Statistics for Business and Social Sciences

Table of contents

Welcome

Welcome to **Statistics for Business and Social Sciences**, a comprehensive guide to understanding and harnessing the power of statistics in the context of business and the social sciences. This preface serves as an introduction to the rich and practical content that you are about to embark upon.

In a world increasingly driven by data and information, the ability to make sense of statistics is an invaluable skill. This course has been designed to provide students with the fundamental tools and knowledge necessary to not only understand statistics but also to apply them to real-world problems in business and social sciences. Whether you are a student, a professional, or simply someone interested in the world of data, this course will equip you with the essential skills to navigate the statistical landscape.

Course Information

At the heart of this course lies the commitment to empower students with the ability to:

1. **Describe the concepts** involved in solving problems related to statistics for business and social sciences (C2).
2. **Determine the appropriate methods** to address statistical problems encountered in business and social sciences (C5).
3. **Demonstrate interpersonal skills** through group work related to statistics for business and social sciences (A3).

As you dive into the course, these outcomes will become your guiding beacons, helping you measure your progress and the depth of your understanding.

Course Description

Our journey through the world of statistics will begin by introducing you to the basic and intermediate methods of data analysis. You will explore the realms of descriptive and inferential statistics, which encompass numerical descriptive measures, estimation, hypothesis testing, and various statistical techniques. A significant emphasis is placed on practicality, which will be evident as you delve into the use of statistical software and interpretation of output.

Syllabus Content

The course content is divided into five distinct chapters, each building upon the foundation laid by the preceding one:

- **Chapter 1: Introduction to Statistics** provides the groundwork by defining what statistics is, explaining the types of variables, data, and methods of data collection.
- **Chapter 2: Descriptive Statistics** explores the art of organizing data and calculating various measures of central tendency, variation, skewness, and position.
- **Chapter 3: Estimation** delves into the concept of sampling distribution and interval estimation for means, both for independent and dependent samples.
- **Chapter 4: Hypothesis Testing** equips you with the skills needed to test hypotheses about means and variances in various scenarios, including one-way analysis of variance and testing for independence.
- **Chapter 5: Bivariate Analysis** ventures into the world of relationships between variables, covering correlation, simple linear regression, and the estimation of linear regression using the least squares method.

Each chapter serves as a stepping stone, progressively building your knowledge and competence in the world of statistics.

Throughout your journey, you'll find numerous examples, exercises, and real-world applications to reinforce your understanding. Remember, this course isn't just about theory; it's about practical skills that you can apply to real-life situations.

We hope this course becomes your gateway to the world of statistics, enabling you to approach business and social sciences with newfound confidence and precision. So, fasten your seatbelts, embrace the challenges, and get ready to embark on a statistical adventure that will broaden your horizons and empower you with valuable skills.

Preface

Welcome to “Statistics for Business and Social Sciences.” This book is designed to be your comprehensive guide to the world of statistics, tailored specifically to the curriculum of Universiti Teknologi MARA’s undergraduate program. Whether you are a business major, social sciences enthusiast, or a student in any discipline, this book will equip you with the essential knowledge and skills to harness the power of statistics.

In the fast-paced and data-driven world we live in, the ability to make informed decisions is paramount. Statistics is the language of data, and its principles are woven into countless aspects of our daily lives. Understanding statistics is not just an academic endeavor; it’s a practical tool that can empower you to navigate the complexities of your chosen field.

The book begins with “Introduction to Statistics,” setting the stage by exploring the fundamental concepts, the role of statistics in research and decision-making, and the ethical considerations of data analysis. We then delve into “Descriptive Statistics,” where we unveil the art of summarizing and presenting data in a meaningful way.

“Estimation” and “Hypothesis Testing” are two pivotal chapters that bridge the gap between data and decision-making. You’ll learn how to make educated guesses about population parameters and how to rigorously test hypotheses using real-world examples.

The journey culminates in “Bivariate Analysis,” where we explore the relationships between two variables, offering insights into correlation, regression, and their applications in business and social sciences.

Each chapter is designed with clarity in mind, with practical examples and exercises to reinforce your understanding. Additionally, we provide a range of resources, including data sets and online tools, to help you apply your knowledge beyond the textbook.

As you embark on this statistical voyage, remember that learning statistics is not just about mastering formulas and techniques; it's about developing a critical mindset, where you can question, interpret, and draw meaningful conclusions from data. This book is your trusted companion in that journey.

We hope this book becomes an invaluable resource, both in your academic pursuits and in your future professional endeavors. Embrace the world of statistics with confidence, and discover how it can enhance your ability to analyze, interpret, and contribute to the business and social sciences fields.

Statistics may appear daunting at first, but with patience and persistence, it can be a fascinating and empowering discipline. So, let's embark on this enlightening journey together, and may this book be your steadfast guide. Enjoy your exploration of "Statistics for Business and Social Sciences."

Rosidah Ahmad

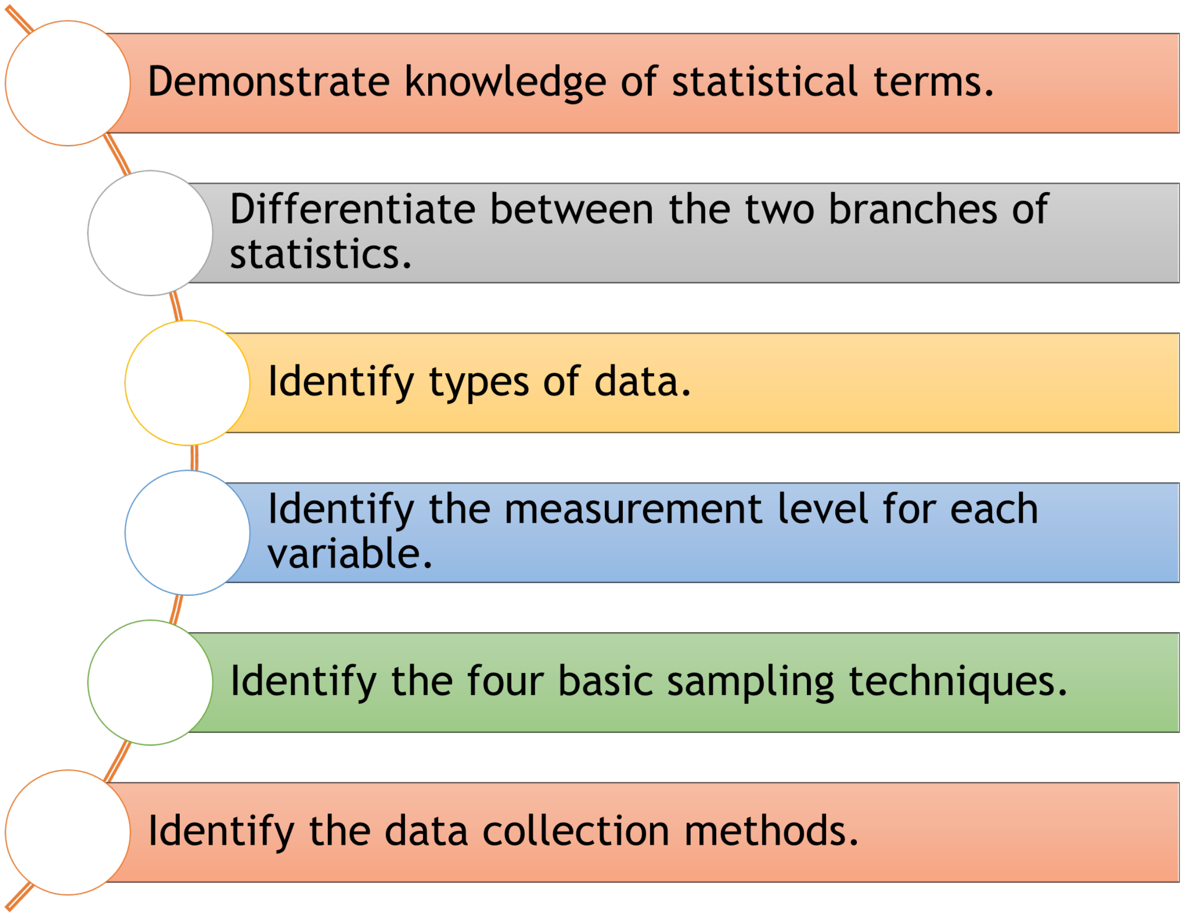
Kamarul Ariffin Mansor

Ida Normaya Mohd Nasir

Norin Rahayu Shamsuddin

1 Introduction to Statistics

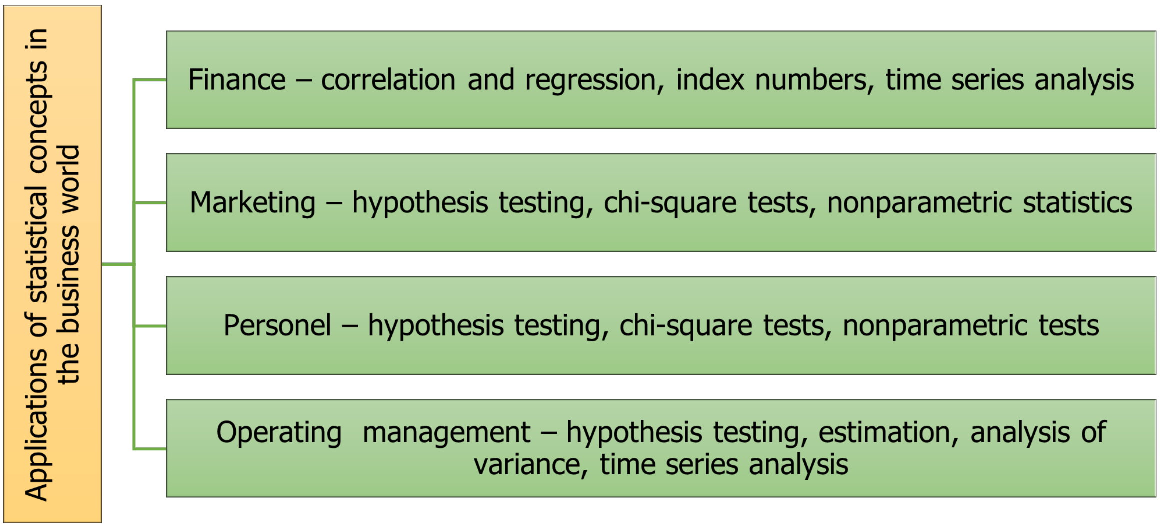
Learning Objectives



- Demonstrate knowledge of statistical terms.
- Differentiate between the two branches of statistics.
- Identify types of data.
- Identify the measurement level for each variable.
- Identify the four basic sampling techniques.
- Identify the data collection methods.

1.1 Introduction

Why study statistics?


Applications of statistical concepts in the business world
Finance – correlation and regression, index numbers, time series analysis
Marketing – hypothesis testing, chi-square tests, nonparametric statistics
Personel – hypothesis testing, chi-square tests, nonparametric tests
Operating management – hypothesis testing, estimation, analysis of variance, time series analysis

1.2 What is Statistics?

Statistics

- Defined as the science of collecting, organizing, presenting, analyzing, and interpreting data to make more effective decisions.

Data

- The values that the variables can assume.

Variable

- An item of interest that can assume different values.

Data

- May consist of more than one variable depend on the study

Data set

- A collection of a set of data values.

Population

- A large of data which consists of all element that are being studied.

Sample

- A subset of population which the elements is selected from a population

Census

- If a researcher would like to make a complete sampling by using all elements in the population.

Statistical analysis

- Used to manipulate, summarize, and investigate data, so that useful decision-making information results.

Example 1.1

Determine the population, sample, and variable(s).

a) The Dean from College YY would like to determine students' performance through online distance learning (ODL). From 1000 students, the dean decides to select only 300 students as a respondent. The information about the number of hours spent and assessment marks are collected.

b) A headmaster of School Y conducted a study on students' satisfaction (strongly disagree=1, disagree=2, neutral=3, agree=4, strongly agree=5) with online distance learning conducted by their teachers. The

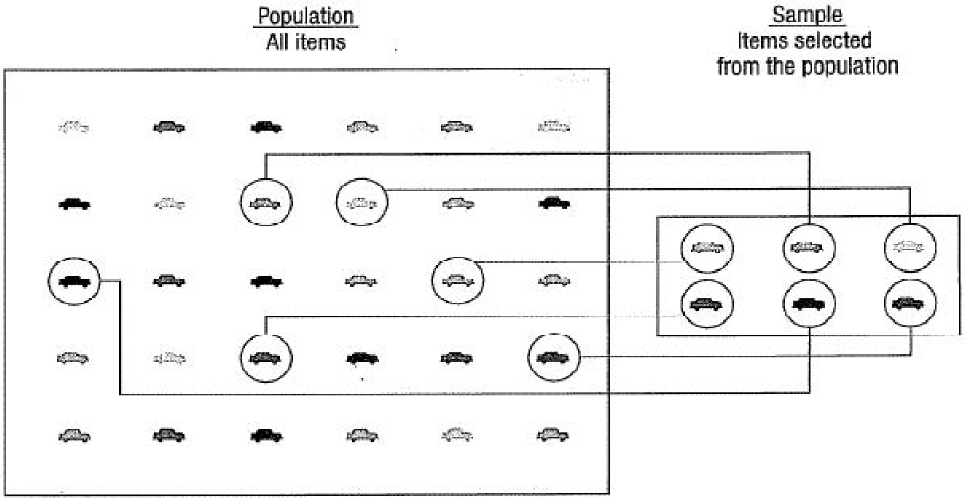


Figure 1.1: Figure 1.1: Population and Sample

headmaster selects only 10 out of 50 classes in School Y. The study included all students in these classes.

Answer

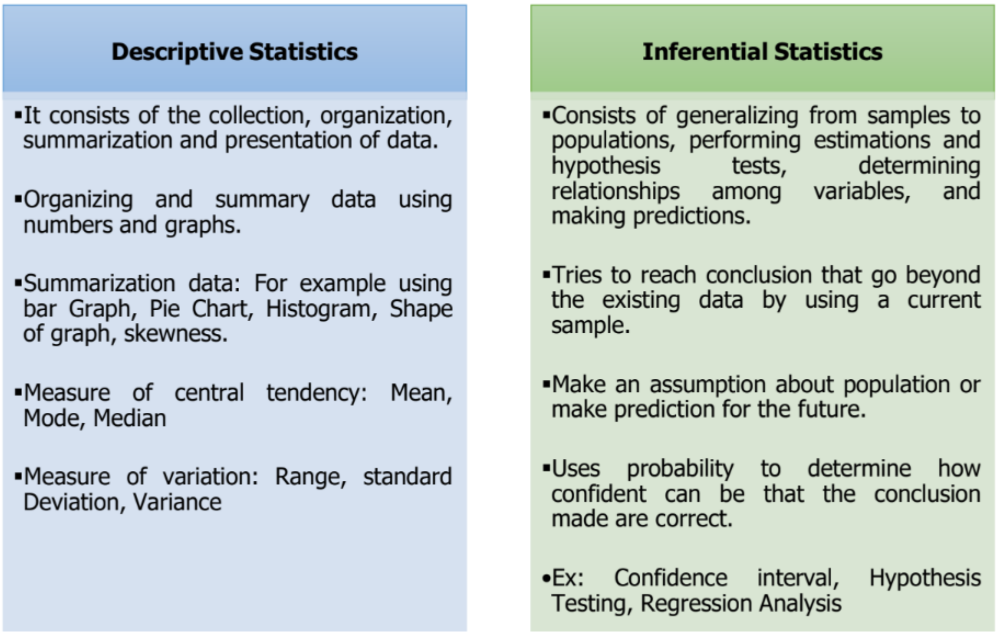
a) Population: All 1000 students from College YY. Sample: 300 students from College YY. Variable: i) Number of hours spent ii) Assessment marks

b) Population: All students from 50 classes of School Y. Sample: Students from selected 10 classes. Variable: Students' satisfaction

1.3 Type of statistics

Generally, **statistics** is divided into two broad categories, depending on how data are used. The two categories are **Descriptive Statistics** and **Inferential Statistics**. A **Descriptive Statistics** describe a *summary* information about variables in data. While, **Inferential Statistics** uses sample data to make an inference or draw a conclusion about the population. The description of Descriptive Statistics and Inferential Statistics shown in Table 1.1.

