.9 = 472.895$	$14.28 \times 34.9 = 498.372$
$7.99 \times 38.2 = 305.218$	$6.99 \times 38.2 = 267.018$
$\Sigma = 1036.41$	$\Sigma = 1032.345$

Paasche Price Index = $(1032.345/1036.41) \times 100 = 99.61$.
Hence, the total expenditure on fruits in 2010 has decreased by 0.39% (100-99.61) compared to 2009.

4 The following table shows the price ($/kg) and quantity (kg) of four grocery items purchased by a family for the years 2008 and 2009.

Item	2008		2009	
	Price	Quantity	Price	Quantity
Sugar	1.25	100	1.45	120
Flour	1.10	120	1.25	160
Rice	3.50	180	3.90	220
Cooking Oil	2.50	80	2.75	60

(a) Determine the average of relative price index for the four items for 2009.

$$\textit{Average of Relative Price Index} = \frac{\Sigma \frac{p_{209}}{p_{2008}} \times 100}{k}$$

$= [(1.45/1.25) \times 100 + (1.25/1.10) \times 100 + (3.90/3.50) \times 100 + (2.75/2.5) \times 100]/4$
$= (116 + 113.64 + 111.43 + 110)/4 = 112.77.$

(b) Compute the Paasche quantity index for 2009.

$$\textit{Paasche Quantity Index} = \frac{\Sigma\, q_{2009} p_{2009}}{\Sigma\, q_{2008} p_{2009}} \times 100$$

$q_{2008}p_{2009}$	$q_{2009}p_{2009}$
$100 \times 1.45 = 145$	$120 \times 1.45 = 174$
$120 \times 1.25 = 150$	$160 \times 1.25 = 200$
$180 \times 3.90 = 702$	$220 \times 3.90 = 858$
$80 \times 2.75 = 220$	$60 \times 2.75 = 165$
$\Sigma = 1\,217$	$\Sigma = 1\,397$

Paasche Quantity Index = $(1\,397/1\,217) \times 100 = 114.79$.

(c) Compute the Laspeyres price index for 2009. Interpret your answer.

$$\textit{Laspeyres Price Index} = \frac{\Sigma\, p_{2009} q_{2008}}{\Sigma\, p_{2008} q_{2008}} \times 100$$

$p_{2008}q_{2008}$	$p_{2009}q_{2008}$
$1.25 \times 100 = 125$	$1.45 \times 100 = 145$
$1.10 \times 120 = 132$	$1.25 \times 120 = 150$
$3.50 \times 180 = 630$	$3.90 \times 180 = 702$
$2.50 \times 80 = 200$	$2.75 \times 80 = 220$
$\Sigma = 1\,087$	$\Sigma = 1\,217$

Laspeyres Price Index = $(1\,217/1\,087) \times 100 = 111.96$.
Hence, total expenditure on four items has increased by 11.96% from 2008 to 2009.

5 The table below shows the average monthly price and quantity for four models of washing machines of a particular brand sold by a home appliances company in 2009 and 2010.

Washing Machine Model (capacity)	Price ($)		Quantity	
	2009	2010	2009	2010
A (10 kg)	790	590	125	95
B (12 kg)	890	790	220	310
C (14 kg)	1190	990	170	210
D (16 kg)	1290	1190	95	115

(a) Calculate the relative price index for model B for 2010 using 2009 as the base year.

Relative price index (model B) $= \dfrac{p_{2010}}{p_{2009}} \times 100 = (790/890) \times 100 = 88.76.$

(b) Calculate the aggregate price index for 2010 using 2009 as the base year, and explain its meaning.

$$\textit{Aggregate Price Index} = \frac{\Sigma p_{2010}}{\Sigma p_{2009}} \times 100$$

$= [(590 + 790 + 990 + 1\,190)/(790 + 890 + 1\,190 + 1\,290)] \times 100 = (3\,560/4\,160) \times 100$
$= 85.58.$

Hence, the price of washing machines has decreased by 14.42% from 2009 to 2010.

