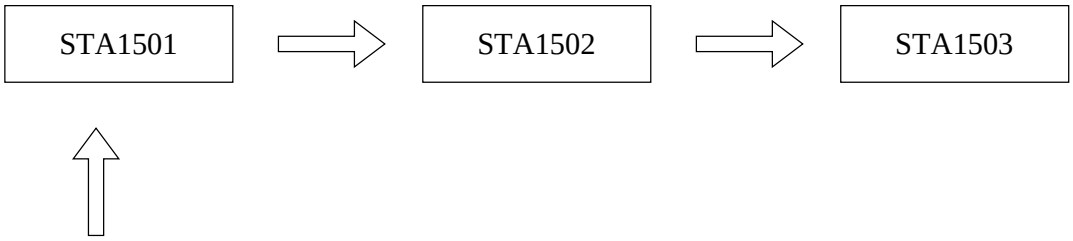
Contents

ORIENTATION

Introduction

Welcome to STA1501. This module consists of the first half of the first-year statistics modules for students in the College of Economic and Management Sciences and the first module out of three first-year statistics modules for students in the College of Science Engineering and Technology. The modules STA1501 and STA1502 form an integrated whole and are focused on the following objective: *To collect, organise, analyse and interpret data for the purpose of making better decisions.*

First−year Statistics



This is where you are

The first part of the module covers the *"Descriptive Statistics"* part, which is earthly and real and the focus is on the presentation of data. The first step is to carefully think about the **type of variable** that each measurement represents. This is extremely important as the type dictates what you can or can't do in the rest of your data analysis. Then we will also consider the **collection** of data (which most often, for the social sciences and in business applications, involve administering questionnaires and/or survey data, and **sampling** plays an important role in this regard). Between the **collection of data** and the ultimate goal of **analysis of data** lies the very important step of **organising and summarising the data.** So, in this module we discuss how we *organise and summarise the gathered information intelligibly and efficiently.*

The second part of the module covers the *"Probability and Probability Distributions"* part where we leave the practical familiarity of data and turn to the less familiar abstract concept of probability. This is almost like a shift in gears! A proper understanding of the laws of probability is essential to ensure a proper understanding of the mechanisms underlying statistical data analysis. **Probability theory is the tool that makes statistical inference possible**.

The third part of the module covers the applications of the probability distributions. You have learned that the shape of the normal distribution is determined by the value of the mean μ and the variance σ^2, whilst the shape of the binomial distribution is determined by the sample size n and the probability of a success p.

These critical values are called *parameters*. We most often don't know what the values of the parameters are and thus we cannot "utilise" these distributions (i.e. use the mathematical formula to draw a probability density graph or compute specific probabilities) unless we somehow *estimate these unknown parameters*. It makes perfect logical sense that to estimate the value of an unknown population parameter, we compute a corresponding or comparable characteristic of the sample.

The objective is to *draw inference* about a population (a complete set of data) based on the limited information contained in a sample. In dictionary terms, *inference* is the act or process of inferring; to *infer* means *to conclude or judge from premises or evidence;* meaning to derive by reasoning. In general, the term implies a conclusion based on experience or knowledge. More specifically in statistics, we have as evidence the limited information contained in the outcome of a sample and we want to conclude something about the unknown population from which the sample was drawn. The set of principles, procedures and methods that we use to study populations by making use of information obtained from samples is called *statistical inference.*

Learning objectives

There are very specific outcomes for this module which we list below. Throughout your study of this module you must come back to this page, sit back and reflect upon these outcomes, think them through, digest them into your system and feel confident in the end that you have mastered them.

- Analyse data considering different types of data and how they relate to relevant graphical and tabular presentations e.g. pie charts, bar charts, histograms, stem-and-leaf displays, line charts, scatter diagrams and box-and-whisker plots
- Analyse data by calculating accurate numerical measures of central location, variability, relative standing and linear relationship.
- Differentiate between simple random sampling, stratified random sampling and cluster sampling and implement a sampling plan for a given research problem with an awareness for the effect of sampling errors.
- Describe the different concepts and laws of probability and apply definitions of joint, marginal and conditional probability.
- Apply the complement, multiplication and addition rules and probability trees for calculation of more complex events and calculate complicated events from the probabilities of related events.
- Understand the role of probability in decision making and the application in basic statistical inference.
- Describe random variables and the probabilities associated with them in the form of a table, formula or graph and also in terms of its parameters, usually the expected value and the variance.
- Describe different probability distributions as either discrete or continuous and know the parameters of expected value and variance

The prescribed textbook

For this module you must study **twelve chapters** from the **prescribed textbook:**

Keller, G. (2018, (11^{th} edition))

STATISTICS FOR MANAGEMENT AND ECONOMICS,

South–Western, Cengage Learning

Chapter 1: WHAT IS STATISTICS?

Chapter 2: GRAPHICAL AND TABULAR DESCRIPTIVE TECHNIQUES

Chapter 3: ART AND SCIENCE OF GRAPHICAL PRESENTATIONS

Chapter 4: NUMERICAL DESCRIPTIVE TECHNIQUES

Chapter 5: DATA COLLECTION AND SAMPLING

Chapter 6: PROBABILITY

Chapter 7: RANDOM VARIABLES AND DISCRETE PROBABILITY DISTRIBUTIONS

Chapter 8: CONTINUOUS PROBABILITY DISTRIBUTION

Chapter 9 : SAMPLING DISTRIBUTIONS

Chapter 10: INTRODUCTION TO ESTIMATION

Chapter 11: INTRODUCTION TO HYPOTHESIS TESTING

Chapter 12: INFERENCE ABOUT A POPULATION

The study guide

The study guide is exactly what its name implies: a guide through the textbook in a systematic way. The textbook will focus on the theoretical contents of the module.. For each separate study unit you should first study the work in the textbook and utilise the guide to assess your progress, test your knowledge and prepare for the examination. In other words, the study guide will provide you with an opportunity to apply your knowledge of the material that is covered in the textbook.

Study units and workload

We realise that you might feel overwhelmed by the volumes and volumes of printed matter that you have to absorb as a student! How do you eat an elephant? Bite by bite! We have divided the twelve chapters of the **textbook** into **11 study units** or "sessions". Make very sure about the sections in each study unit since some sections of the textbook are not included and we do not want you frustrated by working through unnecessary work. The study units vary in length but you should try to spend *on average 12 hours on each unit*. Practically everybody should be able to do statistics. It depends on the amount of TIME you spend on the subject. Regular contact with statistics will ensure that your study becomes personally rewarding.

Try to work through as many of the exercises as possible.

Final word: Attitude

You are the master of your own destiny. Studying through distance education is neither easy nor quick. We know that many of you have some "math anxiety" to deal with, but we will do our best to make your statistics understandable and not too theoretic. Studying statistics is sometimes not "exciting" or "fun" and keep in mind that to master the content of a module can involve considerable effort. However, we do claim that knowledge of statistics will enable you to make effective decisions in your business and to conduct quantitative research into the many larger and detailed data sources that are available. Statistical literacy will enable you to understand statistical reports you might encounter as a manager in your business. We are there to assist you in a process where you shift yourself from a supported school learner to an independent learner. There will be times when you feel frustrated and discouraged and then only your attitude will pull you through!

In a paper by Sue Gordon[1] (1995) from the University of Sydney, the following metaphor is given: "The learning of statistics is like building a road. It's a wonderful road, it will take you to places you did not think you could reach. But when you have constructed one bit of road you cannot sit back and think 'Oh, that's a great piece of road!' and stop at that. Each bit leads you on, shows the direction to go, opens the opportunity for more road to be built. And furthermore, the part of the road that you built a few weeks ago, that you thought you were finished with, is going to develop potholes the instant you turn your back on it. This is not to be construed as failure on your part, this is not inadequacy. This is just part of road building. This is what learning statistics is about: go back and repair, go on and build, go back and repair."

A few logistical problems

Decimal comma or point?

We realise that in the South African schooling system commas are used to indicate the decimal digit values. You have been penalised at school for using a point. Now we sit between two fires: the school system and common practice in calculators and computers! Most computer packages use the decimal point (ignoring the option to change it) and *Keller* (the author) also uses the decimal point in our textbook (*Statistics for Management and Economics*). Thus, we shall use the decimal point in our study guide, assignments and the examination.

Role of computers and statistical calculators

The emphasis in the textbook is well beyond the arithmetic of calculating statistics and the focus is on the identification of the correct technique, interpretation and decision making. This is achieved with a flexible design giving both manual calculations and computer steps.

Every statistical technique that needs **computation** is illustrated in a three-step approach:

Step 1 MANUALLY

Step 2 EXCEL

Step 3 MINITAB

It is a good idea that you initially go through the laborious manual computations to enhance your understanding of the principles and mathematics but we strongly urge you to manage the Excel computations because using computers reflects the real world outside. The additional advantage of using a computer is that you can do calculations for larger and more realistic data sets. Whether you use a computer program or a statistical calculator as tool for your calculations is irrelevant to us. However, the emphasis in this module will always be on the interpretation and how to articulate the results in report writing.

CD Appendixes and **A Study Guide** are provided on the CD-ROM (included in the textbook) in pdf format. Although it will not be to your disadvantage if you do not use the CD we encourage you to try your best to have at least a few sessions on a computer. Statistical software makes Statistics exciting – so, play around on the computer should you have access!

Study Unit 1

WHAT IS STATISTICS?

1.1 Learning outcomes

By the end of this unit, you should be able to

1. define Statistics, population and sample
2. distinguish between statistics and statistic
3. classify data and describe the types of variable

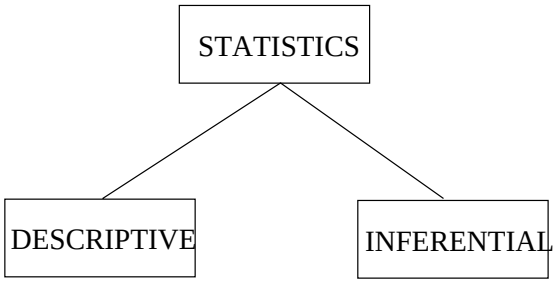
1.2 Introduction

The objective of statistics is collect, analyse data and to draw conclusions about a population based on information contained in a sample. In other word statistics is a method to convert data into information.

You need not panic when all the new terms do not make sense to you, neither need you remember them all at this stage! As we proceed chapter by chapter and you start applying the different techniques you will understand more. *In study unit 5* you will learn most you need to know about data collection and sampling to obtain optimum information and you will learn that there are good and bad ways of obtaining a sample.

1.3 Definitions

Statistics is a science that involves collecting, summarising , analysing and interpretation of data for the purpose of making informed decisions. Statistics is subdivided into two branches namely:



- descriptive statistics and
- inferential statistics.

Descriptive statistics is a branch of statistics that deals with the collection, presentation and characterisation of data in order to properly described the various features of the data.

While *Inferential statistics* is an estimation process which used the result of a sample statistics to make decisions or draw conclusions about a population parameters.

A *population* is the collection of all items, elements or objects that we wish to study. A population does not necessary means a group of people living in a specific area. For instance population of South Africa. In Statistics, a population is referred to a group of all individuals, elements, items of interest to a statistics practitioner.

A *sample* is a fraction or a subset of population

When making conclusions about a population based on a sample, the conclusions and estimates may not be perfect. To minimise the level of uncertainty when making decisions, 2 reliable measures are used in statistical inference: the confidence interval and the level of significance. Details on this measure will be provided at a later stage on Unit 10 and 11.

1.4 Parameter versus a statistic

Parameter is a numerical value that describes or summarises a population, that is, describes the characteristics of a population while a *statistic* is a numerical value that describes the characteristics of a sample or summarises a sample.

1.5 Types of variable and information

There are two types of data, namely quantitative and qualitative data. Quantitative data generates numerical variables and they are usually reported numerically, while qualitative data generates categorical variables.

A quantitative variable can be either discrete or interval. A discrete variable is countable and are referred to as a whole number while an interval variable can assume any given value within a given interval.

1.5.1 Examples of Quantitative variable

The number of students in classroom

Temperature of water

Number of provinces in the country

Salaries

Number of respondent to a consumer survey

1.5.2 Examples of Qualitative variables

Make of a motorcar (Toyota, BMW, Fiat, etc.)