Example 1.2



Descriptive Statistics

- It consists of the collection, organization, summarization and presentation of data.
- Organizing and summary data using numbers and graphs.
- Summarization data: For example using bar Graph, Pie Chart, Histogram, Shape of graph, skewness.
- Measure of central tendency: Mean, Mode, Median
- Measure of variation: Range, standard Deviation, Variance

Inferential Statistics

- Consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.
- Tries to reach conclusion that go beyond the existing data by using a current sample.
- Make an assumption about population or make prediction for the future.
- Uses probability to determine how confident can be that the conclusion made are correct.
- Ex: Confidence interval, Hypothesis Testing, Regression Analysis

Figure 1.2: Table 1.1 Descriptive Statistics and Inferential Statistics

a) The average donation received from Five Top Company was RM50000.
b) In a research study found that anxiety with behavioral intention has a significant inverse causal effect on the behavior of students using communication technology.

Answer

a) Descriptive Statistics
b) Inferential Statistics

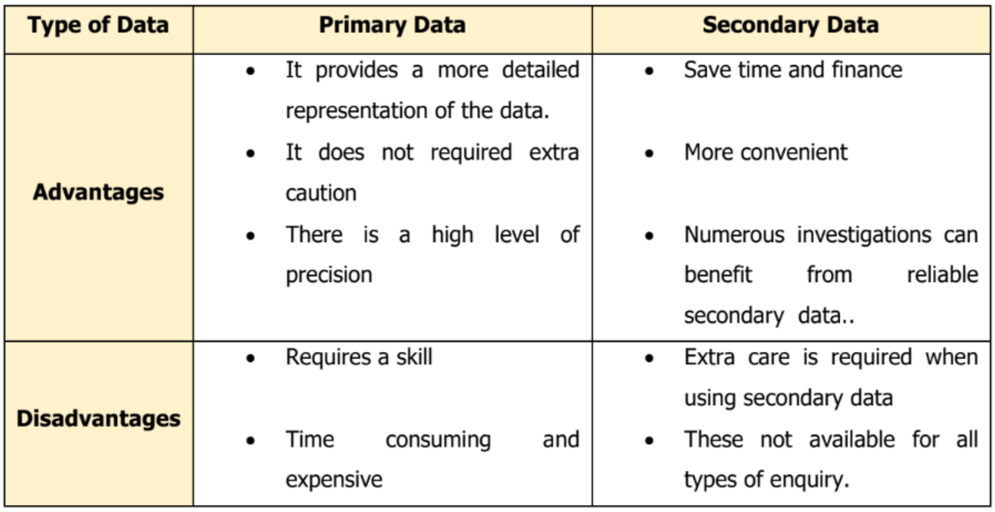
1.4 Type of data

The term “Data” refers to a collection of measurements made on one or more observational units that used to describe situations or events. Statistical data are usually obtained by counting or measuring items. Depending on the sources, statistical data are classified into two types; **Primary Data** and **Secondary Data**.

a) *Primary data* is a data that are collected for the first time and are thus original in nature.

b) *Secondary data* have already been compiled or collected by some other persons and are available for statistical analysis.

The advantages and disadvantages of primary and secondary data shown in Table 1.2.



Type of Data	Primary Data	Secondary Data
Advantages	- It provides a more detailed representation of the data. - It does not required extra caution - There is a high level of precision	- Save time and finance - More convenient - Numerous investigations can benefit from reliable secondary data..
Disadvantages	- Requires a skill - Time consuming and expensive	- Extra care is required when using secondary data - These not available for all types of enquiry.

Figure 1.3: Table 1.2 Advantages and disadvantage of primary and secondary data

Example 1.3

Describe the type of data below.

a) A statistics textbook
b) A mailed questionnaire

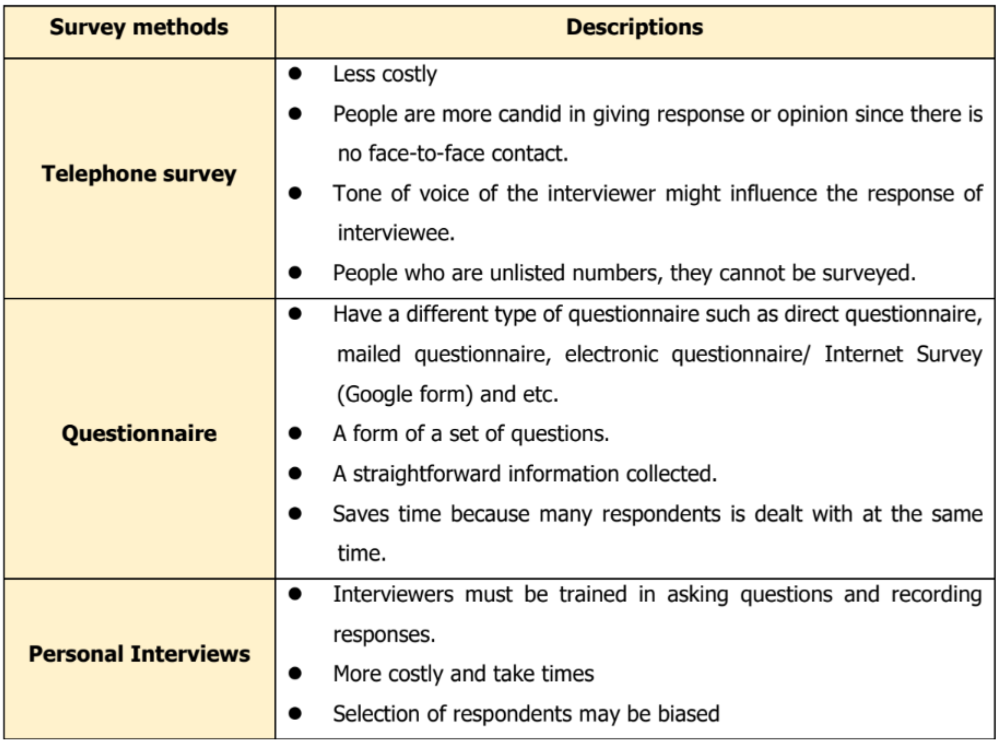
Answer

a) Secondary Data
b) Primary Data

1.5 Data collection

Data are collected in a variety of ways. For **primary data**, one of the most common method is through the use of survey.

Survey is a research process of collecting a data that can be done by using a variety of method. The most common methods are telephone surveys, questionnaire survey and the personal interviews. The descriptions of these methods are shown in Table 1.3.



Survey methods	Descriptions
Telephone survey	● Less costly ● People are more candid in giving response or opinion since there is no face-to-face contact. ● Tone of voice of the interviewer might influence the response of interviewee. ● People who are unlisted numbers, they cannot be surveyed.
Questionnaire	● Have a different type of questionnaire such as direct questionnaire, mailed questionnaire, electronic questionnaire/ Internet Survey (Google form) and etc. ● A form of a set of questions. ● A straightforward information collected. ● Saves time because many respondents is dealt with at the same time.
Personal Interviews	● Interviewers must be trained in asking questions and recording responses. ● More costly and take times ● Selection of respondents may be biased

Figure 1.4: Table 1.3 Descriptions of telephone surveys, questionnaire survey and personal interviews

Secondary data are collected from readily available sources such as websites, article journals, books, etc. It supplies second-hand data collection from other sources, either from individuals or an organization. Among the top sources of secondary data are:

a) Journal articles that comment on or analyse research
b) Textbooks
c) Dictionaries and encyclopaedias
d) Book and interpret, analyse
e) Political commentary
f) Biographies
g) Dissertations
h) Newspaper editorial/opinion pieces

1.6 Variables and types of variables

Variables can be classified as Quantitative Variable or Qualitative Variable.

1.6.1 Quantitative Variables

Quantitative data is an observation that are measured on a numerical scale. Basically, the quantitative data are in the form of values, percentage, frequency, or numbers. This type of data can be visualized using diagram such as tables, graphs, and histogram.

If the variables in the data are being studied, the variables that are report numerically is called quantitative variable.

Quantitative variables are always in numbers and are the result of counting or measuring attributes of a population. Quantitative variable can be separated into two subgroups:

> **Discrete** can assume only integer value (if it is the result of counting, examples the number of students of a given ethnic group in a class, the number of books on a shelf)
>
> **Continuous** can assume any value over a continuous range of possibilities (if it is the result of measuring, examples distance traveled, weight of luggage)

1.6.2 Qualitative Variables

Qualitative data is opposite with quantitative data. Qualitative data provide varieties of items in terms of categories base. It is generally described by words or letters. For example, in qualitative data may contains information about gender, age category or pass or fail.However, if the characteristics or variables being studied is in categorical or non-numerical it is called as qualitative variable.

Qualitative variables can be separated into two subgroups:

> **dichotomic** (if it takes the form of a word with two options (gender - male or female)

polynomic (if it takes the form of a word with more than two options (education - primary school, secondary school, and university)

Usually, a researcher will represent a qualitative variable with a proportion or percentage, while a pie chart, pareto chart, or bar chart will use to visualize a qualitative variables.

> ℹ Note
>
> Remember, in dealing with qualitative variable, calculating the mean or average makes no sense.

Example 1.5

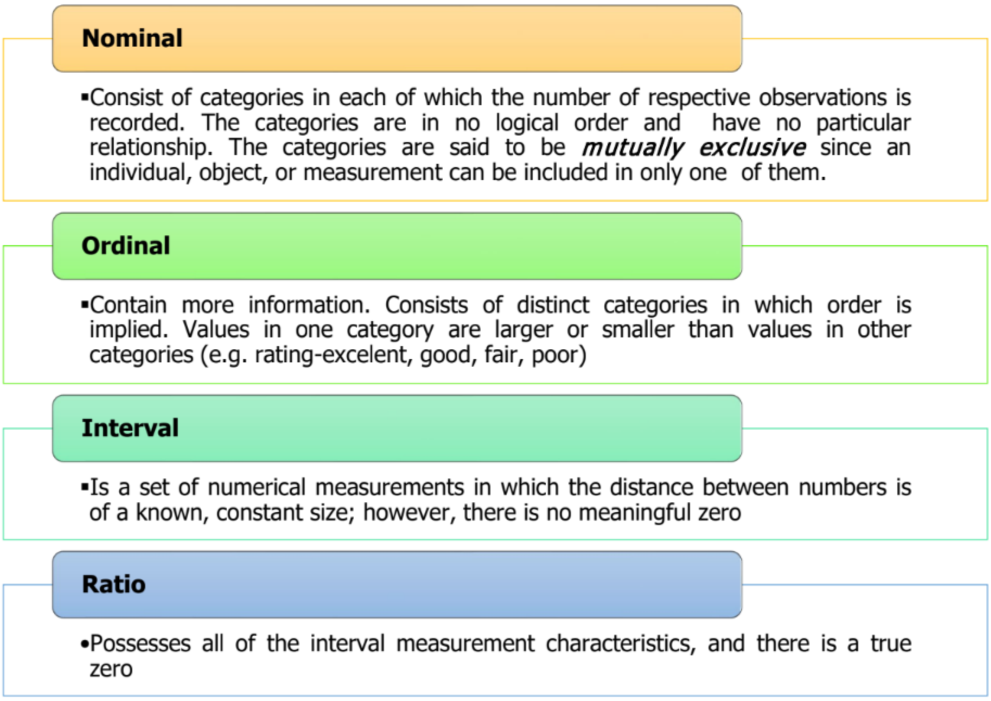
State whether the following quantitative or qualitative variable.

a) Number of diabetes patients
b) Pizza sizes (small, medium, and large)
c) Cholesterol count

Answer

a) Quantitative (Discrete)
b) Qualitative
c) Quantitative (Continuous)

1.7 Level of measurement



Nominal

- Consist of categories in each of which the number of respective observations is recorded. The categories are in no logical order and have no particular relationship. The categories are said to be ***mutually exclusive*** since an individual, object, or measurement can be included in only one of them.

Ordinal

- Contain more information. Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. rating-excelent, good, fair, poor)

Interval

- Is a set of numerical measurements in which the distance between numbers is of a known, constant size; however, there is no meaningful zero

Ratio

- Possesses all of the interval measurement characteristics, and there is a true zero

Example 1.6

State the level of measurement for each variable below:

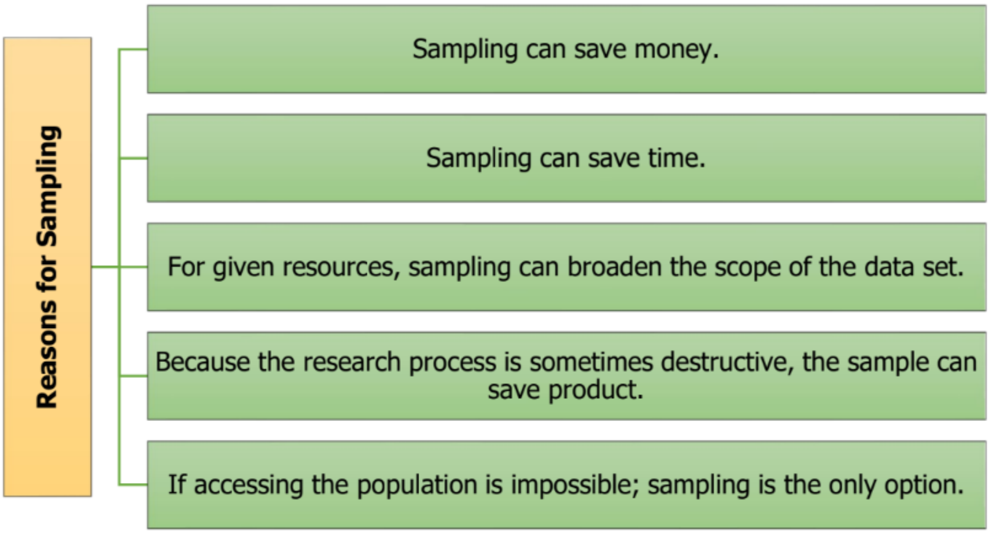
a) The number of students in KAM2283A
b) Temperature in Malaysia
c) Stress level (Mild, Medium, Severe, Very Severe)
d) Food preference

Answer

a) Ratio
b) Interval
c) Ordinal
d) Nominal

1.8 Sampling

Sampling is a process of taking a subset of element from a population. The sample taken from the sampling process have a same characteristic with its population. However, the samples selected are not a perfect representative of population depending on where they are selected. Therefore, there always occur some error in the result of analysis called as sampling error. A **sampling error** is the difference between the results obtained from a sample and the results obtained from the population.