(c) Compute the Laspeyres price index for 2010.

$$\textit{Laspeyres Price Index} = \frac{\Sigma p_{2010} q_{2009}}{\Sigma p_{2009} q_{2009}} \times 100$$

$p_{2009}q_{2009}$	$p_{2010}q_{2009}$
$790 \times 125 = 98\,750$	$590 \times 125 = 73\,750$
$890 \times 220 = 195\,800$	$790 \times 220 = 173\,800$
$1190 \times 170 = 202\,300$	$990 \times 170 = 168\,300$
$1290 \times 95 = 122\,550$	$1190 \times 95 = 113\,050$
$\Sigma = 619\,400$	$\Sigma = 528\,900$

Laspeyres Price Index $= (528\,900/619\,400) \times 100 = 85.39.$

(d) Compute the Paasche quantity index for 2010.

$$\textit{Paasche Quantity Index} = \frac{\Sigma q_{2010} p_{2010}}{\Sigma q_{2009} p_{2010}} \times 100$$

$q_{2009}p_{2010}$	$q_{2010}p_{2010}$
$125 \times 590 = 73\,750$	$95 \times 590 = 56\,050$
$220 \times 790 = 173\,800$	$310 \times 790 = 244\,900$
$170 \times 990 = 168\,300$	$210 \times 990 = 207\,900$
$95 \times 1190 = 113\,050$	$115 \times 1190 = 136\,850$
$\Sigma = 528\,900$	$\Sigma = 645\,700$

Paasche Quantity Index $= (645\,700/528\,900) \times 100 = 122.08.$

6 The prices and quantities of three different consumer items for the period 2008–2010 are shown in the table below.

Item	Price ($/kg)			Quantity (kg)		
	2008	2009	2010	2008	2009	2010
I	14.00	17.80	26.00	2 100	3 400	4 800
II	11.80	15.20	20.40	1 900	1 740	2 000
III	12.60	14.00	20.00	1 060	980	1 400

(a) Using 2009 as the base year, compute the Laspeyres price index for 2010 and interpret the result.

$$\textit{Laspeyres Price Index} = \frac{\sum p_{2010}\, q_{2009}}{\sum p_{2009}\, q_{2009}} \times 100$$

$p_{2009}q_{2009}$	$p_{2010}q_{2009}$
17.80 × 3400 = 60 520	26.00 × 3400 = 88 400
15.20 × 1740 = 26 448	20.40 × 1740 = 35 496
14.00 × 980 = 13 720	20.00 × 980 = 19 600
Σ = 100 688	Σ = 143 496

Laspeyres Price Index = (143 496/100 688) × 100 = 142.52.
Hence, total expenditure on three consumer items has increased by 42.52% from 2009 to 2010.

(b) Using 2008 as the base year, compute the Paasche quantity index for 2010 and interpret the result.

$$\textit{Paasche Quantity Index} = \frac{\sum q_{2010}\, p_{2010}}{\sum q_{2008}\, p_{2010}} \times 100$$

$q_{2008}p_{2010}$	$q_{2010}p_{2010}$
2100 × 26.00 = 54 600	4800 × 26.00 = 124 800
1900 × 20.40 = 38 760	2000 × 20.40 = 40 800
1060 × 20.00 = 21 200	1400 × 20.00 = 28 000
Σ = 114 560	Σ = 193 600

Paasche Quantity Index = (193 600/114 560) × 100 = 168.99.
Hence, total quantity of the three items consumed in 2010 has increased by 68.99% compared to 2008.

(c) Compute the aggregate quantity index for 2009 using 2008 as the base year.

$$\textit{Aggregate Quantity Index} = \frac{\sum q_{2009}}{\sum q_{2008}} \times 100$$

= [(3 400 + 1 740 + 980)/(2 100 + 1 900 + 1 060)] × 100 = (6 120/5 060) × 100 = 120.95.

(d) If the change in price for item I in 2011 is decreased by 10% relative to 2010, determine the price of the item in 2011.

If the price for item I in 2011 is decreased by 10% relative to 2010, then

$$p_{\text{I, 2011}} = p_{\text{I, 2010}} \times (1.00 - 0.10)$$
$$= 26.00 \times 0.90 = 23.40.$$