Level of education (primary, certificate, matric, diploma, degree, etc.)

Sex or gender of individuals (Male or female)

Marital status of an individual (single, married, divorced, widow)

Size of soft drink(500ml, 1litre, 2 litres, 2.25 litres)

1.6 Measurement scale

The measurement scale is a process which assign numerals to objects or events according to rules. Numerals used in the measurement scale are not necessarily numbers. Most books indicate that there are four measurements of scale which are nominal, ordinal, interval and ratio. However the author Keller indicates that there are three main measurement scales namely the nominal, ordinal and interval measurement scales. He has combined the interval and ratio scale as just interval scale.

1.6.1 Nominal level of measurement

Nominal scale applies to names. This measurement scale is used for objects or elements which consists of names. There are codes or labels for this reason we are unable to perform any calculations on these codes or labels.

Examples of nominal measurement scale:

Name of students

Types of vegetables

Number on rugby jersey

Type of a car

1.6.2 Ordinal level of measurement

Ordinal level of measurement presumes that one classification is ranked higher than another. The items or object differ form one to the other one but have more or less of a characteristic than another. In this level of measurement the order of the variables is meaningful.

Example of ordinal level of measurements:

Rate of the nurse at the clinic (effective, very effective, not effective)

Ranking of academic Staff at UNISA (junior lecturer, Lecturer, Senior lecturer, associate professor, professor)

The size of a T-shirt (small, medium, large, extra large)

1.6.3 Interval level of measurement

In the interval scale of measurement all calculations are permitted in the data. The interval data of measurement is used for real numbers. A distiguishing characteristic of this scale is that the differences between the consecutive numbers are of equal intervals and we can interpret differences in the distance along the scale. It can also be referred to as quantitative or numerical (Keller & Gaciu, 2015:12). The authors further indicated that ratio data is a special kind of the interval data and the ratio data measurement scale is used to express the ratio of some of the values of interval data, so that the numbers can be compared as multiples of one another.

Examples of interval level of measurements:

Weight of a new born baby
Number of people in a room
Time taken to write an exam
Temperature of water
Income of a person

Most computer package like IBM SPSS also classifies the measurements as nominal, ordinal and scale where scale represents interval data.

At this stage we only want you to be aware of Microsoft Excel and of the fact that the CD contains an additional *statistical software add-in* for Excel which will enable you to do all the statistical procedures that is covered in the textbook. You can always come back to these pages when you need to know something more about Excel.

We do not expect you to master MINITAB – mainly because it would imply additional costs to obtain the software. However, should you have access, it is a wonderful feeling to be master of a proper all-embracing statistical package such as MINITAB, SPSS or JMP. If you decide to continue with further studies in statistics we will formally introduce you to a statistical package in your second year.

Key terms

descriptive statistics

statistical inference

confidence level

significance level

population

sample

nominal data

ordinal data

interval data

parameter

statistic

1.7 Study Unit 1: Summary

I. Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

II. There are two types of statistics:

   A. *Descriptive statistics* are procedures used to organize and summarize data.

   B. *Inferential statistics* involve taking a sample from a population and making estimates about a population based on the sample results.

      1. A *population* is an entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
      2. A *sample* is part of the population.

III. There are two types of variables:

   A. A *qualitative variable* is categorical or nonnumeric.

      1. Usually we are interested in the number or percentage of observations in each category.
      2. Qualitative data are usually summarized in graphs and bar charts.

   B. There are two types of *quantitative variables* and they are usually reported numerically.

      1. *Discrete* variables can assume only certain values, and there are usually gaps between values.
      2. An *interval* variable can assume any value within a specified range.

IV. There are three levels of measurement:

   A. With the *nominal level*, the data are sorted into categories with no particular order to the categories.

   B. The *ordinal level* of measurement presumes that one classification is ranked higher than another.

   C. The *interval level* of measurement all calculations are permitted in the data.

References

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Keller, Gerald and Gaciu, Nicoleta (2015) *Managerial Statistics*, Europe, Middle East and African Edition, First Edition, Cengage Learning EMEA. Hampshire, United Kingdom

Gordon, Sue (1995) *A theoretical Approach to Understanding Learners of Statistics*. Journal of Statistics Education v. 3, n.3 University of Sydney.]

Study Unit 2

GRAPHICAL DESCRIPTIVE TECHNIQUES

2.1 Learning outcomes

After completing this unit, you should be able to

1. distinguish types of data
2. compile and interpret a frequency table for nominal data
3. present and portray nominal data using a bar chart
4. present and portray nominal data using a pie chart
5. present and portray interval data using graphs and tables
6. describe the relationship between two interval data
7. describe time series data

2.2 Types of data and information

Always remember that the type of data dictates what you can or can't do in the rest of your data analysis.

Knowing about variable types and collecting data is almost like a chicken-and-egg situation! Which one comes first? It is extremely important to carefully think about the **type of variable** that each measurement

represents, because it could influence the manner in which the measurements will be obtained. For example, suppose you compile a questionnaire where the respondent can tick one of the following age categories:

Age	Please Tick
20–25	
25–35	
35–45	
45–65	

The resulting data will be considered as **ordinal measurements** (i.e. if ages are artificially grouped into categories).

But, if the true age of each respondent can be determined to the nearest day, the data can be considered as individual points on an interval scale and we have the strongest level of measurement.

2.2.1 Introduction

We emphasised in study unit 1 how extremely important it was to carefully think about the **type of variable** that each measurement of a data set represents because the type dictates what you can or can't do in the rest of your data analysis. Consciously remind yourself to think about the data type whenever you are busy doing something with data.

The **final mind map** we are working towards (after completion of Statistics 1) and which you have to make part of yourself, is given on the inside cover of your textbook.

2.2.2 Definitions of key terms

For one to be able to gather all the needed information from the data, the following concepts need to be known.

Variable: a variable is characteristic of interest of a population or a sample.

Values of the variables are possible observations of the variable

Data are observed values of a variable

2.2.3 Classification of data

Data can be classified as nominal, ordinal or interval

Intervals data are real numbers such as income, the price of an item, amount of water, number of students. There are also called quantitative or numerical while nominal data are categorical data. There are codes or labels for this reason we are unable to perform any operations including calculations, divisions, multiplication and subtraction. Ordinal data is ranked data.

2.2.4 Examples

1. Name of students: Bongani, James, Zulu is example of nominal data

2. Response to a market research survey measured on the likert scale using the code:1= Strongly agree, 2= Agree, 3 =Neutral , 4= disagree and 5= Strongly disagree is an example of ordinal data.

3. Temperature on the rugby filed during the Super Twelve competition is an example of interval data.

2.3 Describing set of nominal data

The only recognised operation on nominal data is counting, we are able to count each nominal data and the occurrence of each nominal data. The occurrence of each data is referred to as the frequency. We can list nominal data, group them in the table and find the categories of the count of each observations.

2.3.1 Definitions

A *frequency distribution* is a table that records each data and displays the possible categories of each data.

A *relative frequency distribution* is the proportion or the percentage of the observations. In others word, the relative frequency is the frequency of each category divided by the sum of all frequencies.

2.3.2 Example

The blood types for a sample of 20 students were recorded as follows:

$$AB, A, B, O, AB, O, A, A, O, A, A, AB, O, AB, A, B, AB, A, B \text{ and } O.$$

Construct a frequency distribution and calculate the relative frequency.

The table is given below:

Categories	frequencies (f)	relative frequencies (% f)
A	7	$\frac{7}{20} = 0.35$
B	3	$\frac{3}{20} = 0.15$
AB	5	$\frac{5}{20} = 0.25$
O	5	$\frac{5}{20} = 0.25$
Total	20	1

Different techniques are used to summarise and display nominal data including bar graphs and pie charts. The types of graphs depends on the type of data. For nominal data, bar graphs and pie charts are graphical techniques used to summarise and portray them in graphs.

2.3.3 Bar graph

A *bar graph* is a chart drawn using rectangular bar to display the possible categories of data along with the frequencies. Bar charts focus the attention on the frequency of the occurrences of the categories.

2.3.4 Example

Voters participating in an election exit poll in Minnesota (USA) were asked to state their political party affiliation. Coding the data as 1 for Republican, 2 for Democrat, and 3 for Independent, the data collected were as follows:

3 1 2 3 1 3 3 2 1 3 3 2 1 1 3 2 3 1 3 2 3 2 1 1 3

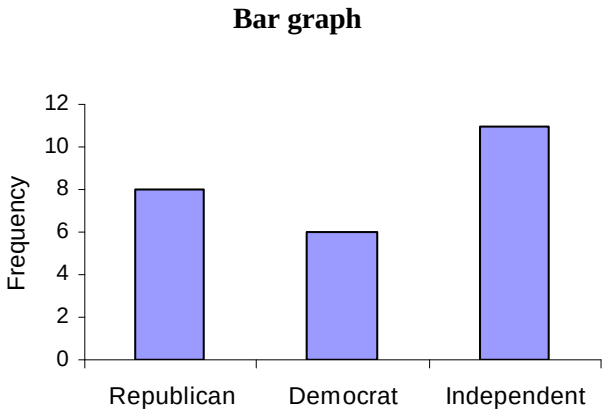
Construct a frequency bar chart.

Manually

We need a *frequency table* before we can draw a *frequency bar chart*.

	Republican (1)	Democrat (2)	Independent (3)
Tally marks	𝍸 |||	𝍸 |	𝍸 𝍸 |
Frequency	8	6	11

Using Excel

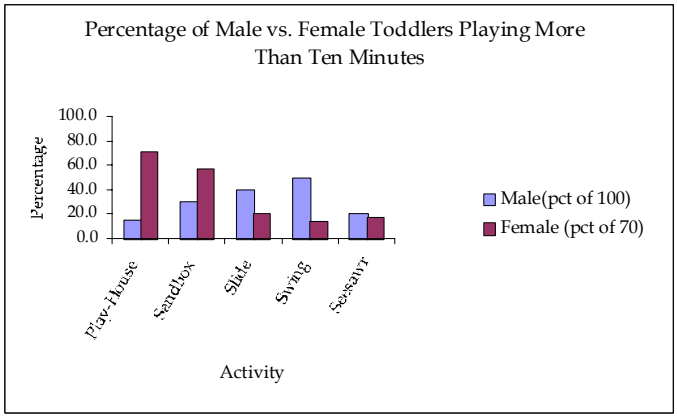


2.3.5 Comparative bar graph

The comparative bar graph is used to compare 2 or more data sets. Bars are grouped together in each categories with spaces on the x-axis intervals.

2.3.6 Example

(a) Cluster bar chart showing, for each play activity, the fraction of all male toddlers who played on each activity for more than ten minutes, as compared to the fraction of female toddlers.



Bar chart displaying the total number of male toddler-play-units for the playhouse and sandbox, versus, the total number of units for the slide and swing.

2.3.7 A pie chart

A circle is drawn and "cut" like a pie. The size of each slice represents a particular frequency. The pie chart emphasises the proportion of occurrences of each category. One of the advantages of a pie chart is that it clearly shows that the total of all the categories of the pie adds to 100%.

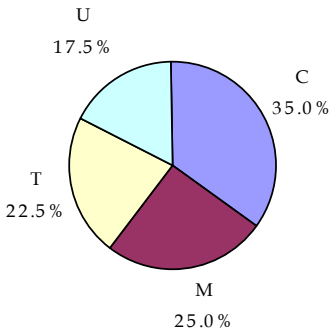
2.3.8 Example

Car buyers were asked to indicate the car dealer they believed offered the best overall service. The four choices were Carriage Motors (C), Marco Toyota (M), Triangle Auto (T) and University Chevrolet (U). The following data were obtained:

T C C C C M T C U U M C M T C M M C M U
T C C T U M M C C T T U C U T M M C U T

Construct a pie chart. Which car dealer offered the best overall service?

Solution:



It seems that Carriage Motors offered the best overall service.

You will not find one chart in the study guide (or textbook) that looks "manually drawn" simply because all printed matter needs to be in electronic format for the production process. In pre-computer days statisticians produced the same pie and bar charts manually and you should not feel discouraged if you do not have the software to produce them, as long as you understand how to construct them.