Reasons for Sampling

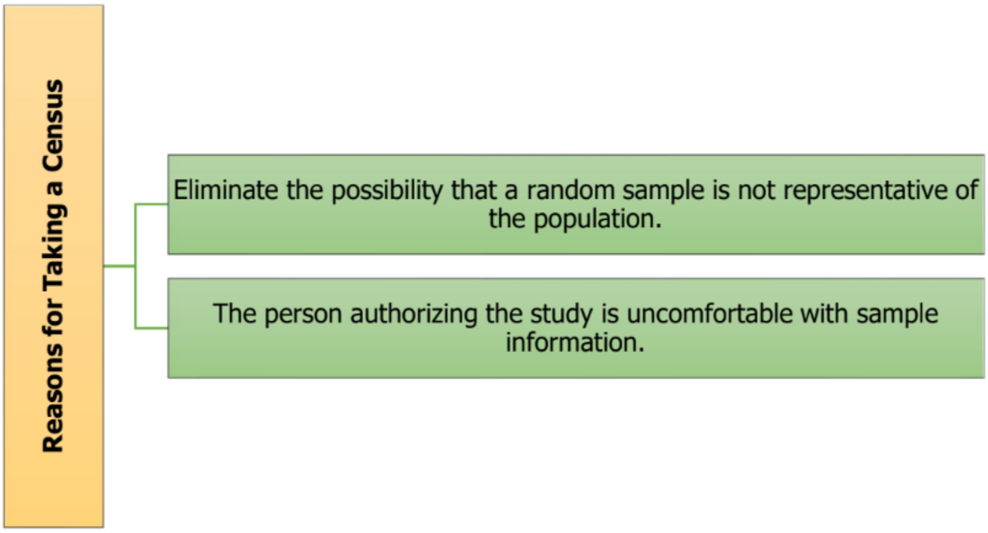
Sampling can save money.

Sampling can save time.

For given resources, sampling can broaden the scope of the data set.

Because the research process is sometimes destructive, the sample can save product.

If accessing the population is impossible; sampling is the only option.



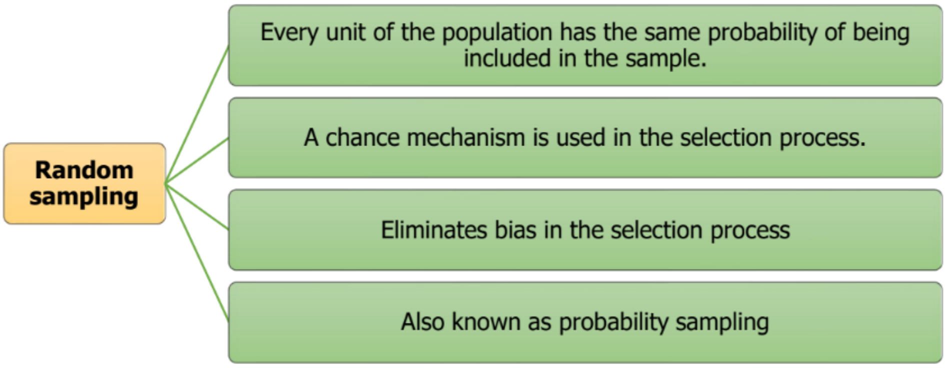
Reasons for Taking a Census

Eliminate the possibility that a random sample is not representative of the population.

The person authorizing the study is uncomfortable with sample information.

1.9 Sampling Method

Random vs Non-Random sampling



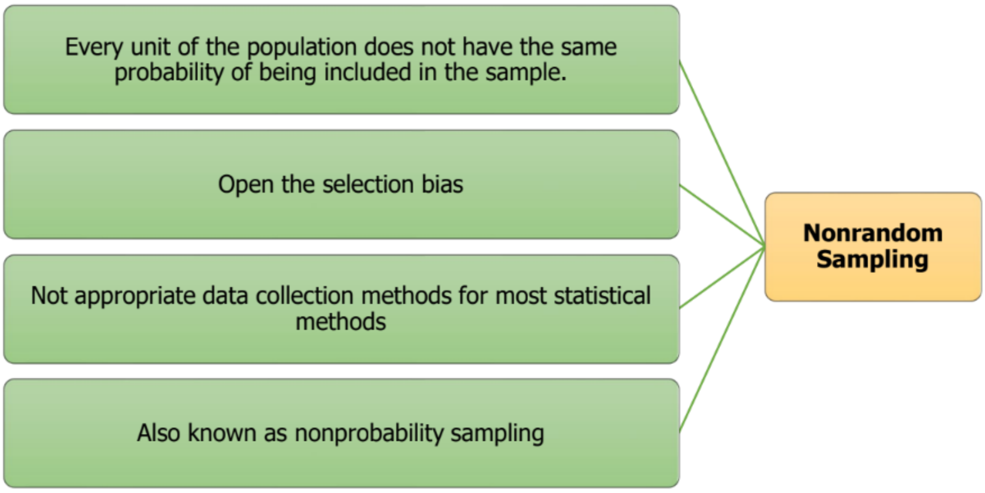
Random sampling

Every unit of the population has the same probability of being included in the sample.

A chance mechanism is used in the selection process.

Eliminates bias in the selection process

Also known as probability sampling



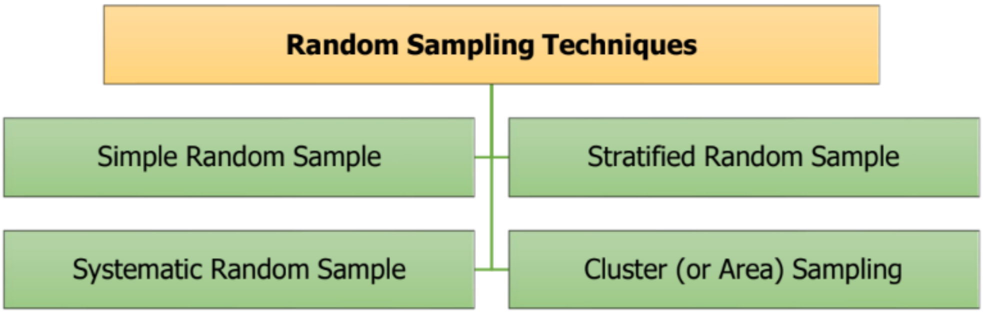
Every unit of the population does not have the same probability of being included in the sample.

Open the selection bias

Not appropriate data collection methods for most statistical methods

Also known as nonprobability sampling

Nonrandom Sampling



Random Sampling Techniques

Simple Random Sample

Stratified Random Sample

Systematic Random Sample

Cluster (or Area) Sampling

1.9.1 Simple Random Sample

- Number each frame unit from 1 to N.
- Use a random number table or a random number generator to select n distinct numbers between 1 and N, inclusively.
- Easier to perform for small populations.

- Cumbersome for large populations

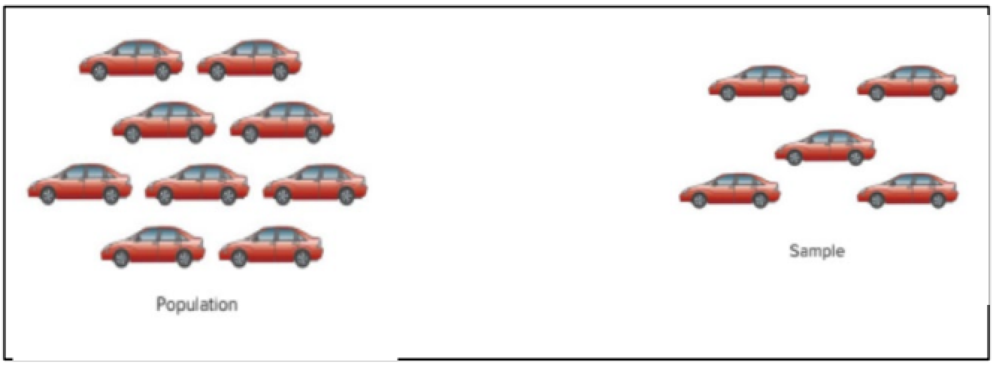


Figure 1.5: Source: Allan G. Bluman (2010)

1.9.2 Systematic Sampling

- Convenient and relatively easy to administer.
- Population elements are an ordered sequence (at least, conceptually).
- The first sample element is selected randomly from the first k population elements.
- Thereafter, sample elements are selected at a constant interval, k, from the ordered sequence frame.

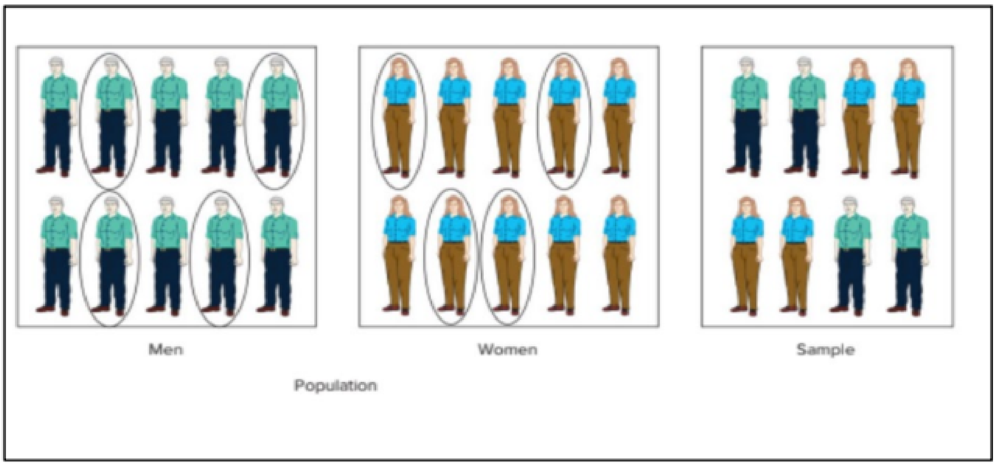


Figure 1.6: Source: Allan G. Bluman (2010)

1.9.3 Stratified Random Sample

- Population is divided into nonoverlapping subpopulations called strata.
- A random sample is selected from each stratum.
- Potential for reducing sampling error.
- Proportionate – the percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the population.
- Disproportionate – proportions of the strata within the sample are different than the proportions of the strata within the population.

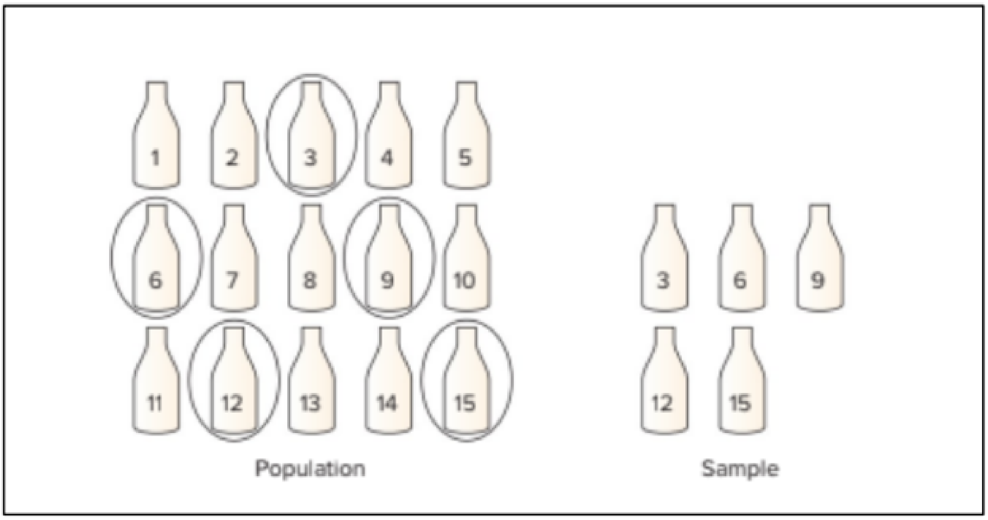


Figure 1.7: Source: Allan G. Bluman (2010)

1.9.4 Cluster Sampling

- Population is divided into non overlapping clusters or areas
- Each cluster is a miniature, or microcosm, of the population.
- A subset of the clusters is selected randomly for the sample.
- If the number of elements in the subset of clusters is larger than the desired value of n, these clusters may be subdivided to form a new set of clusters and subjected

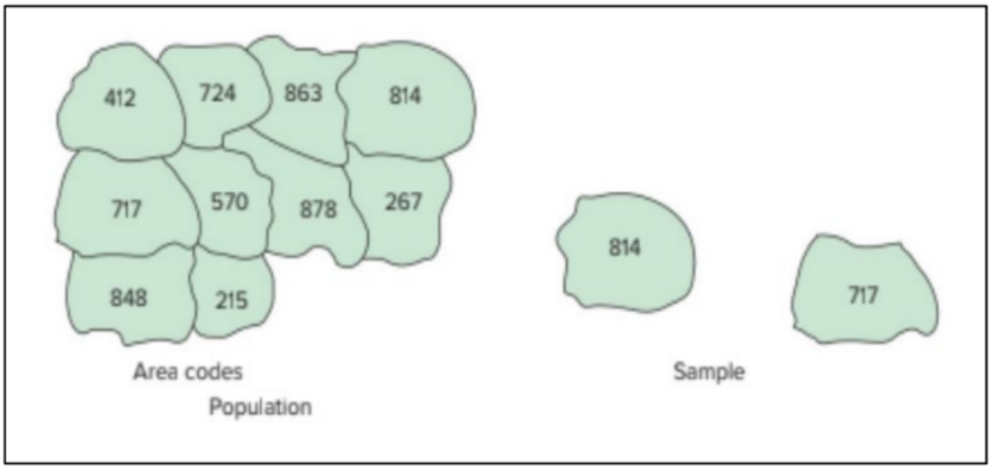


Figure 1.8: Source: Allan G. Bluman (2010)

Advantages

- More convenient for geographically dispersed populations.
- Reduced travel costs to contact sample elements.
- Simplified administration of the survey.
- Unavailability of sampling frame prohibits using other random sampling methods.

Disadvantages

- Statistically less efficient when the cluster elements are similar.
- Costs and problems of statistical analysis are greater than for simple random sampling to a random selection process.

1.9.5 Nonrandom Sampling

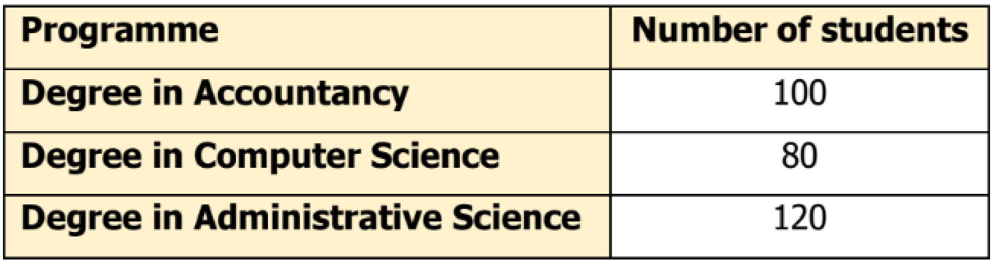
1. **Convenience Sampling**: sample elements are selected for the convenience of the researcher.
2. **Judgement Sampling**: sample elements are selected by the judgment of the researcher.
3. **Quota Sampling**: sample elements are selected until the quota controls are satisfied.
4. **Snowball Sampling**: survey subjects are selected based on a referral from other survey respondents

1.10 Exercise 1

1. A private company manager is interested in studying the relationship between time spent on social media and employee performance. He believed that the more time spent on social media, the more likely the performance drops. The performance of the employee is categorized as excellent, moderate, and low. A random sample of 500 employees was selected for this study, and the time spent on social media was recorded.

a) State the population and the sample for the above study.
b) Identify whether the study was conducted by a census or sample survey. Give a reason.
c) Identify the variable of this study.

2. A lecturer from a private college wanted to estimate how much their students spent (in RM) on reference books for a semester. From a total of 300 students, only 12 students randomly selected as a sample. The distribution of the number of students is shown in the following table.



Programme	Number of students
Degree in Accountancy	100
Degree in Computer Science	80
Degree in Administrative Science	120

a) State the population and sample.
b) State the variable of the study.

3. A registrar in University F would like to study the readiness of students returning to campus after one year staying at home learning through online platform. For this purpose, students are listed according to their student ID. An online questionnaire was distributed randomly to 1000 students. The readiness is scale from 1 (Strongly not ready) to 10 (Strongly ready).

a) State the population, sample and sampling frame.

b) Identify the variable of interest.

4. Determine whether **Descriptive Statistics** or **Inferential Statistics** were used.

a) Based on the bar chart, the highest sales for Company XYZ are on December.
b) There is an association between gender and level of stress.
c) A scientist found that a good laugh significantly reduces person's stress level.
d) The distribution number of male patients received a treatment in Hospital Y is skewed to the right.
e) Sales for Company Y are more consistent than Company X.
f) Based on a sample of 300 students, the dean has enough evidence to conclude that students more prefer face-to-face classes compared to online classes.
g) A study conducted by a research network found that people with fewer than 12 years of education had lower life expectancy than those with more years of education.
h) In 2025, the Save Hypermarket is predicted their sales to be RM 5.3 billion.

5. Decide whether it is primary data or secondary data.

a) Personal Interview.
b) Data from Statistical Department.
c) A study is undertaken to find the labour productivity of Hypermarket Berjaya at two locations. For this purpose, selected laborers in both locations is contacted and their productivity figures are noted.
d) Data obtained from the Kaggle website is being used by students for research courses.
e) Data collected by the Ministry of Health are being used by the scientist, academicians and etc for their further study.

6. State whether the following quantitative (discrete or continuous) or qualitative variable.

a) Distance
b) Litres of petrol
c) Level of anxiety
d) Depression score

e) Sizes of drinks sold by restaurant (small, medium and large)
f) Number of patients waiting for treatment
g) Length of time
h) Temperature at a Hawaii Resort
i) Rating of lecturer

7. State the level of measurement for the following variable.

a) Size of blouse (36,38,40,42)
b) The height of building
c) The number accident case .
d) Stress level (Mild, Medium, Severe, Very Severe)
e) Satisfaction level (1(very unsatisfied) to 5(very satisfied)

8. What types of sampling technique used for each situation?

a) A lecturer conducts a study to collect data on students' performance in the Statistics course. The lecturer randomly selects five classes from ten and samples all students in those five classes.
b) Newton Car operates ten dealerships in ten states, including Sabah and Sarawak. The Head of Service Department is interested to know about the car problems encountered by their customers. Only 150 customers are chosen as respondents for each branch.
c) The manager of Hotel Hibiscus instructed an HR Department to ask about customer satisfaction after a stay at their hotel. The customers are selected randomly.
d) The number of students passed in probability subject decreasing every semester. A lecturer is concerned with the issue and wants to know the problem and identify the issue. From 50 students, only 10 students with weak academic records are chosen.
e) In order to choose a suitable platform for an online class, students were asking for their opinion.

9. State the best sampling method that can be used for the study below:

a) A group of researchers conducted a study on the satisfaction of parents with online learning classes for primary school in Kedah. For that purpose, the researchers set out to conduct a survey and limit a selection of sample, 100 parents from rural areas and 200 parents from urban area.

b) The audit team selects every 20th box out of a total of 100 boxes to test the quality of Product X. All of the product X in the selected box is sampled in order to assess the product's defects.

1.11 Turorial 1

Question 1

A group of researchers wanted to investigate the perception of the married couple towards on the factors that contributed to the marriage problems in Perak. The researchers distributed the questionnaires to 500 randomly selected couples from four districts (I, II, III, IV) in the state. Twenty-five items on the perception towards on the factors contributed to the marriage problems were measured using Likert Scale (strongly agree=1, agree=2, neutral=3, disagree=4, strongly disagree=5). The distribution on the number of married couples is shownin the following table.

District	Number of married couples
I	500
II	450
III	300
IV	350

a) State the population of the above study.
b) State the sampling frame for the above study.
c) State the variable involve for this study. Hence, identify the corresponding scales of measurements.
d) Name the sampling method used in this study. Explain your answer in the context of the study.

Question 2

In the automobile industry, customer service is a crucial factor affecting car sales. The management of a reputed automobile company is interested in determining the level of customers' satisfaction with the service provided by the company's service centres. The company has altogether 40 service centres throughout Malaysia. A sample of eight centres was selected at random. Questionnaires are disseminated to all customers who service their cars at these eight selected services centres on one selected day (the day of the survey). One of the questions asked is satisfaction level on the services provided (using rating: good, fair, poor)

a) State the population of the study.
b) Name the variable of interest for the above study. State its type and level of measurement.
c) Identify the sampling technique used. Explain briefly how the sample is selected.

Question 3

A researcher wishes to study students' satisfaction (strongly disagree=l, disagree=2, neutral=3, agree=4, strongly agree=5) towards the services provided by the Academic Affairs in Nursing College X. The researcher chooses only 10 out of 50 classes. All the students from these 10 classes will be used for the study.

a) Identify the population for the above study.
b) State the sampling frame for the above study.
c) Name the variable of interest for the above study. State its type and its level of measurement.
d) State the sampling technique used in this study.

Question 4

A researcher is interested in studying the career aspirations of students from the Faculty of Electrical Engineering, which consists of 30 classes. The researcher intends to choose all the students from 5 classes for the study.

a) State the population and the sample for the above study.
b) Identify the variable of interest for this study and state the type of variable used.
c) What is the sampling technique used in this study?
d) Name ONE (1) method of data collection suitable for this study? Give ONE (1) advantage of using this method.

Question 5

A survey on the workers' satisfaction levels was carried out at Company XY. The company has 24 branches with the same setting. A sample of 6 branches was selected at random. All workers who work at these 6 branches were then selected for the study.

a) State the population of the study.
b) State the sampling frame for the survey.
c) State the variable for this study. What type of variable is it?
d) Name the sampling technique used in the study.
e) Besides the sampling technique used in (d), briefly explain how the sample of 6 branches can be selected using systematic sampling technique.
f) What is the most suitable data collection method to be used for the study? Give one advantage of the suggested method.

Question 6

Employers were surveyed to determine the level of satisfaction with their employees who are graduated from ICT courses at University YY. This study involved 50 employers from ICT private companies from five randomly chosen states out of 14. All selected employers were asked about gender, length of service (years), how well graduates meet employer expectations (1=Excellent, 2=Average, and 3=Poor), and overall employer's satisfaction with graduates (1=Excellent, 2=Average, and 3=Poor).

a) State the population and sampling frame.
b) Name any TWO (2) variables from the study. Hence, state its type of variable.
c) Name the sampling technique employed in the study.
d) Suggest the data collection method that suitable to the study. Give ONE (1) advantage of the method used.

Question 7

A manager at one of the popular Telco company is currently conducting a survey regarding the service failure at their service counter. The main objective of the survey is to find out the factors that cause the failure. He randomly selected five service counters from ten available service counters all over

Malaysia. A questionnaire is distributed to all the customers at the five selected service counters. The information collected from the customers include age, gender, occupation, income, rating of service (0 to 100) and service quality (poor, moderate and good).

a) State the population in the study.

b) State the sampling technique used in the study.

c) Identify one ordinal variable and one ratio variable obtained from the study.

d) The followings are the statistics produced from the study. Identify whether each statement is a descriptive or inferential statistics.

   i) 45% of the sample customers work in the government sector.
   ii) Based on the sample, it can be concluded that there is an association between gender and service quality.
   iii) We are 90% confident that the average rating of service of for the customers falls between 60 and 90.

1.12 Answer to Tutorial 1

Question 1

a) All married couples in Perak.
b) List of all married couples from four districts in Perak.
c) Perception; Nominal.
d) Stratified Sampling Technique. Determine the number of samples from each districts :

District	Number of married couples	Number of samples
I	500	(500/1600)*500=156
II	450	141
III	300	94
IV	350	109
		500

Question 2

a) All customers at all 40 services centre in Malaysia.
b) Customer satisfaction; qualitative; ordinal
c) Cluster sampling technique.

Question 3

a) All students at Nursing College X.
b) List of all students from ten classes.
c) Students' satisfaction; qualitative ; ordinal
d) Cluster sampling technique.

Question 4

a) Population : All students from Faculty of Electrical Engineering. Sample: All students from five classes.
b) Career aspiration; qualitative; nominal
c) Cluster sampling technique
d) Self-administrative questionnaire. Advantage: quick response

Question 5

a) All workers at Company XY.
b) List of all workers at six branches at Company XY.
c) Satisfaction; qualitative
d) Cluster sampling technique.
e) N=24, n =6; k= 24/6 = 4; in first interval choose any number from 1 until 4. let say, choose number 2. The number 2 will be the first sample. Next interval from 5 until 8. Continue select the samples until you reach 6 samples.
f) Email questionnaires.

Question 6

e) Population: All employers in the ICT private companies who are employed ICT graduates from University YY. Sampling frame: A list name of ICT company who are employed ICT graduated from University YY.
f) Gender (Qualitative); length of services (Quantitative Continuous), employer's expectations (Qualitative), and employer's satisfaction with graduates (Qualitative).
g) Cluster sampling

h) Internet survey (Google form)/ Electronic questionnaire. Fast and short in time span to complete the questionnaire, cheaper. Any relevant answer.

Question 7

a) Population: All customers from 10 service counters all over Malayisa

b) Cluster sampling

c) Ordinal variable : service quality; ratio variable : age/ income

d) i) Descriptive, ii) Inferential iii) Inferential

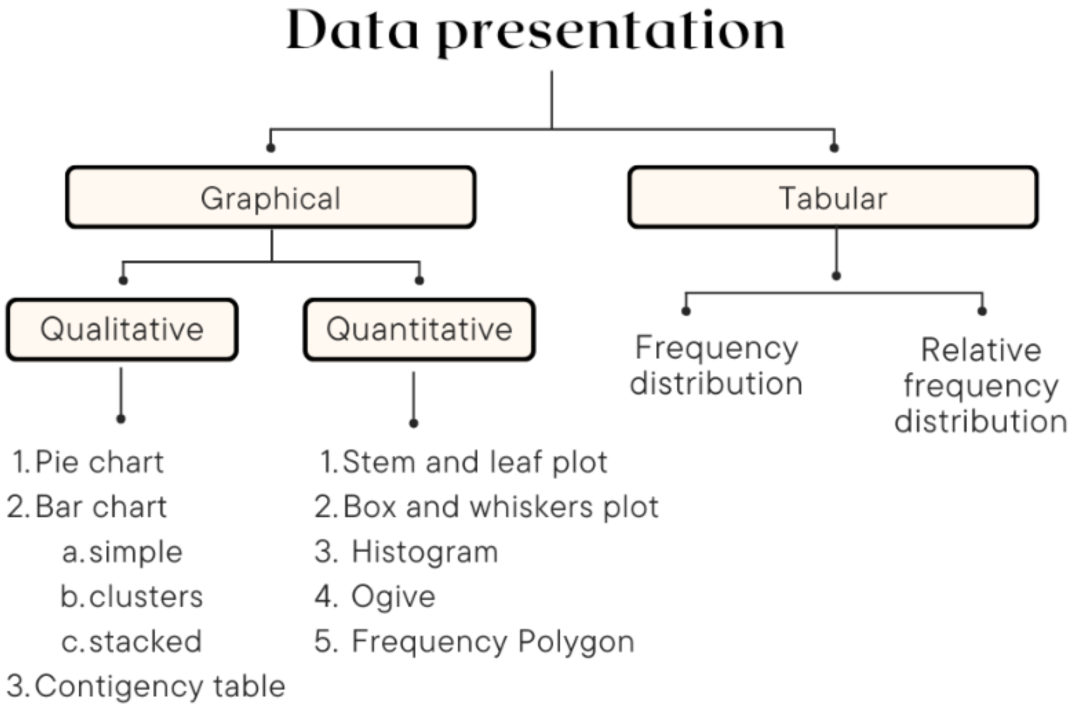
2 Descriptive Statistics

Learning objectives:

1. Identify various ways to present collected data from survey of secondary sources
2. Use appropriate data presentation for a qualitative and quantitative data
3. Calculate the measures of central tendency, measure of variation, measure of skewness and measure of position for ungrouped data.

2.1 Introduction

When conducting a statistical study, the researcher must gather data for the variable under study. For example, if a researcher wishes to study the number of road accidents in Malaysia for the past 2 years, he or she must gather the data from various departments. To describe the situation, draw conclusions, or make inferences about the event, the researcher must organize the data and present it in some meaningful way.



The data can be presented in term of graphical approach or table and numerical descriptive measure for both qualitative dan quantitative data.