For a "manually drawn" pie chart the steps are as follows:

(i) Convert a *frequency table* $\longrightarrow$ *proportional frequency table* $\longrightarrow$ $\frac{\text{proportion}}{100} \times 360$.

(ii) Find a pencil, a compass and a protractor.

(iii) Draw a circle (remember to mark the centre!) and mark the sections according to the calculations in (i).

:-) Easy as pie!

2.4 Describing sets of interval data

2.4.1 Frequency distribution for interval data

A frequency distribution is a table which records each data and the number of times that data occurs in the distribution.

To construct a frequency distribution one need to compute the range, determine the number of class intervals and class width.

Steps on how to construct a frequency distribution for interval data:

1. Calculate the range

   The range is the difference between the largest value and the smallest value. To compute the range we need to identify the lowest and the highest values in the distribution, X Maximum$-X$ Minimum

2. Determine the number of class interval

   The number of class interval $= 1 + 3.3\log(n)$, which is symbolised by k, and n is the number of observations, we need to count the number of data (observations) in the data in order to obtain n. The class intervals must include all the data, the lowest data must fit in the first class interval while the biggest data must fit in the last class interval. The class intervals should be mutually exclusive and no observation can be classified into two different intervals. The number of class intervals depends entirely on the number of observations in the data set. The more observation we have, the larger the number of class intervals we need to use to draw a useful histogram

3. Determine the class width

   The class width is obtained by subtracting the smallest observation from the largest and dividing the difference by the number of classes.

   $$\frac{Range}{K}$$

4. Choose the smallest data or a convenient data that is smaller as the lower boundary of the first class interval.

5. Add the class width to the value chosen on step 4 to get the second lower class boundary, then add the class width to the second lower class boundary to get the third class and so on. List the class on the vertical column.
6. Count the number of tallies in each class
7. The total of the frequencies for all class interval must be equal to the sample size of the original data.

2.4.2 Graphs for interval data

The stem and leaf portrays individual data in a numerical order and summarises data in a table by separating each value into two parts: the stem and the leaf. The stems are the leading digit(s) and are displayed in the vertical position . While the leaf consists of the rests of the digit(s).

The most used graph for interval data is the histogram. The histogram is the most common from of graphical representation of data. it is a series of rectangles of equal widths whose heights represents the frequencies of each class intervals.

For a “manually drawn” histogram you need:

(i) A *frequency distribution* $\longrightarrow$ This is the laborious part.
(ii) A pencil and a ruler.

2.4.3 Example

The ages of a sample of 25 salespersons are as follows:

47	21	37	53	28
40	30	32	34	26
34	24	24	35	45
38	35	28	43	45
30	45	31	41	59

(a) Draw a stem-and-leaf display.
(b) Draw a histogram with four classes.
(c) Draw a histogram with six classes.

Solution:

(a) Stem-and-leaf display

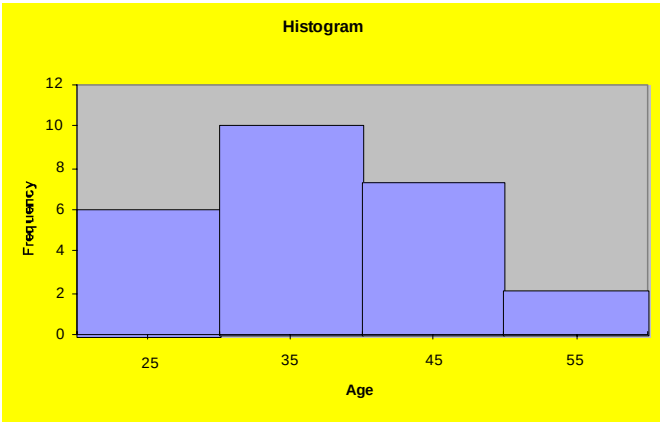
STEM	LEAF
2	1 4 4 6 8 8
3	0 0 1 2 4 4 5 5 7 8
4	0 1 3 5 5 5 7
5	3 9

(b) Histogram with four classes

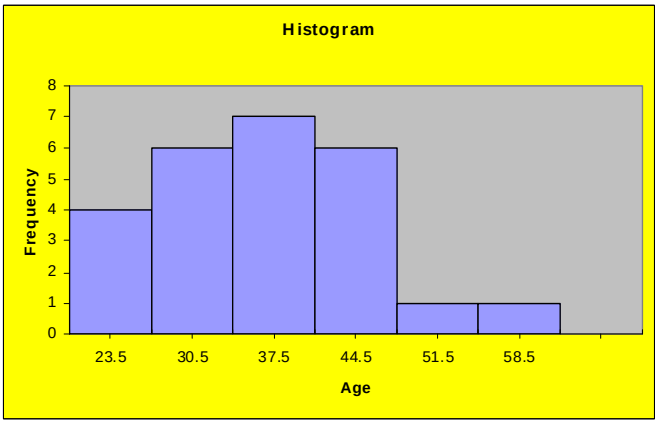
$$\textit{Interval width} = \frac{\text{largest value} - \text{smallest value}}{\text{number of intervals}} = \frac{59-21}{4} = \frac{38}{4} \approx 10$$

Class limits	Number of salespersons
20 and less than 30	6
30 and less than 40	10
40 and less than 50	7
50 and less than 60	2

Solution:



(c) Histogram with six classes

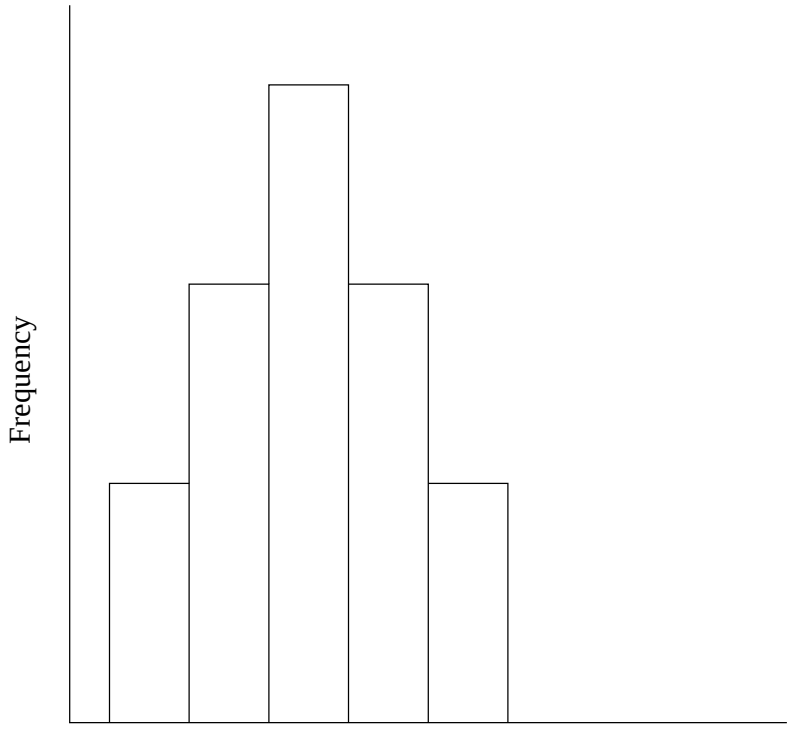


2.4.4 Shapes of histograms

The purpose of drawing histogram is to summarize data graphically in away that an overall picture of the data can be presented and more information can be obtained from the distribution. We describe the shape of the histogram on the basis of the following characteristic:

1. Symmetry

   A histogram is said to be symmetric if, the right side of histogram is identical to the left side. The left and right side of the histogram are identical in shape and size

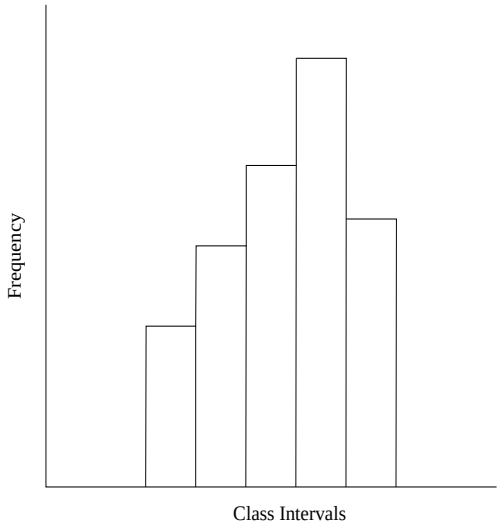


2. Skewness

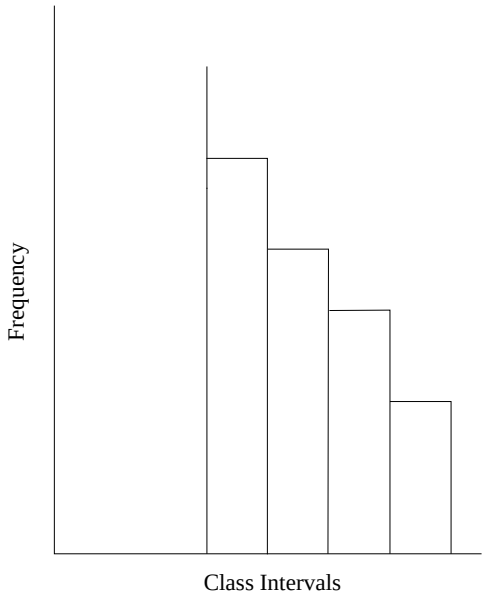
   There are two types of skewed histogram namely skewed positively and skewed negatively

   A positively skewed histogram is a histogram that have long tail extending to the right while a negatively skewed histogram is an histogram that have a long tail extending to the left.

Negatively Skewed Distribution



Positively Skewed Distributions



2.5 Describing the relationship between two variables and describing time series data

(When we have bivariate data on *two numerical variables,* we can also compute something additional which is a **measure of the relationship between them.** This will be formally treated in chapter 4, where the relationship between them is quantified.)

2.5.1 Describing time series data

Time series data are often portrayed on line chart. A line chart is a plot of the variable over time. To construct a line chart, we need to draw a pair on the x axis and y axis and plot the value of variable on the vertical axis and the time period on the horizontal axis.

2.5.2 Describing the relationship between two interval variables

The graph that describe the relationship between two interval variables is called the scatter diagram.

To construct a scatter diagram, we need bivariate data where one is the independent variable symbolised by X and the other the dependent variable symbolised by Y.

2.5.3 Direction

If the independent variables and dependent variables move in the same direction (X increases and Y increases), we conclude that there is positive relationship between the two variables. If they are moving in opposite direction (X increases and Y decreases or Y increases and X decreases, we say that there is negative relationship between the two variables. More details on the relationship between two intervals variables will be covered in the next chapter.

2.5.4 Example

A professor of economics wants to study the relationship between income and education. A sample of 10 individuals is selected at random. The data below shows their income (in R10 000) and education (in years).

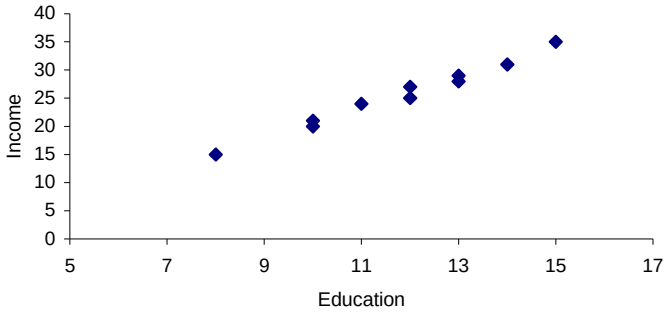
Education	12	14	10	11	13	8	10	15	13	12
Income	25	31	20	24	28	15	21	35	29	27

a. Draw a scatter diagram for the data with the income on the vertical axis.

b. Describe the relationship between income and education.

Solution:

(a)



(b) There is a very strong positive linear relationship between education and income; as years of education increase, there is a definite tendency for income to increase linearly.

2.6 Graphical excellence and graphical deception

2.6.1 Introduction

In the previous study unit you learned about the appropriate graphical and tabular techniques for nominal as well as interval data. The emphasis was on the techniques as such and we did not embroider on the pitfalls. It is important always to remember that the motivation behind graphs is that they add flavour and interest to data organisation. Graphical presentations most often catch a reader's attention and are usually more easily interpreted than tables, but keep in mind that they never create new information. Graphs could have the effect of leading one to conclusions that are more extreme than the pure facts of a table! In fact, they could actually lead to mis-interpretations! Whenever you are in a decision-making situation you should train yourself to see through the visual image into the underlying set of facts. The proper (and safe) way to read a graph of any kind is to carefully think about the *scales* on the vertical and horizontal axes because "blowing up" of a scale could make differences look greater. The cheapest shot to try and "lie with statistics" is to deceive with a graph!

To stress the importance of graphical excellence and the danger of possible graphical deception, Keller devotes a whole chapter (however short it might be!) to this topic.

2.6.2 Graphical excellence and graphical deception

Some of the characteristics of excellent graph are:

1. Salient feature of the data set must be easily understood by the viewer.
2. The graph encourages the viewer to compare two or more variables.
3. The graph must contain worthwhile information.
4. There is no distortion of what the data reveal.
5. Graphical excellence induces the viewer to address the substance of the data and the greatest number of ideas.

Some of the characteristics of graphs deception are; graphs without scale, graphs with different captions, distorted bar graphs, distorted histograms, graphs with no labels.

Key terms

frequency table

pie chart

bar chart

stem-and-leaf display

histogram

univariate data

bivariate data

clustered bar chart

scatterplot

2.7 Study Unit 2: Summary

I. A *frequency table* is a grouping of qualitative data into mutually exclusive classes showing the number of observations in each class.

II. A *relative frequency table* shows the fraction of the number of frequencies in each class.

III. A *bar chart* is a graphic representation of a frequency table.

IV. A *pie chart* shows the proportion each distinct class represents of the total number of frequencies.

V. A *frequency distribution* is a grouping of data into mutually exclusive classes showing the number of observations in each class.

   A. The steps in constructing a frequency distribution are as follows:

      1. Decide on the number of classes.
      2. Determine the class interval.
      3. Set the individual class limits.
      4. Tally the raw data into classes.
      5. Count the number of tallies in each class.

   B. The class frequency is the number of observations in each class.

   C. The class interval is the difference between the limits of two consecutive classes.

   D. The class midpoint is halfway between the limits of consecutive classes.

VI A *relative frequency distribution* shows the percentage of observations in each class.

VII. There are three methods for *graphically* portraying a frequency distribution.

   A. A histogram portrays the number of frequencies in each class in the form of a rectangle.

   B. A frequency polygon consists of line segments connecting the points formed by the intersection of the class midpoint and the class frequency.

   C. A cumulative frequency distribution shows the number or percentage of observations below given values.

References

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 3

NUMERICAL DESCRIPTIVE TECHNIQUES

3.1 Learning outcomes

After completing this unit, you should be able to:

1. define and compute measures of central tendency including the mean, median and mode
2. define and compute measures of variability including the range, variance and standard deviation
3. define and compute measures of relative standing including the quartiles and interquartiles
4. define and compute measures of linear correlation including the covariance, the coefficient of correlation, the least squares line and the coefficient of determination
5. explain the correlation, the coefficient of determination and the linear correlation

3.2 Measures of central tendency

Measures of central tendency are measures of location. They are typical values that describes or summarise the distribution, The most used measures of location are the arithmetic mean, the median and the mode.

3.2.1 The arithmetic mean

The *arithmetic mean* (the average) is the sum of all the values in the data set divided by the number of observations.

The population mean which is symbolised by μ and is given by $\frac{\sum x}{N}$. The sample mean is symbolised by $\bar{X}$ and is given by $\frac{\sum x}{n}$.

Steps on how to compute the mean;

1. Add all the given observations and find the total ($\sum x$) (i.e., the sum)
2. Count the number of observations (n)
3. Divide the total number of observations (sum) by the number of observations (n)

3.2.2 Example

The following are the exam marks obtained by a sample of 10 students:

75 76 52 43 83 24 52 96 47 92

Calculate the mean

$$\begin{aligned} \sum x &= 75+76+52+43+83+24+52+24+52+96+47+92 \\ n &= 10 \end{aligned}$$

The mean $= \frac{640}{10} = 64$

The average mark obtained by the sample of 10 students is equal to 64%.

3.2.3 The median

The *median* is the centre of the distribution when data is arranged in ascending or descending order. It divides the data set in two parts, n halves below and above for an ordered data set. It is that value with 50% of the observations less or equal to it and 50% of the observations above or equal to it.

Steps to follow when computing the median

1. Arrange data in a numerical order; starting from the lowest data to the highest
2. Count the number of observations (n)
3. Determine the median position : $\frac{n+1}{2}$
4. Read the median value from the ordered data

Note that if the number of observations (n) is odd , the median is the value that is exactly in the middle of the list. But if the number of observations is even, the median is computed by calculating the average of the two middle numbers.

3.2.4 Examples

The time (in hours) spent to complete the second assignment by a sample of 7 honours students is given by :

4 9 5 6 7 5 8

Find the median

Step 1: 4 5 5 6 7 8 9

Step 2: $n = 7$

Step 3: the position of the median is $\frac{7+1}{2} = 4$

Step 4: the median is the value which is on the fourth position and 6 is the median.

The ages(in years) of a sample of 10 accountants are as follows:

45 46 60 32 43 29 30 47 63 50

Find the median

5.5

Step 1 :

29 30 32 43 45 46 47 50 60 63

5 6

Step 2 : $n = 10$

Step 3 : $\frac{10+1}{2} = 5.5$

Step 4: 5.5 is between 5 and 6, we need to add the value on position 5 and the value on position 6 and divide the sum by 2, $\frac{45+46}{2} = 45.5$

The median is 45.5

3.2.5 The mode

Definition

The *mode* is the most frequent value in the data set

3.2.6 Example

Using example 4.1.2

75 76 52 43 83 24 52 96 47 76

The most frequent value in the data set is 76, the mode is 76.

It is extremely important that you know how to compute the sample mean, $\overline{x}$, and that you feel comfortable with the mathematical expression: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The mean plays an important part in many of the statistical analyses you will encounter in *Statistical Inference I,* i.e. in STA1502.

3.2.7 Mean, Mode, Median: Choosing the best measure of central tendency

There are several factors to take into considerations when choosing the best measure of location. The mean is the most preferred measure of central location for quantitative data. However for a skewed distribution the median is the best measure to choose. The mode is seldom the best measure of central location. For ordinal and nominal data the mean is not an appropriate measure of location but the mode is the appropriate measure of location for nominal and ordinal data.

3.3 Measures of variability

The measures of variability describes the amount of variation or spread in a data set. There are also called measures of dispersion. The most used measures of variability are the range, the variance, the standard deviation and the coefficient of variation.

3.3.1 The range

The range is the difference between the largest value and the smallest value. To compute the range we need to identify the lowest and the highest values in the distribution,

$$X \text{ Maximum} - X \text{ Minimum}$$

3.3.2 The variance

The variance and its related measure, the standard deviation measure the amount of variation of the data around the mean. They measure " how far each data value is far from the mean".

It is extremely important that you know how to compute the sample variance, s^2, and that you

feel comfortable with the mathematical expression

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

The variance plays an important role in many of the statistical analyses you will encounter in *Statistical Inference*.

3.3.3 The standard deviation

The sample standard deviation is simply the positive square root of the variance which is symbolised by $s = \sqrt{variance}$

3.3.4 The coefficient of variation

The coefficient of variation of a distribution is the standard deviation of the data set divided by their mean. It is a relative measure of dispersion, which is expressed as a percentage and symbolised by CV.

$$CV = \frac{mean}{standard\ deviation}$$

3.3.5 Example

The following data represent the *mass in kilograms* of a sample of 25 business class passengers plus their luggage on an aeroplane:

164, 148, 137, 157, 173, 156, 177, 172, 169, 165, 145, 168, 163, 162, 174, 152, 156, 168, 154, 151, 174, 146, 134, 140, and 171.

(a) Compute the sample variance and sample standard deviation.

(b) Compute the range and coefficient of variation.

(c) Is it possible for the standard deviation of a data set to be larger than its variance? Explain.

Solution

(a) $s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 156.1233$ and $s = \sqrt{s^2} = 12.49$ (These values were obtained using Excel.)

Manual computation: [**:-)** Even if we use the "**Shortcut for Sample Variance**", it is still extremely laborious!]

$s^2 = \frac{1}{n-1}[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}] = \frac{1}{24}[636090 - \frac{(3976)^2}{25}] = \frac{1}{24}[636090 - \frac{(3976)^2}{25}] = \frac{3746.96}{24} = 156.1233$

	Weight	squared value
	164	26896
	148	21904
	137	18769
	157	24649
	173	29929
	156	24336
	177	31329
	172	29584
	169	28561
	165	27225
	145	21025
	168	28224
	163	26569
	162	26244
	174	30276
	152	23104
	156	24336
	168	28224
	154	23716
	151	22801
	174	30276
	146	21316
	134	17956
	140	19600
	171	29241
SUM	3976	636090

Please take note of the following comments:

- If we are lucky and $\frac{\sum x_i}{n}$ works out an integer (very rare in real life!), it is possible to compute an alternative column with the values $(x_i - \overline{x})^2$.

- $\sum x_i^2 \neq \left(\sum x_i\right)^2$ Repeat, **they are not equal!** Do not fall into the trap to think that it is a handy short cut!

  For example, if $x_1 = 2, \quad x_2 = 4$, and $x_3 = 5$

  $\Longrightarrow \sum x_i^2 = 4 + 16 + 25 = 45 \qquad$ but $\qquad \left(\sum x_i\right)^2 = (2 + 4 + 5)^2 = 121.$
- If you do not have Excel, the big problem is to either compute $\sum x_i^2$ or $\sum (x_i - \overline{x})^2$.
- $\sum (x_i - \overline{x})$ (the sum of the deviations that are not squared) will always add up to zero.

(b) (i) Range $= 177 - 134 = 43$

(ii) $cv = \dfrac{s}{\overline{x}} = \frac{12.49}{159.04} = 0.079$

(c) Yes. A standard deviation could be larger than its corresponding variance when the variance is between 0 and 1 (exclusive).

For example if $s^2 = 0.5 \implies s = \sqrt{0.5} = 0.707\,11 > s^2$.

3.4 Summary

I. A *measure of location* is a value used to describe the centre of a set of data.

A. The *arithmetic mean* is the most widely reported measure of location.

1. It is calculated by adding the values of the observations and dividing by the total number of observations.
   a. The formula for a population mean of ungrouped or raw data is

$$\mu = \frac{\Sigma x}{N}$$

   b. The formula for the mean of a sample is

$$\overline{X} = \frac{\Sigma x}{n}$$

2. The major characteristics of the arithmetic mean are as follows:
   a. At least the interval scale of measurement is required.
   b. All the data values are used in the calculation.
   c. A set of data has only one mean, that is, it is unique.
   d. The sum of the deviations from the mean equals 0.

B. The *median* is the value in the middle of a set of ordered data.

1. To find the median, sort the observations from smallest to largest and identify the middle value.
2. The major characteristics of the median are as follows:
   a. It requires at least the ordinal scale of measurement.

b. It is not influenced by extreme values.

c. Fifty percent of the observations are larger than the median.

d. It is unique to a set of data.

C. The *mode* is the value that occurs most often in a set of data.

1. The mode can be found for nominal-level data.
2. A set of data can have more than one mode.

II. The *dispersion* is the variation or spread in a set of data.

A. The *range* is the difference between the largest and the smallest value in a set of data.

1. The formula for the range is

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

2. The major characteristics of the range are as follows:
   a. Only two values are used in its calculation.
   b. It is influenced by extreme values.
   c. It is easy to compute and to understand.

B. The *variance* is the mean of the squared deviations from the arithmetic mean.

1. The formula for the population variance is

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

2. The formula for the sample variance is

$$s^2 = \frac{\Sigma \left(x - \overline{X}\right)^2}{n - 1}$$

3. The major characteristics of the variance are as follows:
   a. All observations are used in the calculation.
   b. It is not unduly influenced by extreme observations.
   c. The units are difficult to work with; they are the original units squared.