2.2 Organizing Data

Charts and graphs

- Frequency distributions are good ways to present the essential aspects of data collections in concise and understandable terms
- Pictures are always more effective in displaying large data collections

Histogram

- Frequently used to graphically present interval and ratio data
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

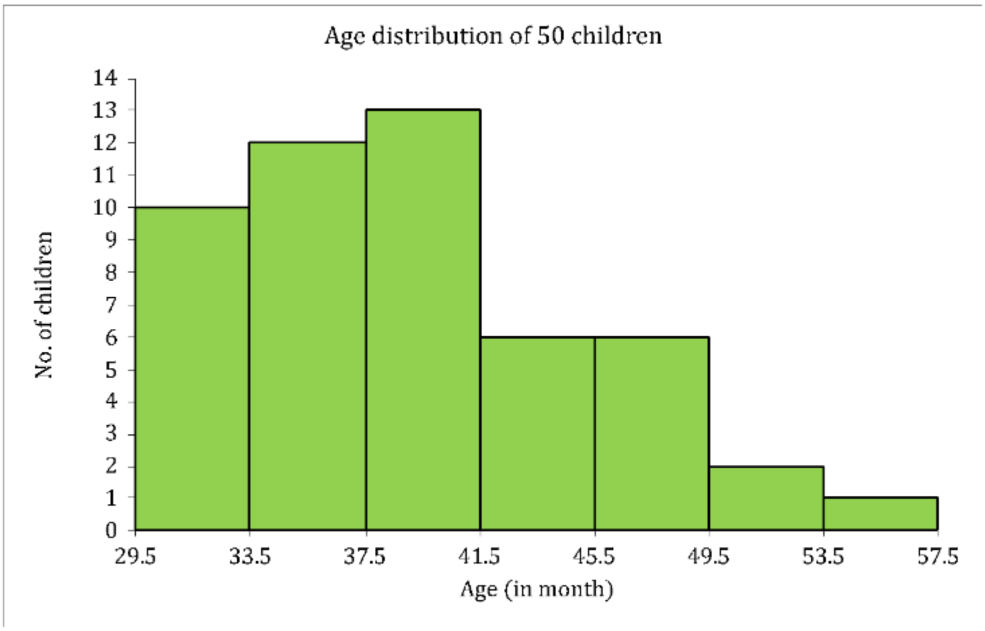


Figure 2.1: Figure 1: Histogram – distribution of the ages (in months) of 50 children

Frequency Polygon

- Another common method for graphically presenting interval and ratio data
- To construct a frequency polygon mark the frequencies on the vertical axis and the values of the variable being measured on the horizontal axis, as with the histogram.
- If the purpose of presenting is to compare with other distributions, the frequency polygon provides a good summary of the data

Ogive

- A graph of a cumulative frequency distribution
- Ogive is used when one wants to determine how many observations lie above or below a certain value in a distribution.
- First cumulative frequency distribution is constructed
- Cumulative frequencies are plotted at the upper-class limit of each category
- Ogive can also be constructed for a relative frequency distribution.

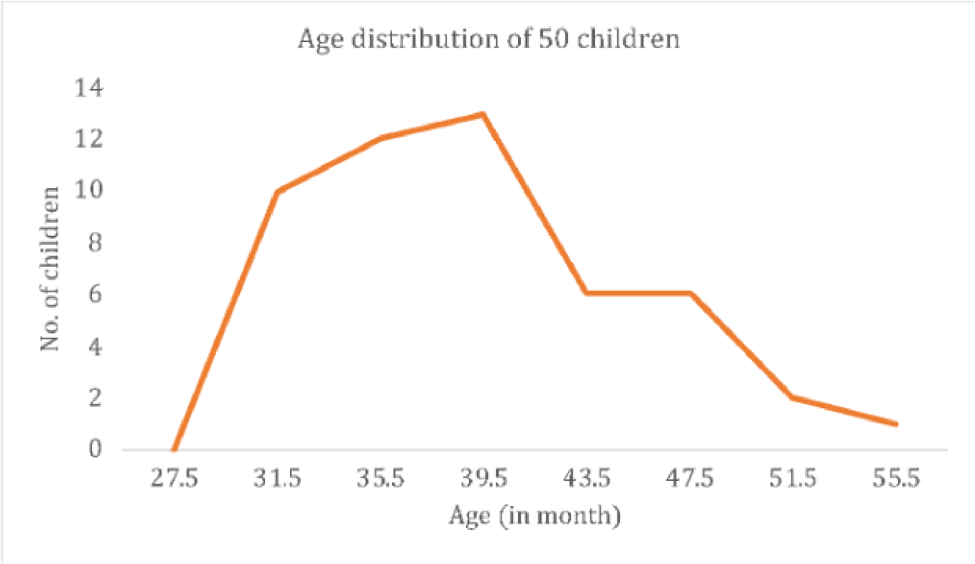


Figure 2.2: Figure 2: Frequency Polygon – distribution of the ages (in months) of 50 children

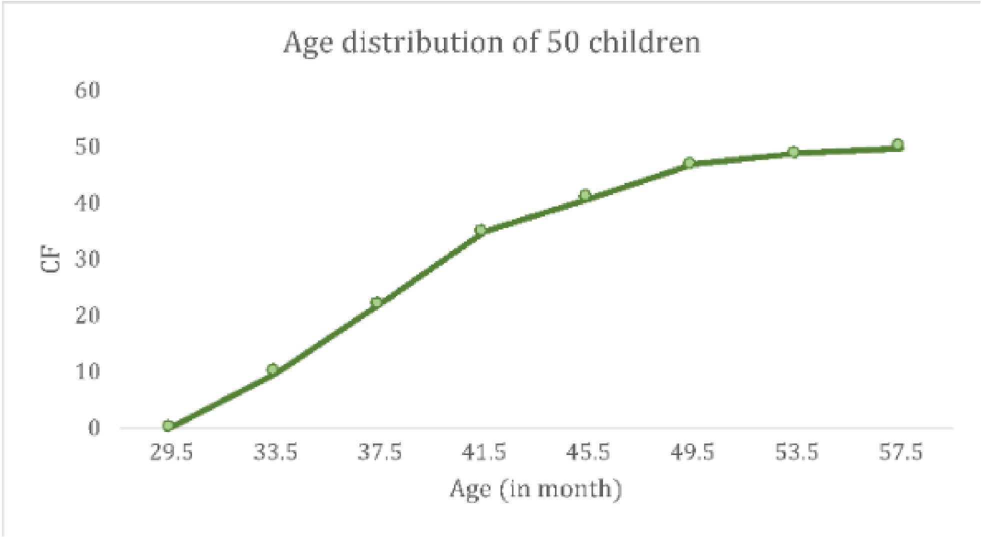
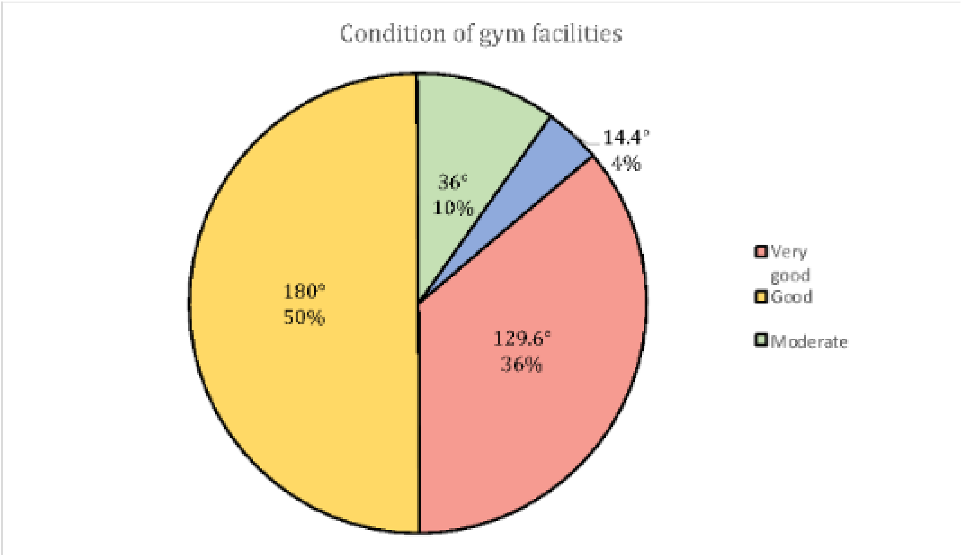


Figure 2.3: Figure 3: Ogive distribution of the ages (in months) of 50 children

Pie Chart

- The pie chart is an effective way of displaying the percentage breakdown of data by category.
- The size of angle (°) is determined based on the frequency of group/category
- Useful if the relative sizes of the data components are to be emphasized Pie charts also provide an effective way of presenting ratio- or interval-scaled data after they have been organized into categories
- How to compute angle/degree?

$$x^\circ = \frac{f}{\sum x} \times 360$$



Bar Charts

- Another common method for graphically presenting nominal and ordinal scaled data
- One bar is used to represent the frequency for each category
- The bars are usually positioned vertically with their bases located on the horizontal axis of the graph
- The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data – the separation emphasize the plotting of frequencies for distinct categories

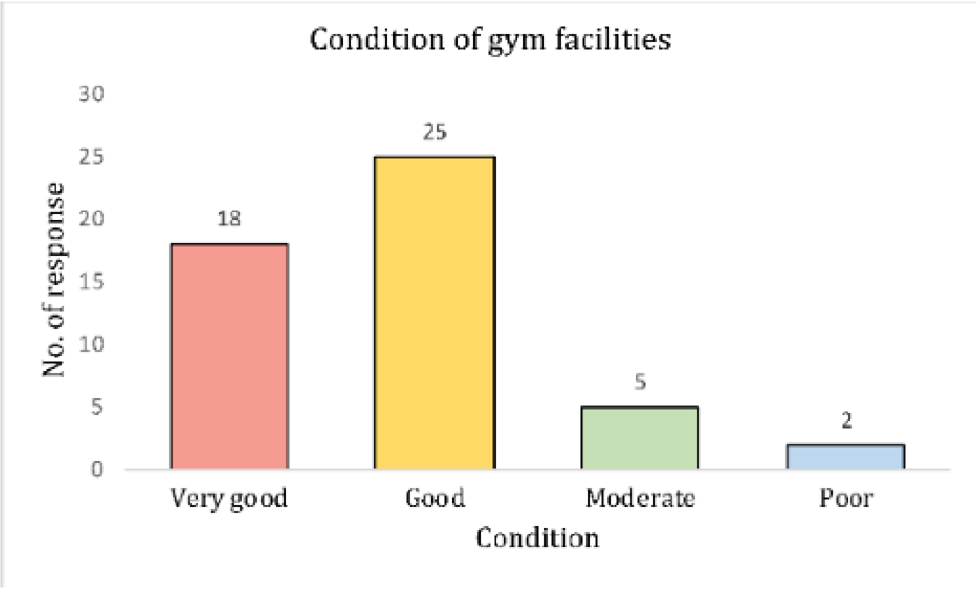
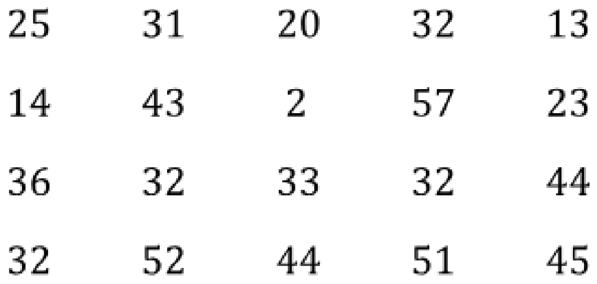


Figure 2.4: Figure 5: Bar Chart – Condition of gym facilities

Stem and Leaf

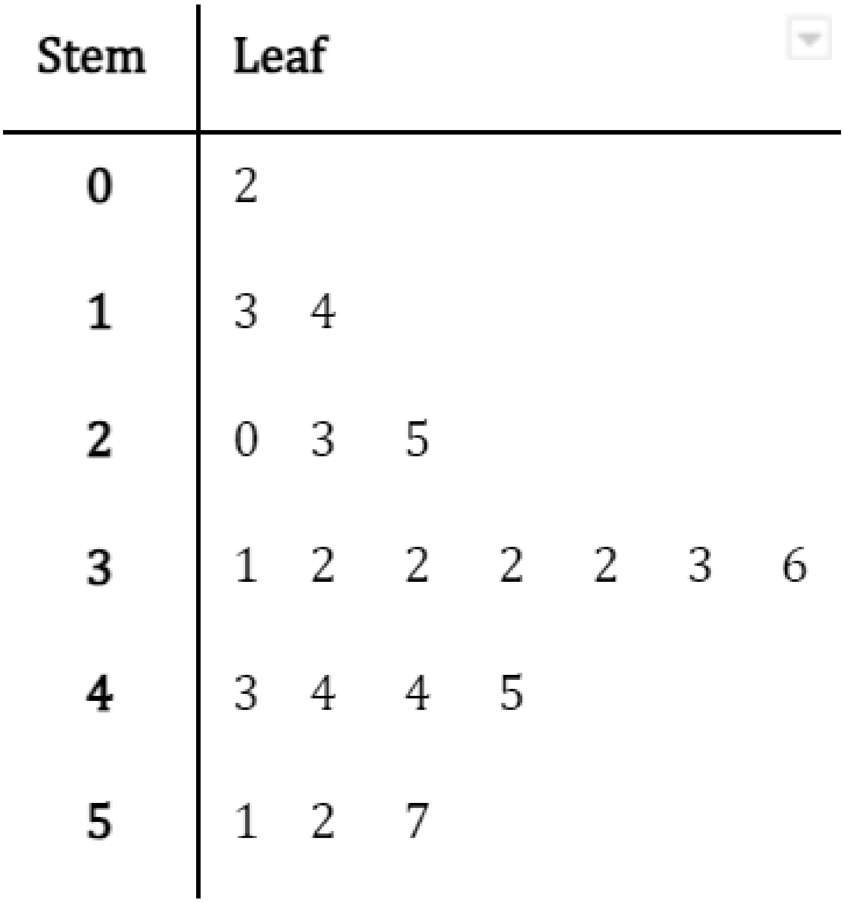
A *stem and leaf plot* is a data plot that uses part of a data value as the stem and part of the data value as the leaf to form groups or classes.