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

1. The major characteristics of the standard deviation are as follows:
   a. It is in the same units as the original data.
   b. It is the square root of the average squared distance from the mean.
   c. It cannot be negative.
   d. It is the most widely reported measure of dispersion.
2. The formula for the sample standard deviation is

$$s = \sqrt{\frac{\Sigma \left(x - \overline{X}\right)^2}{n - 1}}$$

3.5 Measures of relative standing and box plots

Always keep in mind that similar to a parameter being a descriptive measurement about a population, a statistic is a descriptive measurement about a sample. In the previous study unit we introduced you to numerical descriptive techniques and discussed measures of *central location* which are meaningful and appropriate for nominal and interval data. You learned how to compute the mean, the median, and the mode for a given data set and what each statistic tells you about the data. The geometric mean is used when we wish to find the average growth rate, or rate of change, in a variable over time.

We also introduced you to measures of *variability* and concluded that the range, the variance and the standard deviation are meaningful and appropriate only for interval data. Is there a possible measure of variability for ordinal data?

In this study unit we continue with descriptive measurements concerning a sample and we introduce you to **quantiles** which is a collective name for *measures of relative standing or measures of position*. Our aim is answer the question: Where does an observation stand relative to the sample it comes from? We also explore whether measures of relative standing can be used as measures of centrality and variability.

We conclude this study unit with **measures of linear relationship**. The added important concept of this section is the idea of **bivariate data**. In other words, we are not focusing on the variables when they are studied one at a time but we are interested in **studying them together,** i.e. the association or the relationship between *two* variables is our main focus. (Studying three or more variables jointly, i.e. their interactions and relations, is more complicated than studying two variables jointly and falls outside the scope of our first-year syllabus but it is touched upon in chapter 17 of the textbook.)

Do you recall that we emphasised that the type of variable dictates what we can or can't do? In the rest of the chapters to follow, you will come across different combinations of type of variables for bivariate data, for example both are categorical or both are interval or one is categorical and the other interval! In this study unit we merely introduce you to bivariate data on **two interval variables** whilst this problem is continued in chapter 16 which is covered in STA1502.

I. A *dot plot* shows the range of values on the horizontal axis and a dot is placed above each of the values.

   A. Dot plots report the details of each observation.

   B. They are useful for comparing two or more data sets.

II. *Measures of location* also describe the shape of a set of observations.

3.5.1 Quartiles and percentiles

The lower quartile (Q_1)

The *lower quartile* for set of measurement which has been ordered from the lowest to the largest value is the value that has 25% of the observations below or equal to it 75% of the observations above or equal to it. It is also referred to as a 25^{th} percentile.

The second quartile (Q_2)

The *second quartile* for a set of measurement is the median of the observations. For a set of ordered measurement, the second quartile is the value that has 50% of the observations below or equal to it and 50% percent of the observation above it. It is also referred to as the 50^{th} percentile.

The third Quartile (Q_3)

The *third quartile* for an ordered set of measurement is the value that has 75% of the observations below or equal it and 25% of the observations above or equal to it.

Location for a percentile is given by :

$L_p = (n+1)\frac{p}{100}, therefore$

$Q_1 = L_{25} = (n+1)\frac{25}{100}$

$Q_2 = L_{50} = (n+1)\frac{50}{100}$

$Q_3 = L_{75} = (n+1)\frac{75}{100}$

A. *Quartiles* divide a set of observations into four equal parts.
    1. Twenty-five percent of the observations are less than the first quartile, 50 percent are less than the second quartile, and 75 percent are less than the third quartile.
    2. The interquartile range is the difference between the third and the first quartile.

B. *Deciles* divide a set of observations into ten equal parts and *percentiles* into 100 equal parts.

C. A *box plot* is a graphic display of a set of data.
    1. A box is drawn enclosing the regions between the first and third quartiles.
        a. A line is drawn inside the box at the median value.
        b. Dotted line segments are drawn from the third quartile to the largest value to show the highest 25 percent of the values and from the first quartile to the smallest value to show the lowest 25 percent of the values.
    2. A box plot is based on five statistics: the maximum and minimum values, the first and third quartiles, and the median.

3.5.2 Interquartile range

The interquartile range measure the spread of the middle 50% of the observations, it given by

$$Q_3 - Q_1$$

A manually drawn box plot can look almost identical to a computer produced plot, but it would require much more computation! If you **do not have the software** to produce it, you need to do the following:

(i) Compute values for Q_1, Q_2 $(= median)$ and Q_3.

(ii) Identify the minimum and the maximum value in the observations.

3.5.3 Example

The following data represent the ages in years of a sample of **25** employees from a government department:

31, 43, 56, 23, 49, 42, 33, 61, 44, 28, 48, 38, 44, 35, 40, 64, 52, 42, 47, 39, 53, 27, 36, 35, and 20.

(a) Find the lower quartile of the ages.

(b) Find the upper quartile of the ages.

(c) Find the 60th percentile of the ages.

(d) Compute the interquartile range of the data, and interpret its meaning.

(e) Calculate the 4th decile of the data.

(f) Compute the 8th decile of the data.

(g) Calculate the 1st quantile.

(h) Construct a box plot for the ages and identify any outliers.

(i) What does the box plot tell you about the distribution of the data?

(j) Construct a relative frequency distribution for the data, using five class intervals and the value 20 as the lower limit of the first class.

(k) Construct a relative frequency histogram for the data.

(l) What does the histogram tell you about the distribution of the data?

Solution:

Note: The observations must first be arranged in ascending order.

20, 23, 27, 28, 31, 33, 35, 35, 36, 38, 39, 40, 42, 42, 43, 44, 44, 47, 48, 49, 52, 53, 56, 61, 64

(a) Location $L_{25} = (n+1)\frac{25}{100} = (26)\frac{1}{4} = 6.5$, and value of $Q_1 = 33 + 0.50(35 - 33) = 34$ years.

(b) Location $L_{75} = (n+1)\frac{75}{100} = (26)\frac{3}{4} = 19.5$, and value of $Q_3 = 48 + 0.50(49 - 48) = 48.50$ years.

(c) Location $L_{60} = (n+1)\frac{60}{100} = (26)\frac{3}{5} = 15.6$, and value of the 60^{th} percentile = $43 + 0.60(44 - 43) = 43.6$ years.

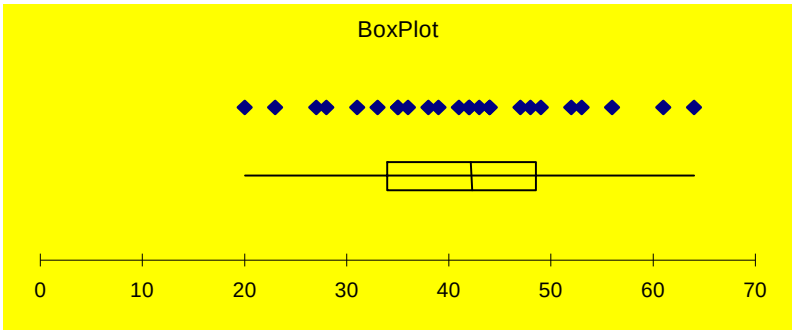
(d) IQR $= Q_3 - Q_1 = 48.5 - 34 = 14.5$. This means that 50% of the ages of employees from that government department are between 34 and 48.5 years.

(e) Location $L_{40} = (26)\frac{2}{5} = 10.4$, and value of 4th decile $= 38 + 0.40(39 - 38) = 38.40$

(f) Location $= L_{80} = (26)\frac{4}{5} = 20.8$, and value of 8th decile $= 49 + 0.80(52 - 49) = 51.4$

(g) Location $= L_{20} = (26)\frac{1}{5} = 5.2$, and value of 1st quantile $= 31 + 0.20(33 - 31) = 31.40$

(h) Box plot for the ages:

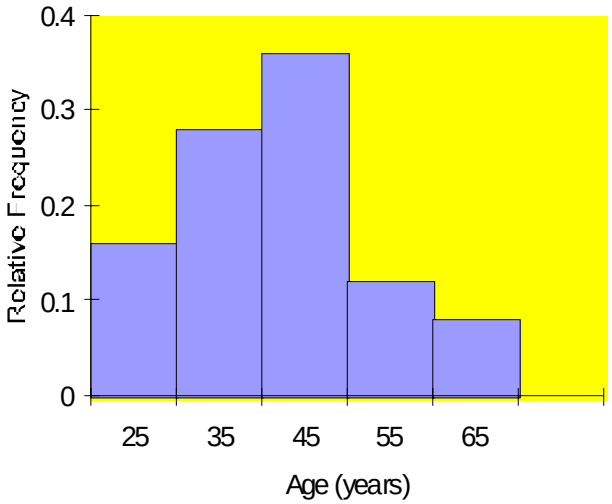


(i) There are no outliers. The box plot indicates symmetry.

(j) Relative frequency distribution for the data (using five class intervals and the value 20 as the lower limit of the first class)

Class Limits	Relative Frequency
20 up to 30	0.16
30 up to 40	0.28
40 up to 50	0.36
50 up to 60	0.12
60 up to 70	0.08
Total	1.00

(k) Relative frequency histogram for the data



(l) **The histogram incorrectly indicates positive skewness** as a result of the class limits chosen in (j). A histogram using a class width of 9 would have indicated symmetry.

3.6 Measures of Linear Relationship

This section introduces three numerical measures of the relationship: the covariance, the coefficient of correlation and the coefficient of determination.

3.6.1 Covariance

The **sample covariance** of the x-values and the y-values is defined as

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}.$$

A *scatter diagram* is a graphic tool to portray the relationship between two variables.

A. The *dependent variable* is scaled on the Y-axis and is the variable being estimated.

B. The *independent variable* is scaled on the X-axis and is the variable used as the estimator.

3.6.2 The coefficient of correlation

The *coefficient of correlation* measures the strength of the linear association between two variables.

A. Both variables must be at least the interval scale of measurement.

B. The coefficient of correlation can range from -1.00 up to 1.00.

C. If the correlation between two variables is 0, there is no association between them.

D. A value of 1.00 indicates perfect positive correlation, and -1.00 perfect negative correlation.

E. A positive sign means there is a direct relationship between the variables, and a negative sign means there is an inverse relationship.

F. It is designated by the letter r and found by the following equation:

$$r = \frac{\Sigma \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{(n-1)\, s_x s_y} = \frac{Sxy}{S_x S_y}$$

where

$$\begin{aligned} S_{xy} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{(n-1)} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right) \end{aligned}$$

$$\begin{aligned} s_x^2 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right) \end{aligned}$$

and

$$\begin{aligned} s_y^2 &= \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}{n-1} \\ & \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right) \end{aligned}$$

and $s_x = \sqrt{s_x^2}$ and $s_y = \sqrt{s_y^2}$.

3.6.3 The coefficient of determination

The coefficient of determination is the square of the coefficient of correlation.

The *coefficient of determination* is the fraction of the variation on one variable that is explained by the variation on the other variable.

A. It ranges from 0 to 1.

B. It is the square of the coefficient of correlation.

In *regression analysis* we estimate one variable based on another variable.

A. The variable being estimated is the dependent variable.

B. The variable used to make the estimate is the independent variable.

   1. The relationship between the variables must be linear.
   2. Both the independent and the dependent variables must be interval or ratio scale.
   3. The least squares criterion is used to determine the regression equation.

The *least squares regression line* is of the form $\hat{Y} = b_0 + b_1 X$.

A. $\hat{Y}$ is the estimated value of Y for a selected value of X.

B. a is the constant or intercept.

   1. It is the value of $\hat{Y}$ when $X = 0$.

2. a is computed using the following equation:

$$b_0 = \overline{Y} - b_1\overline{X}$$

C. b_1 is the slope of the fitted line.

1. It shows the amount of change in $\hat{Y}$ for a change of one unit in X.
2. A positive value for b_1 indicates a direct relationship between the two variables, and a negative value an inverse relationship.
3. The sign of b_1 and the sign of r, the coefficient of correlation, are always the same.
4. b_1 is computed using the following equation:

$$b_1 = r\left(\frac{s_y}{s_x}\right) = \frac{s_{xy}}{s_x^2}$$

D. X is the value of the independent variable.

3.6.4 Example

1. Given the following sample data:

x	420	610	625	500	400	450	550	650	480	565
y	2.80	3.60	3.75	3.00	2.50	2.70	3.50	3.90	2.95	3.30

(a) Calculate the covariance and the correlation coefficient.

(b) Comment on the relationship between x and y.

(c) Determine the least squares line.

(d) Draw the scatter diagram and plot the least squares line.

2. A sample of eight employees yield observations of variable x (years of experience) and variable y (number of projects involved in over the last two years) as shown below:

x	5	3	7	9	2	4	6	8
y	20	23	15	11	27	21	17	14

(a) Calculate the covariance between x and y.

(b) Calculate the coefficient of correlation, and comment on the relationship between x and y.

(c) Determine the least squares line, and use it to estimate the value of y for $x = 6$.

(d) Draw the scatter diagram and plot the least squares line.

Solutions:

1. $n = 10$ $\quad \sum x_i = 5250$ $\quad \sum x_i^2 = 2826250$ $\quad \overline{x} = 525$

$\sum x_i y_i = 17171.25$ $\quad \sum y_i = 32$ $\quad \sum y_i^2 = 104.455$ $\quad \overline{y} = 3.2$

(a)

$$\begin{aligned} cov(X,Y) &= \frac{\sum x_i y_i - n\overline{xy}}{n-1} \\ &= \frac{17171.25 - 10\,(525)\,(3.2)}{10-1} \\ &= \frac{371.25}{9} \\ &= 41.25 \end{aligned}$$

$$\begin{aligned} s_x &= \sqrt{\frac{\sum x_i^2 - n\overline{x}^2}{n-1}} \\ &= \sqrt{\frac{2826250 - 10\,(525)^2}{10-1}} \\ &= \sqrt{\frac{70000}{9}} \\ &= \sqrt{7777.7778} \end{aligned}$$

$$\begin{aligned} s_y &= \sqrt{\frac{\sum y_i^2 - n\overline{y}^2}{n-1}} \\ &= \sqrt{\frac{104.455 - 10\,(3.2)^2}{10-1}} \\ &= \sqrt{\frac{2.055}{9}} \\ &= \sqrt{0.2283} \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} r &= \frac{Cov\,(x,y)}{\sqrt{Var\,(x)\,Var\,(y)}} \\ &= \frac{41.25}{\sqrt{7777.7778}\sqrt{0.2283}} \\ &= 0.979 \end{aligned}$$

(b) There is a very strong positive linear relationship between X and Y.

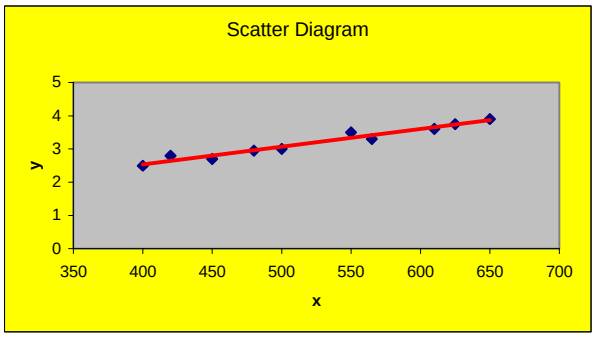
(c)

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{41.25}{7777.7778} \approx 0.0053$$

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{y} \\ &= 3.2 - (0.005304 \times 525) \\ &= 3.2 - 2.7844 \\ &= 0.4156 \end{aligned}$$

The regression line is $\hat{y} = 0.4156 + 0.0053x$

(d)



2. $n = 8$ $\sum x_i = 44$ $\sum x_i^2 = 284$ $\overline{x} = 5.5$

$\sum x_i y_i = 725$ $\sum y_i = 148$ $\sum y_i^2 = 2930$ $\overline{y} = 18.5$

(a) The covariance is

$$\begin{aligned} cov(X,Y) &= \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{n-1} \\ &= \frac{725 - 8\,(18.5)\,(5.5)}{8-1} \\ &= \frac{-89}{7} \\ &= -12.7143 \end{aligned}$$

(b)

$$\begin{aligned} s_x &= \sqrt{\frac{\sum x_i^2 - n\overline{x}^2}{n-1}} \\ &= \sqrt{\frac{284 - 8\,(5.5)^2}{8-1}} \\ &= \sqrt{\frac{42}{7}} \\ &= \sqrt{6} \end{aligned}$$

$$\begin{aligned} s_y &= \sqrt{\frac{\sum y_i^2 - n\bar{y}^2}{n-1}} \\ &= \sqrt{\frac{2930 - 8\,(18.5)^2}{8-1}} \\ &= \sqrt{\frac{192}{7}} \\ &= \sqrt{27.4286} \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} r &= \frac{Cov\,(x,y)}{\sqrt{Var\,(x)\,Var\,(y)}} \\ &= \frac{-12.7143}{\sqrt{6}\sqrt{27.4286}} \\ &= -0.9911 \end{aligned}$$

There is a very strong (almost perfect) negative linear relationship between X and Y.

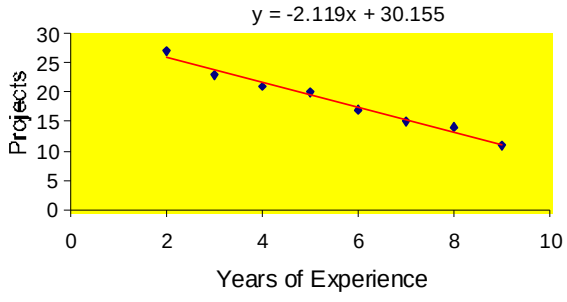
(c)

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{-12.7143}{6} \approx -2.119$$

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{y} \\ &= 18.5 - (-2.11905 \times 5.5) \\ &= 18.5 + 11.6548 \\ &= 30.155 \end{aligned}$$

$\widehat{y} = 30.155 - 2.119x$. When $x = 6$. $\hat{y} = 30.155 - 2.119(6) = 17.441$

(d) Scatter diagram plotted with the least squares line



Key terms

interquartile range

P^{th} percentile

box plot

sample covariance

sample correlation coefficient

linear relationship

linear equation

References

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 4

DATA COLLECTION AND SAMPLING

4.1 Learning outcomes

After completing this unit, you should be able to

1. define and explain data collection methods
2. define the terms: target population and sampled population
3. describe the advantages and disadvantages of telephone interview, personal interview and self administered survey
4. construct a questionnaire design
5. define and explain sampling plan
6. define and explain sampling and nonsampling errors

4.2 Introduction

In the previous study units you learned about the appropriate graphical and tabular techniques for sample data. This means you had a specific fixed data set (sample) which you tried to describe or organise. This entailed that you *graphically displayed* the data to show the shape of the sample and *computed descriptive measures* to summarise the sample. This mainly included *measures of locality* (mean, mode etc.) and *measures of dispersion* (range, standard deviation, etc.). All these techniques and methods were applied to real life (or empirical) examples. The emphasis was on the techniques as such and although we sensitised you to the fact that nominal and interval data must be treated differently, **we did not embroider on how the data was obtained**.

In this study unit you will learn that there are different ways to collect data as well as good and bad ways to obtain a sample. Always keep in mind that the overall objective of Statistics is to draw conclusions about a population based on the limited information contained in a sample. (In the module STA1502 you will learn how to link the information from a sample to a population.)

So, the better the sampling, the better the results that you produce! How will you collect data or design a sampling scheme to obtain optimum information?

4.3 Methods of collecting data and sampling

Sampling should be a component of any method of collecting data, unless it is possible to conduct a census. The separate heading of "Methods of Collecting Data" is to underline the **different ways** in which you could gather data. Each separate method will most probably entail sampling as well – indicating that only **"a part"** of all the possible observations of a specific variable will be obtained. So, whether you obtain data by direct observation, experiments or surveys you need to think carefully about the possible problems evolving from the sampling process. Especially when statistical inference is the final aim, you must always ask yourself whether the **target population** and the **sampled population** are the same.

The data collection method is a method which the researcher uses to gather the required information from the respondents. Statistical studies may be classified in three categories, depending on the way in which data are obtained;

1. direct observation,
2. experiments,
3. and surveys.

Each one of these methods has advantages and disadvantages. The most appropriate method depends to a large extent on the objectives of the specific survey, the time and money available, the sensitivity of the questions and the characteristics of the target population. The researcher should take all the advantages and disadvantages of the various techniques into consideration and choose the method which best suits the specific study.

4.3.1 Direct observation

In direct observation studies, a sample is obtained from the population and its characteristics observed. Data can be collected by directly observing the number of respondents. The direct observation methods can be simplified in three steps: listen, observe and record all the required information. An observational studies, if properly planned and executed, may usually be much more informative, although there are many limitations.

One of the main limitation is that it is difficult to produce reliable and useful information using a direct observation method of data collection. One of the advantages of the direct observation method is that the approach is not costly.

4.3.2 Experiments

Experiments are performed with the specific aim of determining the effect of treatment on members of the population. Two or more treatments can be compared. Sometimes a treatment is tested against a "control" which can be:

- no treatment at all,
- a standard treatment,
- comparison of different dosages, etc.

4.3.3 Examples of experiments

- The effect of different types of packing on the sales of a product.
- The comparison of the effect of different treatment for hypertension on the blood pressure of the patients.

4.3.4 Surveys

A survey is an investigation about the characteristic of a given population by means of collecting data from a sample of the population and estimating the characteristics. It is one of the most used technique of collecting data. The response rate plays a crucial role on the survey method.

Surveys can be conducted through personal interviews, telephone interview, self administered survey and questionnaire design.

Personal interview

The researcher collect data by asking prepared questions to respondent. Generally, the respondent is asked questions by the researcher who records the responses.

Advantages of the personal interview

1. High response rate
2. More clarification on the questions

Disadvantages:

1. Costly
2. Time consuming

Telephone interview

The interviewer solicits informations through a telephone conversation.

Advantages

1. Less expensive
2. Large sample can be obtained in a shorter period of time

Disadvantages

No verbal response can be observed

Self administered survey

Questions are mailed, posted or delivered to a sample.

Advantages

Cheaper

Can be used for a large sample

Disadvantages

Low response rate

4.3.5 Questionnaire design

The questionnaire is one of the main source of data collection. As such, the questionnaire design is very crucial in order to produce reliable data analysis and achieve the objectives of the study. Some characteristics of a good questionnaire are

1. Questions must be short ,
2. Questions must be clear,
3. The flow of ideas is very important,
4. Avoid leading , difficult and negatives questions,
5. Questionnaire must contribute to the aim of the study,
6. Time management,
7. Design both close and open-ended questions.

4.4 Sampling

Examining a population can be costly and time consuming. Inferential statistics permit us to draw conclusions or make inferences about a population. The idea is "to identify a population, breakdown the population in order to obtain a sample, analyse a sample and extend the result of the sample to the population".

Target population: population about which a research want to make inference.

A *sample* is fraction or a subset of the population. The sample must be representative of the population.

Self- selected samples are almost biased because the elements, individuals who are part of the sample are more interested in the issue that are other members of the populations.

Sampling plans

There are several probability sampling methods including the simple random, the stratified and the cluster sampling plans. In an unbiased sample all members of the population have a chance of being selected for the sample.

Simple random sampling plan

A *simple random sample* is a sample selected such that every possible, with the same number of observations has equal chance of being selected.

Stratified sampling plan

A *stratified random sample* is obtained by dividing the population in into several groups called strata, and then drawing simple random samples form each stratum.

A cluster sampling

A *cluster sample* is a simple random sample in which each sampling unit is a collection /groups/cluster of elements.

4.5 Sampling and non sampling errors

The population is different form the sample, two majors errors arise when a sample of observations is taken from the population in order to make statement about a population based on the characteristics of the sample: sampling error and non-sampling error.

Sampling error is the difference between the sample and the population.

Non sampling error result from the mistakes made in the process of obtaining data or data being selected or chosen without following a proper process such as errors in data acquisition, nonresponse error and selection bias.