25	31	20	32	13
14	43	2	57	23
36	32	33	32	44
32	52	44	51	45

2.3 Numerical descriptive measures (ungrouped)

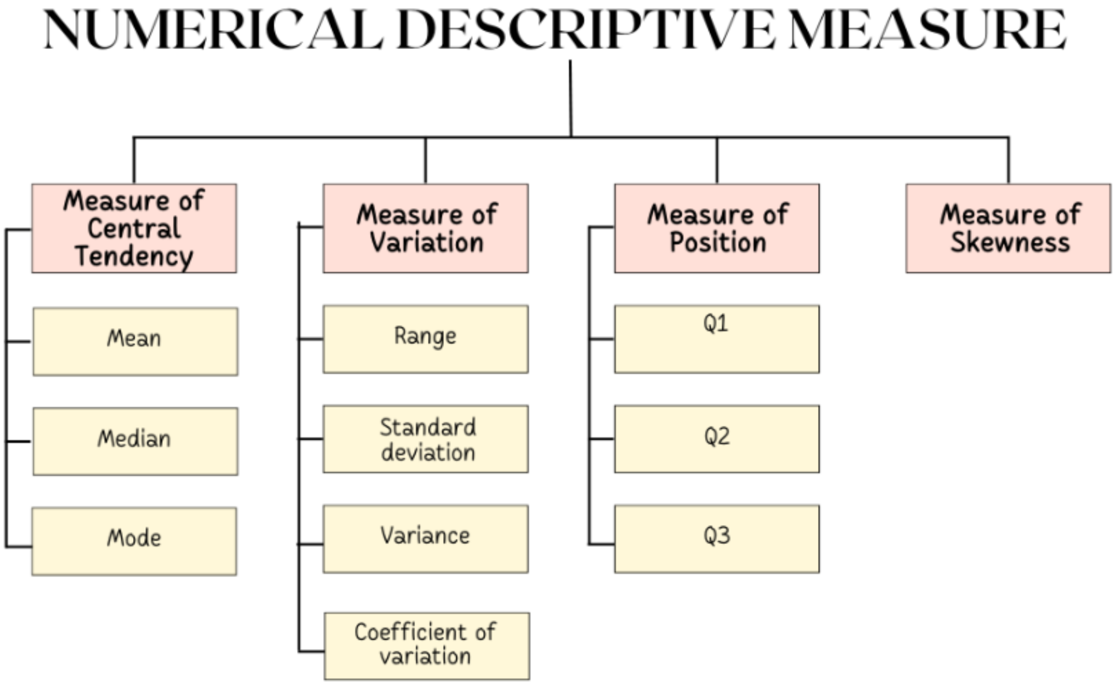
- Measures of central tendency (mean, median, mode)



Stem	Leaf						
0	2						
1	3	4					
2	0	3	5				
3	1	2	2	2	2	3	6
4	3	4	4	5			
5	1	2	7				

Figure 2.5: Figure 6: Stem-and-Leaf Diagram

- Measure of variation (range, standard deviation, variance, coefficient of variation
- Measure of skewness
- Measures of Position (Q1, Q2, and Q3)



2.3.1 Measures of Central Tendency

Mean

The mean is the quotient of the sum of the values and the total number of values. The symbol $\bar{x}$ is used for **sample mean** and is given by the following formula:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_{n-1} + x_n}{n} = \frac{\sum x}{n}$$

For a population, the Greek letter μ (mu) is used for the mean.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + ... + X_{n-1} + X_N}{N} = \frac{\sum X}{N}$$

Example 2.1

The number of calls that a local police department responded to for a sample of 9 months is shown. Find the mean.

475, 447, 440, 761, 993, 1052, 783, 671, 621

Answer

$$\begin{aligned}\bar{x} = \frac{\sum x}{n} &= \frac{475 + 447 + 440 + 761 + 993 + 1052 + 783 + 671 + 621}{9} \\ &= \frac{6243}{9} \approx 693.7\end{aligned}$$

Interpretation: On average, the number of calls responded by local police department is approximately 694 calls.

Example 2.2

A climatologist recorded daily temperature in Sungai Petani for twelve days. The data points are as follows (in degrees Celsius):

30.5, 31.2, 29.8, 30.0, 30.7, 31.5, 29.9, 30.3, 30.1, 31.0, 30.4, 31.8

Calculate the mean of daily temperature, and interpretation of the meaning of the mean value.

Answer

$$\begin{aligned}\bar{x} &= \frac{30.5 + 31.2 + 29.8 + 30.0 + 30.7 + 31.5 + 29.9 + 30.3 + 30.1 + 31.0 + 30.4 + 31.8}{12} \\ &= 30.75^{\circ}C\end{aligned}$$

On average, the daily temperatures in Sungai Petani for twelve days is $30.75^{\circ}C$.

Example 2.3

In a bakery, the number of doughnuts sold in 7 days is recorded as follows:

53, 60, 45, 77, 58, 42, 68

Answer

$$\begin{aligned}\bar{x} &= \frac{53+60+45+77+58+42+68}{7}\\ &= 57.57\end{aligned}$$

On average, the number of doughnuts sold in seven days is approximately 58 doughnuts.

Properties of the Mean

- Found by using all the values of data.
- Varies less than the median or mode.
- Used in computing other statistics, such as the variance.
- Unique, usually not one of the data values.
- Cannot be used with open-ended classes.
- Affected by extremely high or low values, called outliers.

Median

The **median** is the midpoint of the data array. The symbol for the median is $\tilde{x}$. How to identify median value?

i. Sort the data in ascending order
ii. Identify the location of median using
iii. Determine the value of median

- If the number of data (n) is ODD, the value is in the middle of sequence
- If the number of data (n) is EVEN, the value of median is average of 2 middle values

Example 2.4

The number of police officers killed in the line of duty over the last 11 years is shown. Find the median.

175 152 121 142 188 154 160 165 148 156 239

Answer:

i. Sort the data in ascending order

121, 142, 148, 152, 154, 156, 160, 165, 175, 188, 239

ii Identify the location of median:

$$\frac{11+1}{2} = 6^{th}$$

iii. Select the middle value, $\tilde{x} = 156$

Example 2.5

The number of tornadoes that have occurred in the certain country over an 8-year period follows.

684, 764, 656, 702, 856, 1133, 1132, 1303

Find the median.

Answer:

Since the data given is even-number, there will be two values in the middle.

i. Sort the data in ascending order

656, 684, 702, 764, 856, 1132, 1133, 1303

ii Identify the location of median:

$$\frac{8+1}{2} = 4.5^{th}$$

iii. Select the middle value, $\tilde{x} = \frac{764+856}{2} = \frac{1620}{2} = 810$

Interpretation of median: Half of the tornadoes that have occurred in a certain country is less than or equal to 810 of tornadoes.

Properties of Median

- Give the midpoint.
- Used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- Can be used for an open-ended distribution.
- Affected less than the mean by extremely high or extremely low values.

Mode

The **mode** is the value that occurs most often in a data set. It is sometimes said to be the most typical case. There may be no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal). Symbol of the mode is $\hat{x}$.

Example 2.6

Find the mode of the signing bonuses of eight football players for a specific year. The bonuses in thousands of ringgit Malaysia (RM) are:

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

Answer:

You may find it easier to sort first.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5

Select the value that occurs the most.

Interpretation: Most of the players signed the bonuses of 10 thousand ringgits Malaysia.

Example 2.7

The data show the number of licensed nuclear reactors in the certain country for a recent 15-year period. Find the mode.

103 103 103 103 103 106 108 108 108 110 108 112 112 112 108

Answer:

103 and 108 both occur the most. The data set is said to be bimodal. The modals are 103 and 108.

Properties of mode

- Used when the most typical case is desired.
- Easiest average to compute.
- Can be used with nominal data.
- Not always unique or may not exist.

Distributions Shape

2.3.2 Measure of variation

Measures of variation include range, standard deviation, variance, and coefficient of variation.

The **range** is the difference between the highest and lowest values in a data set.

Range = Highest – Lowest

The **population variance** is the average of the squares of the distance each value is from the mean.

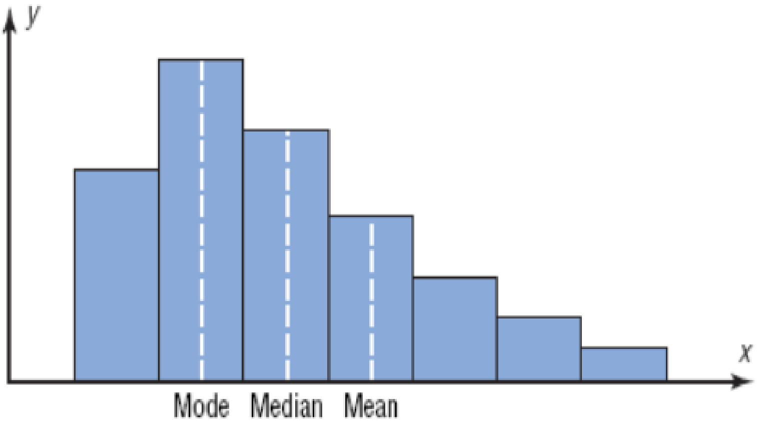


Figure 2.6: Figure 7: a) Positively skewed or right-skewed

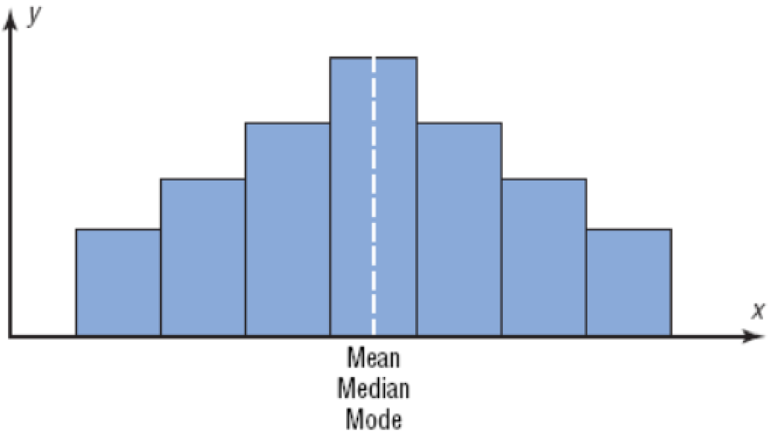


Figure 2.7: Figure 7: b) Symmetrical distribution since mean, mode and median are equal

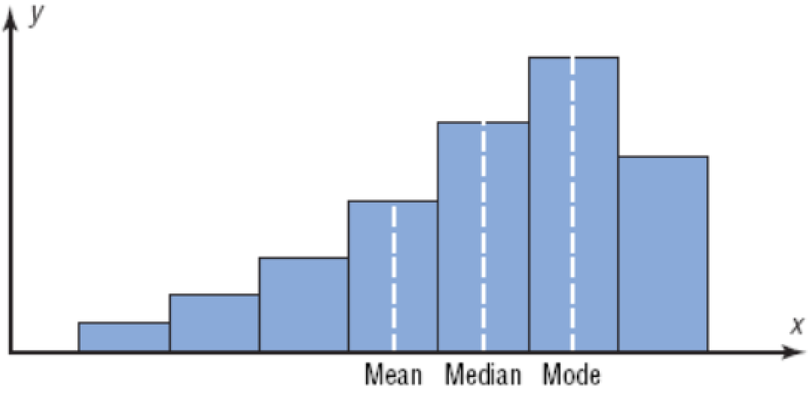


Figure 2.8: Figure 7: c) Negatively skewed or left-skewed

The **standard deviation** is the square root of the variance.

Example 2.8

The table below shows the life-long (in months) of two brands of paint.

Brand XX	Brand YY
10	35
60	45
50	30
30	35
40	40
20	25

Determine the mean and standard deviation for the above data.

Answer:

Brand XX: $\mu = mean = \frac{\sum x}{N} = \frac{210}{6} = 35$ and $Range = 60 - 10 = 50$.

Brand YY: $\mu = mean = \frac{\sum x}{N} = \frac{210}{6} = 35$ and $Range = 45 - 25 = 20$.

The average for both brands is the same, but the range for Brand XX is much greater than the range for Brand YY. Which brand would you buy?

Uses of the Variance and Standard Deviation

- To determine the spread of the data.
- To determine the consistency of a variable.
- To determine the number of data values that fall within a specified interval in a distribution (Chebyshev's Theorem).
- Used in inferential statistics.

Population variance:

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$$

Sample variance:

$$s^2 = \frac{\sum (X-\bar{X})^2}{n-1} \quad or \quad s^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]$$

Sample Standard Deviation:

$$s = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}} \quad or \quad s = \sqrt{\frac{1}{n-1}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]}$$

The **coefficient of variation** is the standard deviation divided by the mean, expressed as a percentage.

$$Coefficient\ of\ Variation\ (CV)\ = \frac{s}{x} \times 100\%$$

Example 2.9

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is \$5225, and the standard deviation is \$773. Compare the variations of the two.

Answer:

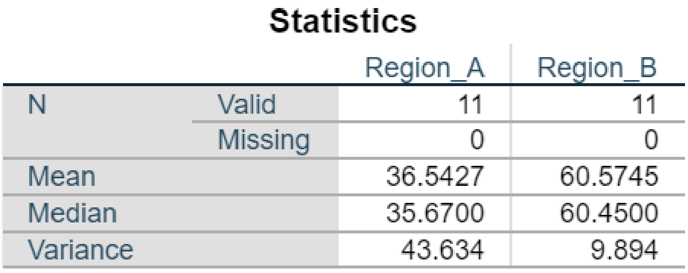
$CV_{Sales} = \frac{5}{87} \times 100\% = 5.7\%$

$CV_{Commissions} = \frac{773}{5225} \times 100\% = 14.8\%$

The data distribution for Commissions are more dispersed than sales. Similarly, Sales data distribution is more consistent than the data distribution in commissions.

Example 2.10

A meteorologist is studying the monthly rainfall patterns in two different regions, Region A and Region B, over the past year. They have collected data for 11 months (in mm). The goal is to determine which region has more dispersed rainfall. Calculate the coefficient of variation for each region's monthly rainfall data and identify which region's data is more dispersed.



Statistics

		Region_A	Region_B
N	Valid	11	11
	Missing	0	0
Mean		36.5427	60.5745
Median		35.6700	60.4500
Variance		43.634	9.894

Figure 2.9: SPSS Output

Answer:

$CV_{RegionA} = \frac{\sqrt{43.634}}{36.5427} \times 100\% = 18.08\%$

$CV_{RegionB} = \frac{\sqrt{9.894}}{60.5745} \times 100\% = 5.19\%$

Region A has a more dispersed rainfall pattern than Region B.

2.3.3 Measure of skewness

Skewness is the measurement of the lack of symmetry of the distribution. There are several measures for expressing the amount of skewness, but the only important one is the Pearson Coefficient of Skewness:

$$PCS = \frac{mean - mode}{standard\ deviation} \quad or \quad PCS = \frac{3(mean - median)}{standard\ deviation}$$

Skewness

It is the *degree of distortion* from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution. It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.

There are two types of Skewness: **Positive** and **Negative**.

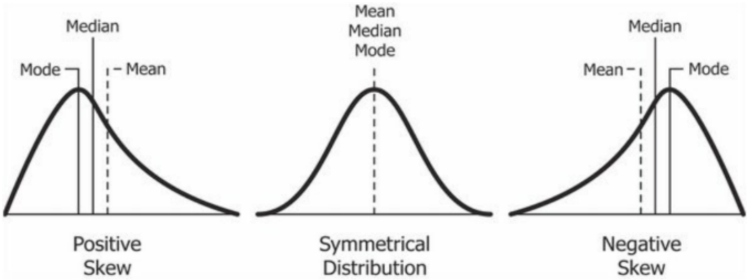
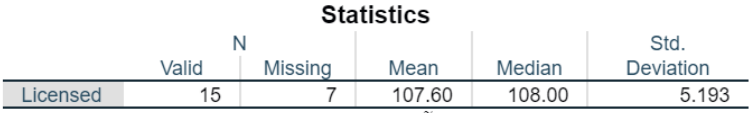


Figure 2.10: Figure 8: Types of Skewness

Example 2.11

The SPSS output gives information on descriptive statistics on the number of licensed nuclear reactors in the certain country for a recent 15-year period. Find the Person Coefficient of Skewness. Hence identify the shape of distribution.