4.6 Example

An organisational psychologist wants to randomly split up 20 personnel of a small firm into two teams of 10 personnel each to compete against each other in a team building endeavour. **How would the psychologist randomly assign the 20 personnel to the two teams?** Describe a simple random sampling plan for this experiment.

……

……

……

Figure 5.2: Table 1

Column 1	Column 2	Column 3
0.269082	26.90817	27
0.088626	8.862575	9
0.394574	39.45738	40
0.590594	59.05942	60
0.937895	93.78948	94
0.759453	75.94531	76
0.196326	19.63256	20
0.205756	20.57558	21
0.642384	64.23841	65
0.162145	16.21448	17
0.941496	94.1496	95
0.60448	60.44801	61
0.061342	6.13422	7
0.507035	50.70345	51
0.759636	75.96362	76
0.617115	61.71148	62
0.546068	54.60677	55
0.713462	71.34617	72
0.473189	47.31895	48
0.534745	53.47453	54
0.448469	44.84695	45
0.550218	55.02182	56
0.801599	80.15992	81
0.621387	62.13874	63
0.209265	20.92654	21
0.496536	49.65361	50
0.403394	40.33937	41
0.112827	11.28269	12
0.986908	98.69076	99
0.979949	97.99493	98
0.78753	78.75301	79
0.607776	60.77761	61
0.108829	10.8829	11
0.147374	14.73739	15
0.624958	62.4958	63
0.257881	25.78814	26
0.126408	12.64077	13
0.8258	82.58003	83
0.491958	49.19584	50
0.662954	66.29536	67
0.754387	75.4387	76
0.828455	82.84555	83
0.576281	57.6281	58
0.877377	87.73766	88
0.476547	47.65465	48
0.263009	26.30085	27
0.377514	37.7514	38
0.434156	43.41563	44
0.098544	9.854427	10
0.99823	99.82299	100

Solution

Usually the "most primitive" but simple random sampling plan that pops up in my mind is to assign a number to each member of the firm and simply write the twenty numbers 1 to 20 on twenty pieces of paper and place them in a hat (or other container). Draw ten numbers and assign these numbers to the first team and the remaining ten numbers in the hat will be assigned to the second team. Easy as pie! But some people are uncomfortable with pieces of paper and hats and prefer a proper scientific method where they make use of **Random Numbers**. If we want to force you into a specific method we could phrase the question as: *How would the psychologist assign the 20 personnel to the two teams by making use of random numbers?*

Using Random Numbers

If we use Table 2 of the random numbers (given in Figure 5.2 and generated by Excel), the idea is to write down the first ten numbers (in value below twenty) and consider them as the first team. But what happens? There are *only nine values* between one and twenty! They are

9; 20; 17; 7; 12; 11; 15; 13; 10.

What do we do now?

(a) We can either *repeat the Excel process* but this time we generate more two-digit numbers (say 100 values) to make sure there will be 10 to pick, or....

(b) we try to make better use of the available 50 numbers. How?

OPTIONAL INFORMATION:

You can skip this section if you are not into serious random sampling!

Making better use of the random numbers:

Refer to Example 4.6 above. You noticed that we could only draw nine numbers but we needed ten! Is there a clever plan to make better use of the generated random numbers? If we *use only* the digits **01, 02, ..., 20** it will force us to work through a *very long list* of random numbers. (We have to ignore 21, 22, up to 99.)

We could use the following scheme: (Note that we make a list of every possible two-digit number from 01 to 99 as well as 00.)

Number of Member	Random numbers
1	01, 21, 41, 61, 81
2	02, 22, 42, 62, 82
3	03, 23, 43, 63, 83
4	04, 24, 44, 64, 84
5	05, 25, 45, 65, 85
6	06, 26, 46, 66, 86
7	07, 27, 47, 67, 87
8	08, 28, 48, 68, 88
9	09, 29, 49, 69, 89
10	10, 30, 50, 70, 90
11	11, 31, 51, 71, 91
12	12, 32, 52, 72, 92
13	13, 33, 53, 73, 93
14	14, 34, 54, 74, 94
15	15, 35, 55, 75, 95
16	16, 36, 56, 76, 96
17	17, 37, 57, 77, 97
18	18, 38, 58, 78, 98
19	19, 39, 59, 79, 99
20	20, 40, 60, 80, 00

The final step is to take all the two-digit random numbers from Table 2 and "connect each number with a member" of which we then assign the first ten members to the first team.

Random number:	27	9	40	60	94	76	20	21	65	17	95	61	7	51
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Associated member:	07	09	20	20	14	16	20	01	05	17	15	01	07	11 etc

Note that **20** is repeated in the fourth and the seventh associated members which we ignore, **01** is repeated in the twelfth associated member which we ignore and **07** is repeated in the thirteenth associated member which we also ignore.

Thus the first ten randomly chosen personnel will be **07 09 20 14 16 01 05 17 15 11** which will all be assigned to the first team.

Example

Select the correct option.

Question 1

A pharmaceutical company interested in measuring how often physicians prescribe a certain drug, has selected a simple random sample from each of two groups: General Practitioners and Psychiatrists. What is this type of sampling called?

(a) simple random sampling

(b) cluster sampling

(c) stratified random sampling

(d) biased sampling

(e) none of the above

Question 2

When every possible sample with the same number of observations is equally likely to be chosen, the selected sample is called

(a) a simple random sample

(b) a stratified sample

(c) a cluster sample

(d) a biased sample

(e) none of the above

Solution

Question 1

ANSWER: option (c).

General Practitioners and Psychiatrists are mutually exclusive groups of doctors and because significantly different feedback from each group may be expected, stratification is the better option.

Question 2

ANSWER: option (a).

This is a different interpretation of the definition of simple random sampling. We are used to saying "each observation is equally likely to be chosen".

4.7 Study Unit 4: Summary

In an unbiased or probability sample all members of the population have a chance of being selected for the sample. There are several *probability sampling methods*.

A. In a *simple random sample* all members of the population have the same chance of being selected for the sample.

B. In a *systematic sample* a random starting point is selected, and then every *k*th item thereafter is selected for the sample.

C. In a *stratified sample* the population is divided into several groups, called strata, and then a random sample is selected from each stratum.

D. In *cluster sampling* the population is divided into primary units, then samples are drawn from the primary units.

Reference

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 5

BASIC PROBABILITY

5.1 Learning outcomes

After completing this unit, you should be able to

1. define the term probability, random experiment and its outcomes and sample space,
2. define the concept of exhaustive and mutually exclusive events,
3. explain and apply the properties of probabilities,
4. discuss the approaches to assign probabilities,
5. describe the difference between joint, marginal and conditional probability,
6. apply the complement, addition, multiplication rules of probabilities,
7. draw a probability tree

5.2 Introduction

At this stage you should be familiar with the different ways in which data can be presented and eager to learn more about the abstract concept of probability.

Do you realise how many things in life can be determined with certainty? An example would be the change you receive if you buy an item marked $R42.60$ and you pay with $R50$ – the change will be $R7.40$. Call paying for the item an *event* and the change you get the *outcome of the event.* It is possible to *predict* the single outcome of the event because the change will always be $R7.40$ if you pay with $R50$. Now, think in terms of an event with more than one possible outcome. Let me help you: think in terms of the examination results for STA1501 at the end of this year. The results for each student can be one of three possibilities.

Can you list them? They are {Pass; Fail with the option of a supplementary examination; Fail}. This list of possible outcomes is called the *sample space*. The lecturer knows that the results for each student can only be one of these outcomes, but it is not possible to predict for a randomly chosen student what happened in the examination. Aren't you glad that results are based on knowledge and not simply randomly allocated?

Think in terms of a competition where the names of ten finalists are placed in a hat and one name is randomly drawn from the hat. The person who matches the name will be the winner. Although one cannot predict the name of the winner, you know that it will be one of the ten finalists. Suppose you were a finalist, can you determine the probability or chance that you would be the winner? Of course - you have a *one in ten chance* of being the winner. How did you calculate that? You know that there were *ten* finalists, so there are ten names in the hat, each with an equal chance of one in ten to be the winner. This uncertainty in games and selections based on chance excites the average person and forms the basis of the principles of **probability** that we are going to discuss.

Probability has become an integrated part of everyday life and it is freely used in terms of lotteries, government planning, financial predictions, medical test results, etc. Have you noticed how manufacturers use chance to sell their products? Two stores selling colour laser printers for the same price place their advertisements simultaneously; the one making an offer that one lucky customer can win a TV and the other without this offer. Where do you think most of the people will buy? The interesting thing is that you cannot determine the probability of being the winner of the TV, because you do not know how many printers were sold at that shop. Unfortunately, the more customers, the smaller the probability of being the lucky one! The words *chance*, *luck* and maybe even *random* are rather commonly used in modern speech, though often more as conceptions of coincidence and/or superstition.

The English word "probable" refers to "something that is relatively likely, but not certain". Do you know that a statistician who specializes in probability is called a **probabilist**? Probability theory, as applied in Statistics, involves more, much more than working out the chance you have of winning the lottery! Probability Theory is a theoretical part of Statistics involving complex mathematics and abstract reasoning. In this first-year course we are only going to address the basic laws of probability and look into some elementary probability calculations.

5.3 Basis of probability

How do you feel about definitions? Make sure that you reflect on the *facts* contained in each definition. Without knowledge of concepts such as Assigning Probabilities to Events, Joint Probability, Marginal Probability, Conditional Probability, Probability Trees, etc. you cannot become a "master of probability"!

We would like to make some important comments about the two requirements (or rules) of probability:

A probability of an outcome of a random experiment. It is the chance, likelihood of the outcome occurring. In other words, a probability is a chance that an events will happen. A probability is always expressed as a percentage, a fraction or a decimal. The value of a probability is always between 0 and 1.

Think about , what is the probability that a student will pass the exam? What are the chances that there will be rain today?

5.3.1 A random experiment

A *random experiment* is a process or an action which generates one or multiple outcomes.

5.3.2 Examples

1. Flipping a coin is an action which lead to a head or a tail.
2. Tossing a die is an action which generates 1, 2,3,4,5 or 6.

5.3.3 A sample space

A *sample space* is a list of all outcomes of a random experiment and is denoted by S.

5.3.4 Examples

1. If you flip a coin once, there are 2 possible outcomes namely the head and the tail.

   The sample space: $S = \{head, tail\}$
2. If a die is tossed once, there a six possible outcomes;1,2,3,4,5,6.

   The sample space of the toss of a fair die is : $S = \{1, 2, 3, 4, 5, 6\}$

5.4 Requirements of probability

- If you *see* $0 \leq P(O_i) \leq 1$ what do you say in your mind?

  You should say: *The probability* P, *on every possible outcome* O_i, *of a random experiment, namely* $P(O_i)$, *must always be any number from* 0 *to* 1. Yes, I do not say "between" 0 and 1, because both 0 and 1 are included as possibilities. If the probability of a particular outcome is equal to 0 that outcome is *impossible,* while a probability outcome of 1 implies that it is a *certain* outcome.

- Probability is usually expressed as a fraction, e.g. $\frac{3}{10}$ (or in the decimal form 0.3).
- If you *see* $\sum_{i=1}^{k} P(O_i) = 1$ what do say in your mind?

  You should say: *If the probabilities on all possible outcomes in the sample space of a random experiment are added, the answer must be* 1.

5.5 Approaches to assigning probabilities

The three approaches to assigning probability are:

- Classical approach
- Relative frequency approach
- Subjective approach

5.5.1 Classical Approach

$$P(E) = \frac{\text{Number of successes}}{\text{Total number of outcomes}}$$

5.5.2 Example

In rolling a fair die once, the probability of the event obtaining an odd number (1, 3, 5) is 3 out of (1, 2, 3, 4, 5, 6) .

Number of successes is 3.

Total number of outcomes is 6, and the probability is $\frac{3}{6}$.

5.5.3 Relative frequency approach

$$P(E) = \frac{\text{Number of times the events occurred}}{\text{Total number of observations}}$$

5.5.4 Example

If you toss a coin 15 times and get 3 heads, the relative frequency approach gives: $\frac{3}{15}$.

5.5.5 Subjective probability

Probability is assigned based on intuition, past experience, educated guess or opinion and give a numerical estimate of the likelihood that a particular outcome will occur.