Answer:



Statistics

	N		Mean	Median	Std. Deviation
	Valid	Missing			
Licensed	15	7	107.60	108.00	5.193

Figure 2.11: SPSS Output

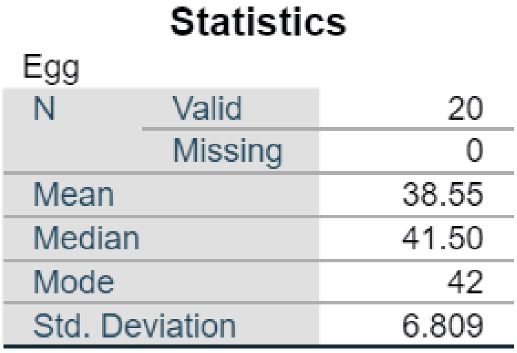
$$\begin{aligned} PCS &= \frac{3(\bar{x} - \tilde{x})}{s} \\ &= \frac{3(107.6 - 108)}{5.193} \\ &= -0.2311 \end{aligned}$$

Negatively skewed.

Example 2.12

The table below shows the number of eggs laid by 20 chickens over a 30-day period in a small poultry farm. Calculate the Pearson Coefficient of Skewness. Hence identify the shape of distribution.

Answer:



Statistics

Egg

N	Valid	20
	Missing	0
Mean		38.55
Median		41.50
Mode		42
Std. Deviation		6.809

Figure 2.12: SPSS Output

$$\begin{aligned} PCS &= \frac{\bar{x} - \hat{x}}{s} \\ &= \frac{38.55 - 42}{6.809} \\ &= -0.5067 \end{aligned}$$

Negatively skewed.

2.3.4 Measures of Position (Q, Q2, and Q3)

Quartiles separate the data set into 4 equal groups. $Q_1 = P_{25}$, $Q_2 = MD$, $Q_3 = P_{75}$

The **Interquartile Range**, $IQR = Q_3$–Q_1.

Example 2.13

Find Q_1, Q_2, and Q_3 for the data set.

15, 13, 6, 5, 12, 50, 22, 18

Answer:

Sort in ascending order.

5, 6, 12, 13, 15, 18, 22, 50

Location: $Q_1 = \frac{n+1}{4} \rightarrow Q_1 = \frac{8+1}{4} = 2.25^{th}$ in the array.

Value of Q1: $X_{Q_1} = 6 + 0.25(12 - 6) = 7.5$

Location: $Q_2 = \frac{n+1}{2} \rightarrow Q_2 = \frac{8+1}{2} = 4.5^{th}$ in the array.

Value of Q2: $X_{Q_2} = 13 + 0.5(15 - 13) = 14$

Location: $Q_3 = 3(\frac{n+1}{4}) \rightarrow Q_3 = 3(\frac{8+1}{4}) = 6.75^{th}$ in the array.

Value of Q3: $X_{Q_3} = 18 + 0.75(22 - 18) = 21$

An **outlier** is an extremely high or low data value when compared with the rest of the data values. A data value less than Q1 – 1.5(IQR) or greater than Q3 + 1.5(IQR) can be considered an outlier.

2.3.5 Box-and-whisker plot

Also known as box plots, they display a **five-number summary** of a set of data. Consist of - (i) minimum, (ii) first quartile, (iii) median, (iv) third quartile, and (v) maximum values.

Able to identify the centre, the spread, and the skewness of the data

Example 2.14

The number of meteorites found in 10 U.S. states is shown. Construct a boxplot for the data. Hence, determine its skewness.

89, 47, 164, 296, 30, 215, 138, 78, 48, 39 30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Five-Number Summary: 30; 47; 83.5; 164; 296

Example 2.15

A meteorologist is studying the monthly rainfall patterns in two different regions, Region A and Region B, over the past year. They have collected data for 11 months (in mm). They determine the skewness of the two regions based on box and whiskers plot.

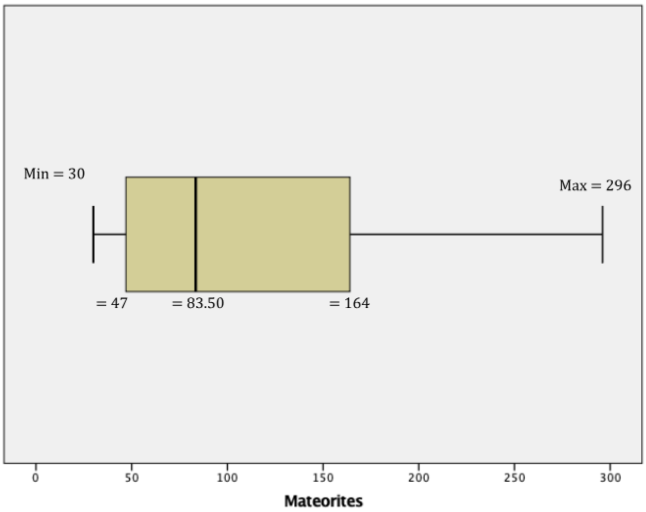
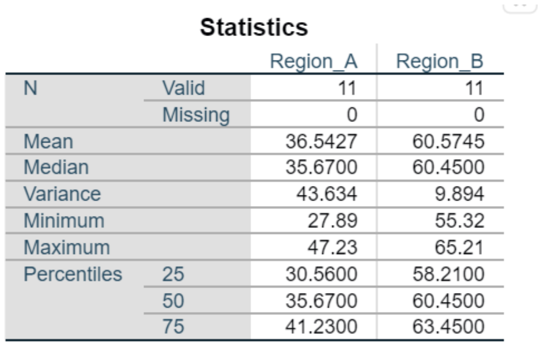


Figure 2.13: Figure 9: Skewed to the right / right skewness / positively skewed

Answer :



Statistics

		Region_A	Region_B
N	Valid	11	11
	Missing	0	0
Mean		36.5427	60.5745
Median		35.6700	60.4500
Variance		43.634	9.894
Minimum		27.89	55.32
Maximum		47.23	65.21
Percentiles	25	30.5600	58.2100
	50	35.6700	60.4500
	75	41.2300	63.4500

Figure 2.14: SPSS Output

2.4 SPSS Output on Numerical Descriptive Statistics

IBM statistical software provides a wide range of statistical analysis and graphing capabilities. We use dataset from SPSS's samples so that students can have free access to it and be able to run it by themselves to produce the SPSS output on descriptive statistics.

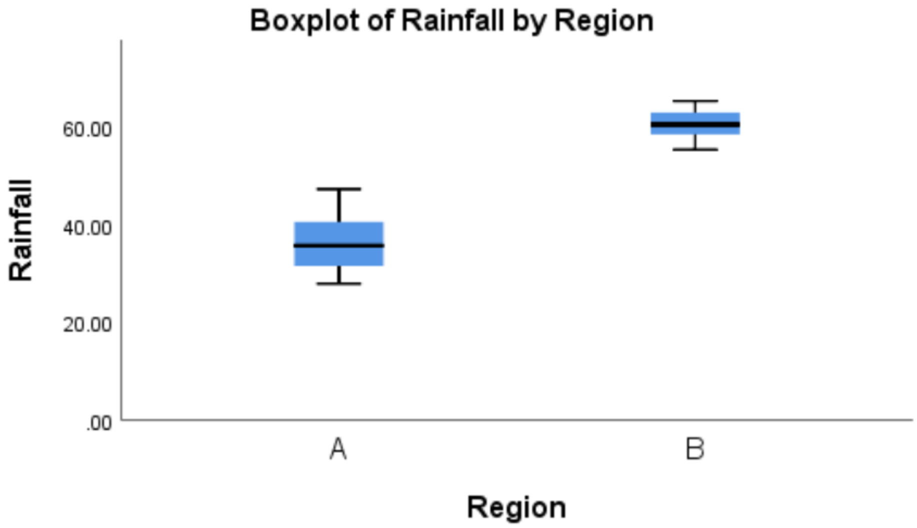
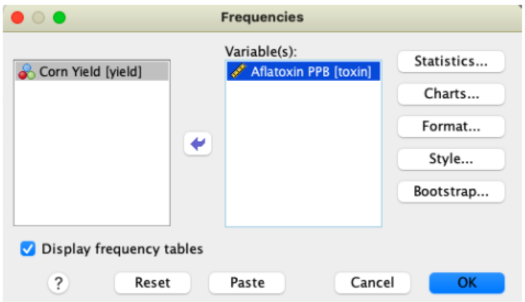


Figure 2.15: Figure 9: Region A – positively skewed; Region B – Approximately normal

Data: **aflatoxin20.sav** Description: This data consists of 32 poison concentrations which are measured in parts per billion (PPB) on corn crops. This poison is known as aflatoxin. The concentration varies widely within crop yields.

To open a data file: From the menus choose: File > Open > Data... A dialog box for opening files is displayed and choose **aflatoxin20.sav**

To obtain the descriptive analysis of the data:
1. From the menus choose:
Analyze > Descriptive Statistics > Frequencies...
The Frequencies dialog box is displayed.
2. Double click on ***Aflatoxin PPB [toxin]***

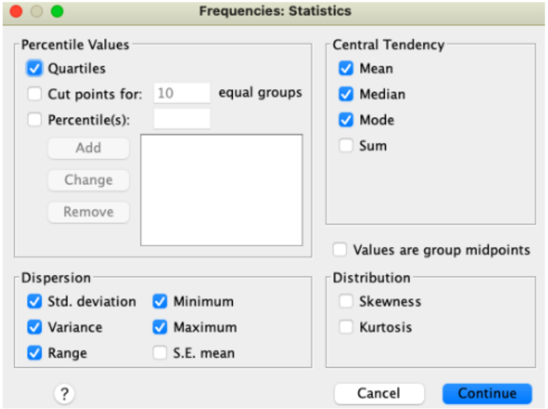


Corn Yield [yield]

Variable(s):
Aflatoxin PPB [toxin]

☑ Display frequency tables

3. Click on **Statistics** and choose as per figure below:



Frequencies: Statistics

Percentile Values
- ☑ Quartiles
- □ Cut points for: 10 equal groups
- □ Percentile(s):

Central Tendency
- ☑ Mean
- ☑ Median
- ☑ Mode
- □ Sum

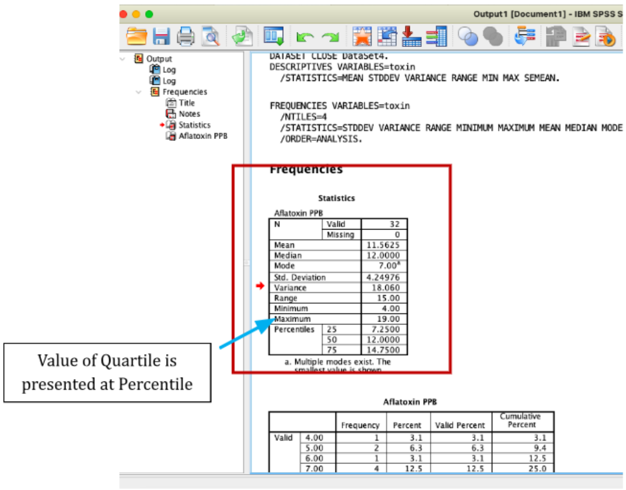
□ Values are group midpoints

Dispersion
- ☑ Std. deviation
- ☑ Variance
- ☑ Range
- ☑ Minimum
- ☑ Maximum
- □ S.E. mean

Distribution
- □ Skewness
- □ Kurtosis

4. Click **Continue**

5. Click **OK** to run the procedure. Results are displayed in the Viewer window.



```
DATASET CLOSE DataSet4.
DESCRIPTIVES VARIABLES=toxin
  /STATISTICS=MEAN STDDEV VARIANCE RANGE MIN MAX SEMEAN.

FREQUENCIES VARIABLES=toxin
  /NTILES=4
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM MEAN MEDIAN MODE
  /ORDER=ANALYSIS.
```

Frequencies

Statistics

Aflatoxin PPB

N	Valid	32
	Missing	0
Mean		11.5625
Median		12.0000
Mode		7.00[a]
Std. Deviation		4.24976
Variance		18.060
Range		15.00
Minimum		4.00
Maximum		19.00
Percentiles	25	7.2500
	50	12.0000
	75	14.7500

a. Multiple modes exist. The smallest value is shown.

Value of Quartile is presented at Percentile

Aflatoxin PPB

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4.00	1	3.1	3.1	3.1
	5.00	2	6.3	6.3	9.4
	6.00	1	3.1	3.1	12.5
	7.00	4	12.5	12.5	25.0

2.5 Exercise 2

1. The weight (kg) of 12 customers registered for WX fitness studio are as follows.
   110, 90, 65, 100, 95, 75, 90, 85, 80, 70, 75, 90

a. Find the mode and comment on the value obtained.
b. Calculate the mean and standard deviation.
c. Describe the shape of distribution by using an appropriate measurement.

2. A researcher studied petrol usage (in littles) for 14 cars travelling from town A to town B in a month. The data collected is shown below:
   81, 80, 65, 105, 144, 75, 150, 96, 91, 68, 135, 134, 95, 124

a. Find the first quartile, median, and third quartile.
b. Construct a box-and-whiskers plot for the above data.
c. Determine the shape of the data distribution based on the box-and-whiskers plot.

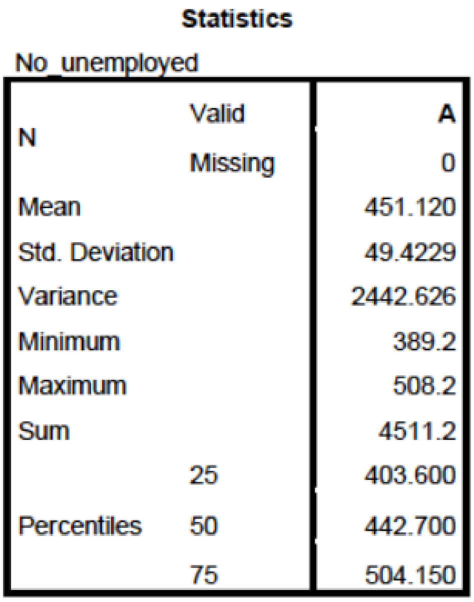
3. The number of current issues books sold in a week by 12 bookstores at town A is as follows.
   11, 23, 12, 15, 22, 10, 16, 15, 7, 12, 26, 14

a. Find the mean and the mode.
b. By comparing the mean and the mode in (a), determine the shape of the distribution for the above data.
c. Calculate the first quartile, median and third quartile of the above data.

4. The chemistry test marks of eight randomly selected students are given below:
   85, 65, 48, 70, 30, 80, 92, 70

a. Calculate the mean, median and mode for the test score.
b. Find the variance of the test score for the selected students.
c. Determine the shape of the distribution of the test score using Pearson's coefficient of skewness.

5. The output below refers to National unemployment rate prior to the coronavirus outbreak (Covid-19).



Statistics

No_unemployed

N	Valid	A
	Missing	0
Mean		451.120
Std. Deviation		49.4229
Variance		2442.626
Minimum		389.2
Maximum		508.2
Sum		4511.2
Percentiles	25	403.600
	50	442.700
	75	504.150

a. Determine the value of **A**.

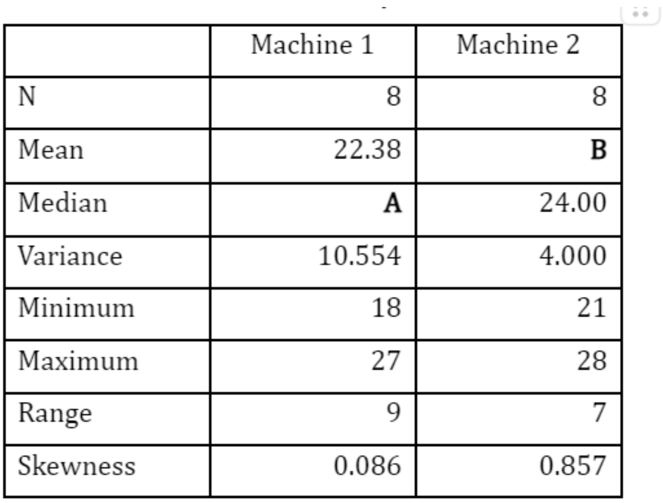
b. Find the value of Pearson Coefficient of Skewness. Then, identify the shape of distribution.

c. Based on information given in the output, sketch the appropriate diagram to represent the shape of distribution for unemployment rate.

6. A company is considering installing a new machine to assemble its product. The company is considering two types of machines, but it will buy only one type. The company selected eight assembly workers and asked them to use these two types of machines to assemble products. The data for two random samples is shown in Table 1. Table 2 shows the summary of the statistics for the time taken for the two types of machines based on data in Table 1.

Table 1: Time taken (in minutes) to assemble one unit of product on each machine

Machine 1	23	26	19	24	27	22	20	18
Machine 2	21	24	23	25	24	28	24	23

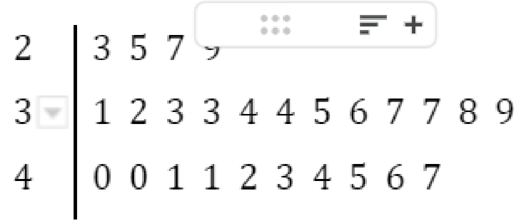


	Machine 1	Machine 2
N	8	8
Mean	22.38	**B**
Median	**A**	24.00
Variance	10.554	4.000
Minimum	18	21
Maximum	27	28
Range	9	7
Skewness	0.086	0.857

Table 2: Summary Statistics

a. Compute the value of **A** and **B**.
b. Determine which machine is more consistent in their time taken to assemble one-unit product.

7. The weight of chickens (in kg) in small poultry in Klebang were summarize as follows:



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

Key: 3|3 means

a. Calculate the mean and standard deviation.
b. Determine the median value.
c. Interpret the value obtain in (b).
d. State the name of the plot.

2.6 Answers to Exercise 2

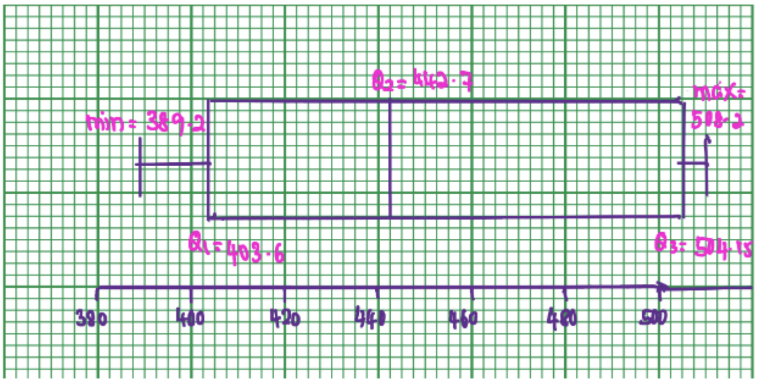
1. a) Mode = 90 kg. Most frequently weight of the customers for WX fitness studio is 90 kg.
   b) Mean = 85.42, s = 13.0486
   c) PMS = -0.3510. The distribution of data is skewed to the left.

2. a) Median (Q2) = 95.5, Q1 = 78.75, Q3 = 134.25
   b) The shape of the data distribution is skewed to the right.

3. a) Mean = 15.25, Mode = 12 and 15
   b) Since Mean > Mode, the shape of the distribution for the above data is skewed to the right.
   c) Q1 = 11.25, Q2 = 14.5, Q3 = 20.5

4. a) Mean = 67.5, median = 70, Mode = 70.
   b) Variance = 409.714
   c) PCS = -0.1235. The shape of the distribution of the test score is skewed to the left.

5. a) A = 10
   b) PCS = 0.5111 (positive skewness)
   c)



6 a) A = 22.5; B = 24.0 b) CV1 = 14.52%; CV2 = 8.33% c) Machine 2 is more consistent in time taken to assemble one unit of product compared to Machine 1.

7. a) Mean = 3.71 kg, s = 0.7022
   b) Median = 3.65 kg
   c) Half of the chicken's weights are less than 3.65kg and another half of the chickens weigh more than 3.65 kg.
   d) Stem and leaf plot.

2.7 Tutorial 2

1. The descriptive statistics for the life span (in years) of brand AA washing machine are summarized as below.

Descriptive Statistics

	N	Mean	Mode	Std. Deviation	Minimum	Maximum
Life span (years)	89	6.59	7.04	0.74	5.2	8.1

a. Calculate the coefficient of skewness. Hence, comment on the shape of the distribution.
b. Explain the meaning of the mode value.
c. Given the mean and variance for the life span (in years) of brand BB were 7.1 and 12.3 respectively. Using an appropriate measurement, determine which brand has a more consistent life span.

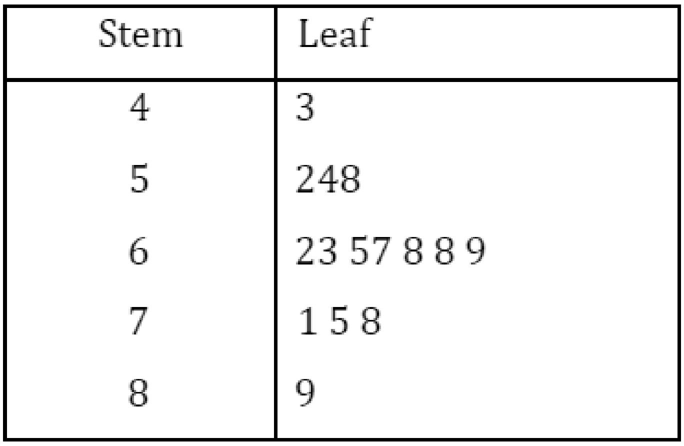
3. Waiting times (in minutes) of customers at the Providence Bank (PB) and the Valley Bank (VB) are given as follows.

PB	3.2	4.4	4.8	5.2	6.2	6.7	7.5	8.0	6.5	9.0	5.1	3.3	5.2	4.0	4.9
VB	5.8	5.6	5.7	5.8	6.1	6.4	6.7	6.7	6.7	6.8	6.5	7.0	6.9	6.5	6.4

a. Calculate the mean and standard deviation for Valley Bank waiting times.
b. From the above box-and-whisker plots of waiting times for both Providence and Valley banks, comment on the shape of distribution for each plot.
c. The mean and standard deviation for Providence Bank are 5.6 and 1.692 respectively. Determine which bank shows a more consistent wait time.

3. The stem and leaf plot below represents the History test scores (out of 100) of 15 randomly selected students in Class A.

a. Calculate the mean and standard deviation for the test score.
b. Interpret the mean obtained in a).
c. The summary statistics of the test score for History and Geography students in Class A are summarized in the following table. Using an appropriate measure, determine which distribution of the test score between the subjects is more dispersed.



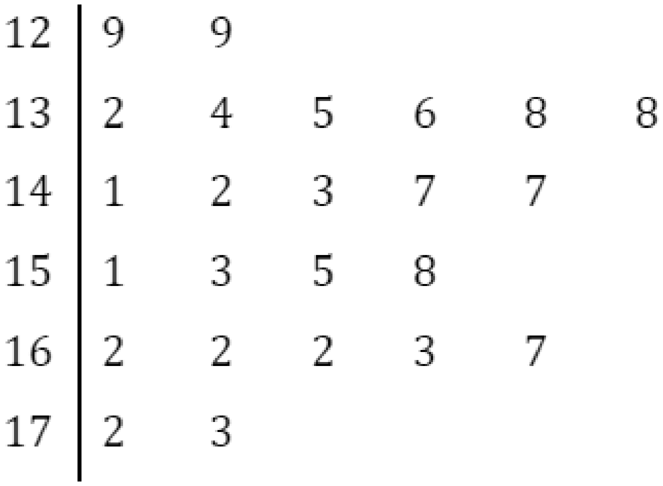
Stem	Leaf
4	3
5	248
6	23 57 8 8 9
7	1 5 8
8	9

Figure 2.16: Key: 4|3 means 43

Descriptive statistics

	N	Mean	Std. Deviation
History	15	65.4667	11.1859
Geography	17	66.9412	17.5623

4. The following chart shows the recorded weekly milk yield (in the nearest kg) for each cow selected at random from Farm A.



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

Figure 2.17: Key: 12|9 means 129

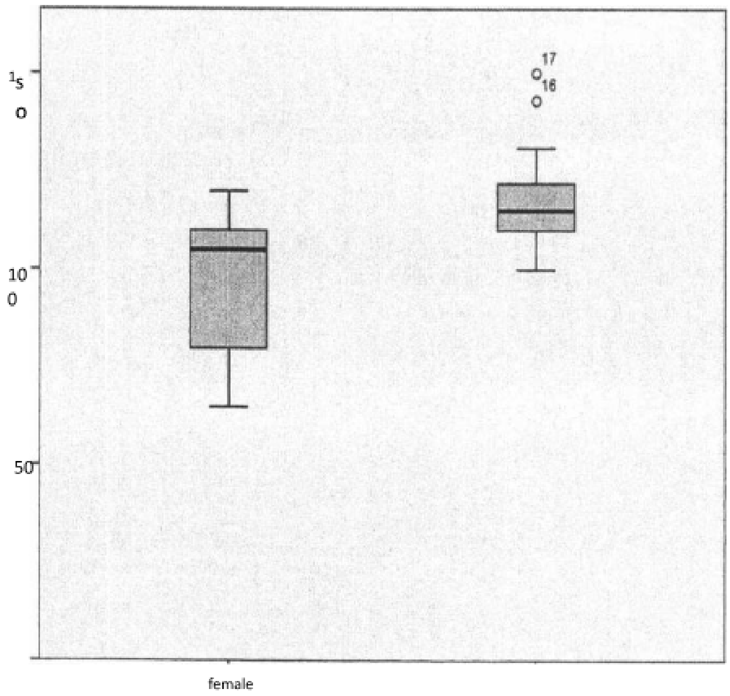
a. State the name of the above chart.
b. Find the median weekly milk yield recorded at Farm A. Hence, interpret the result.
c. The statistics for the weekly milk yields for Farm A and Farm B are summarized in the following table. Using an appropriate measure, determine which farm has more consistent weekly milk yield.

Descriptive statistics

	N	Mean	Std.

5. The following chart summarizes the weight in kilograms of 17 male and female Orang Utans in the Reserves Centre.

Gender	**Mean**	**Std. Deviation**	**Minimum**	**Maximum**
Male	117.94	13.32	100	150
Female	97.65	17.79	65	120



a. Name the diagram given.
b. Identify the outlier(s), if any.
c. Using an appropriate measure, determine which gender is more consistent in distribution of weight.

6. The weekly consumption of cheese (in ounces) for 35 participants in a nutrition study is summarize as follows:

	N	$\sum x$	$\sum x^2$	W
Cheese (in ounces)	25	265	4579	10

a. Calculate the mean and standard deviation
b. Most participants spent 10 ounces of cheese weekly. Based on the given interpretation, name the statistical measure of W.
c. Identify the skewness of the weekly consumption of cheese using appropriate measurement.

2.8 Answers to Tutorial 2

1 a) Coefficient of skewness = -0.6081 The shape of the distribution is skewed to the left. b) Mode =7.04 Most of the life span of Brand AA washing machine is 7.04 years. c) CV for Brand AA = 11.23%, CV for Brand BB = 49.40% Therefore, Brand AA has a more consistent of life span.

2. a) Mean = 6.373, s = 0.462
   b) The shape of the distribution of PB is skewed to the right. The shape of the distribution of VB is skewed to the left.

3. a) Mean = 65.467, n = 15, s = 11.186
   b) The average test score of History is 65.467.
   c) History test scores is more consistent the Geography test scores.

4. a) Stem-and-Leaf Plot
   b) Median = 147 kg 50% of the weekly milk yield is less than 147 kg.
   c) Farm B has more consistent weekly milk yield.

5. a) Box and Whisker Plot
   b) Outliers : 16, 17
   c) Male is more consistent in distribution of weight.

6. a) Mean = 10.6 ounces, s = 8.588
   b) Mode
   c) PCS = 0.07 (approximately normal)

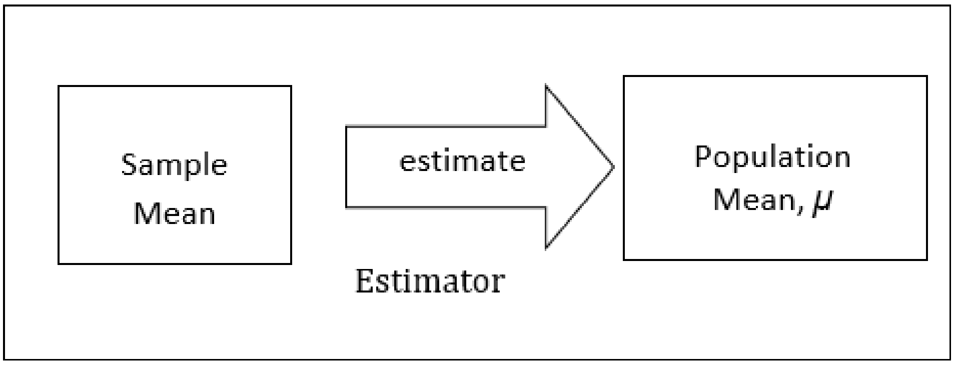
3 Estimation

Learning Objectives:

1. Describe and construct the interval estimation for the mean of the population parameter (both known and unknown) for **large and small sample size**.
2. Describe and construct the interval estimation for the difference between two means (both known and unknown) for **independent** sample.
3. Describe and construct the interval estimation for the difference between two means (both known and unknown) for **dependent** sample.

3.1 Introduction

Estimation is the process in which a numerical value collected from a sample is assigned to a population parameter. The value(s) that is assigned to a population parameter based on the sample statistic is called an estimation while the sample statistic used to estimate population parameter is called an estimator. There are two (2) types of estimates that will be discussed in this chapter name as point estimate and interval estimate.


Sample Mean
estimate
Population Mean, μ
Estimator