5.5.6 Example

Given a student's assignment marks, a lecturer may feel that the student has a 50% change to pass the module.

5.6 Defining events

A *simple event* is an individual outcome of a sample space.

An event is a list or set of one or more simple events in a sample space.

The probability of an event is the total or sum of the probabilities of the simple events that constitute the event.

5.7 Joint, marginal and conditional probability

5.7.1 Intersection

The intersection of events A and B is the event that occurs when both A and B occur. It is denoted by: A and B. The probability of the intersection is also called the joint probability.

5.7.2 Examples

1. The names and gender (M for male and F for female) of the ten finalists in a competition are as follows:

Mokiti (M)	Jerome (M)
Dimakatso (F)	Marius (M)
Denise (F)	Peter (M)
Solomon (M)	Margaret (F)
Kai (M)	Nkosinathi (M)

   The winner will be determined by writing these names on individual pieces of paper, placing them all in a hat and then randomly drawing one piece of paper. The name on the paper will be the winner.

   Think carefully: May I ask the following questions? If your answer is "no", give reasons.

1. Do you think that this was a beauty competition?
2. Find the average of these ten names.
3. Arrange the names from the shortest to the longest and determine the *median name*.

Real questions:

The sample space for this competition is a list of the ten names. Assume that each person has an equal chance of being the winner, i.e., each simple event has the same probability. Find the probability on the following events:

1. The winner is male.
2. The winner's name has six letters.
3. The winner's name has four letters.
4. The name of the winner starts with the letter M.
5. Nkosinathi is the winner.

2. Choose the correct answer:

If A and B are mutually exclusive events with $P(A) = 0.70$, then $P(A \text{ and } B)$

1. can be any value between 0 and 1
2. can be any value between 0 and 0.70
3. cannot be larger than 0.30
4. is zero
5. is exactly 0.3

Solutions:

1. The silly questions:

1. Such a question has nothing to do with statistics! Common sense would say that beauty contests are more associated with females and in any case should not be across the two genders!
2. The "average" of *names* cannot be determined. An average only makes sense if you have *number*s (i.e. numerical data). If you feel insecure about my statement, please go back to p.90 and read the *Introduction.*
3. The median is the middle value in an ordered list and names cannot be ordered (numerically). Of course, if the question asked you to determine the median *number of letters* in the names, it would

be totally different.

Name	No. of letters in name
Kai	3
Peter	5
Mokiti	6
Jerome	6
Marius	6
Denise	6
Solomon	7
Margaret	8
Dimakatso	9
Nkosinathi	10

Above is now an ordered list of the number of letters in the names, with Kai being the shortest name with only 3 letters and Nkosinathi with 10 letters the longest name. What is the median number of letters? Because there are an even number of names, it is the value between the 5th and 6th name, i.e. between Marius and Denise. These names both have 6 letters and therefore the median number of letters in these ten names is 6.

Be sure to pick up such nonsense as discussed in these three questions in presentations by incompetent persons. Keep in mind that nonsense is not always in such a blatant form as mine!

The real questions and their probabilities:

1. Seven of the ten names are those of males (Kai, Peter, Mokiti, Jerome, Marius, Solomon and Nkosinathi). Therefore $P(\text{male}) = \frac{7}{10}$ (or 0.7).
2. Four of the ten names have six letters (Mokiti, Jerome, Marius and Denise). Therefore $P(\text{six letters in name}) = \frac{4}{10}$ (or 0.4).
3. There is no name with four letters in the group. In such a case there would be no winner. We write $P(\text{four letters in name}) = \frac{0}{10} = 0$. Remember? This is an impossible outcome.
4. Three names start with the letter M (**M**okiti, **M**arius and **M**argaret), so $P(\text{name starts with an M}) = \frac{3}{10}$ (or 0.3).
5. Nkosinathi has the same chance as any of the other nine to win, i.e. $P(\text{Nkosinathi wins}) = \frac{1}{10}$ (or 0.1).

2. The correct option is (3). Why?

- Option (1) is true in general for all probabilities, but with the information that A and B are mutually exclusive outcomes in a particular event, this cannot be true. E.g. if $P(B) = 0.8$ (which is a value between 0 and 1), then A and B cannot be mutually exclusive.
- Option (2) is wrong for the same reason, because $P(B)$ may not be between 0.7 and 0.3.
- Option (4) is just nonsense!
- Careful with this one – "exactly 0.3" would be true only if A and B were the only two possible outcomes of the event and such information was not given.

5.8 Sophisticated methods and rules in probability theory

A reminder of notation you will come across in the text and in certain questions:

Symbol	Meaning of the symbol	
$\geq$	greater than or equal to	at least
$\leq$	less than or equal to	at most
$>$	greater than	more than
$<$	smaller than	less than

Mathematical set notation supplies the building blocks for Probability Theory, as illustrated clearly and in the principles applied in joint, marginal and conditional probability. Although you can learn the meaning of these statistical concepts without acknowledging the mathematical principles underlying it, many students find explanations very interesting and quite logical if the relevant set theory is included in a very down-to-earth way. You will see that I have given tiny bits of mathematical notation and information for those who are interested (they are clearly marked). Remember that you will not be tested on this!

5.8.1 Joint probability

Suppose that there are certain events and each event has a specific number of possible outcomes. In Activity 5.3.2 you saw that the probabilities on such selected events could be determined (we counted the number of favourable outcomes to determine the relevant probabilities). Suppose that, instead of considering one event at a time, you want to join events, then you will need to know the details of the particular combination. Should your interest be in determining whether any events have outcomes that "overlap", then you have to study the *intersection* of the events that you are considering. You are *joining* the events in such a way that the different outcomes in the *joint event* all realised in each and every original event. The names given in **Example 5.7.2**. Consider the two events:

(2) The winner's name has six letters. Call this event A.

(4) The name of the winner starts with the letter M. Call this event B.

Our interest is now in the joint event of *A and B,* such that each outcome of this joint event was an outcome in event A and also an outcome in event B. Can you list the names that were outcomes in A as well as in B?

The outcomes of A: Mokiti, Jerome, Marius and Denise.

The outcomes of B: **M**okiti, **M**arius and **M**argaret.

The names that occur in both outcomes are Mokiti and Marius.

This implies that the event (A and B) = {Marius, Mokiti} .

See if you can determine the probability of this joint event A and B.

Two of the *ten* names are in A as well as in B. Do you agree that $P(A \text{ and } B) = \frac{2}{10}$? Is this easy, or is it easy?

In set theory a *joint event* is called an *intersection.* Link this to the word "intersection" as used for the place where two roads cross: if you are in the *intersection* you could have come from any of the roads crossing there. The joint probability of the events *A and B* implies that an intersection has to be determined.

For the curious student:

The mathematical symbol for the word *intersection* looks like this: $\cap$.

The notation $P(A \cap B)$ implies "The probability on the intersection of the events A and B".

The symbol $\cap$ therefore implies a joint probability.

5.8.2 Example

Why are some mutual fund managers more successful than others? One possible factor is the university where the manager earned his or her master of business administration (MBA). Suppose that a potential investor examined the relationship between how well the mutual fund performs and where the fund manager earned his or her MBA. After the analysis table of joint probabilities, was developed. Analyze these probabilities and interpret the results.

Table Joint Probabilities

	Mutual Fund out perform market (B_1)	Mutual Fund does not outperform market (B_2)
Top-20 MBA program (A1)	$0.11 = P$ (A_1 and B_1)	$0.29 =$ (A_1 and B_2)
Notop-20 MBA program (A2)	$0.06 = P$ (A_2 and B_1)	$0.54 =$ (A_2 and B_2)

Source: Keller. G 2018

1. The joint probability of A_1 and $B_1 = 0.11$
2. The joint probability of A_2 and $B_1 = 0.06$
3. The joint probability of A_1 and $B_2 = 0.29$
4. The joint probability of A_2 and $B_2 = 0.54$

5. The joint probability of $(A_1) = 0.11 + 0.29 = 0.40$

6. The joint probability of $(A_2) = 0.06 + 0.54 = 0.60$

7. The joint probability of $(B_1) = 0.11 + 0.06 = 0.17$

8. The joint probability of $(B_2) = 0.29 + 0.54 = 0.83$

5.8.3 Marginal probability

Marginal probabilities are computed by adding across or down columns, are so named because they calculated in the margins of the table.

- Using the first row on the table gives the following marginal probabilities

$$P\ (A_1 \text{ and } B_1) + P\ (A_1 \text{ and } B_2) = 0.11 + 0.29 = 0.40$$

- Using the second row on the table gives the following marginal probabilities

$$P\ (A_2 \text{ and } B_1) + P\ (A_2 \text{ and } B_2) = 0.06 + 0.54 = 0.60$$

- Adding down the first column produces the following marginal probabilities

$$P\ (A_1 \text{ and } B_1) + P\ (A_2 \text{ and } B_1) = 0.11 + 0.06 = 0.17$$

- Adding down the second column produces the following marginal probabilities

$$P\ (A_1 \text{ and } B_2) + P\ (A_2 \text{ and } B_2) = 0.29 + 0.54 = 0.83$$

Tables (rows and columns) are used to give a more systematic presentation of information. Every cell in a table belongs to a row as well as to a column. In terms of the previous discussion one can say that every value in a cell falls in the *intersection* of a specific row and specific column and the value written there in the cell is the *joint probability* of the events indicated by that particular row and column. Think in terms of a two-by-two table, where you have two rows and two columns (as Table 6.2 in Keller). For example: the *second row* represents "Not top-20 MBA program" and the *first column* represents "Mutual Fund Outperforms Market". The *joint probability* of these two events represents a mutual fund outperforming the market being managed by a person who did not graduate from a top-20 MBA program. The numerical value of this joint probability is equal to 0.06 (the value written where the second row and first column meet). The great use of tabulation is that you can add the values in any row or in any column to give the so-called *marginal totals*. In this example the values in the table represent probabilities, so we call these totals *marginal probabilities.*

5.8.4 Conditional probability

The *conditional probability* is the probability of an event A given information about the occurrence of the event B.

The probability of event A given event B is given by:

$$P\left(A/B\right) = \frac{P\left(A \text{ and } B\right)}{P\left(B\right)}$$

where $P\left(B\right) > 0$.

The probability of event B given A is given by:

$$P\left(B/A\right) = \frac{P\left(A \text{ and } B\right)}{P\left(A\right)}$$

where $P\left(A\right) > 0$.

5.8.5 Example

You will see in questions how convenient it is to use the marginal probabilities of a table of probabilities when you have to determine a conditional probability. Use the information from Table 6.2 and see if you can *read off* the probability $P\left(B_2 \mid A_1\right)$.

How? Find the total of row A_1, which is 0.40 (the marginal probability of the event A_1). This is the denominator of the fraction (the number in the lower position of the fraction). In the numerator (upper position of the fraction) write the joint probability on A_1 and B_2 which is 0.29

$$P\left(B_2 \mid A_1\right) = \frac{0.29}{0.40} = 0.725$$

The probability that the Mutual Fund does not outperform the market, *if you know that* the Fund Manager graduated from a top-20 MBA program, namely $P\left(B_2 \mid A_1\right)$, is equal to 0.725.

Your next task is to interpret the answer above in words, as Keller did for $P\left(A_1 \mid B_2\right)$.

See if you agree with me:

Thus, 72.5% of managers who graduated from top-20 MBA programs managed funds that were not able to outperform the market.

Why did I ask you to find this particular probability? I want you to see that our answer for $P\left(B_2 \mid A_1\right)$ is not

the same as the book's answer for $P(A_1 \mid B_2)$. Therefore

$$\begin{aligned} & P(A_1 \mid B_2) && \neq P(B_2 \mid A_1) \\ \text{because} \quad & 0.3494 && \neq 0.725 \end{aligned}$$

Always remember that you cannot change the events around in conditional probability!

5.8.6 Independent events

Two events A and B are independent if $P(A/B) = P(A)$ or $P(B/A) = P(B)$

5.8.7 Example

What is the meaning of *independence*? If you know the answer, the concept "*independent events*" will be quite understandable.

- Your future employment is *independent* of the colour of your hair, but it is dependent on your studies.
- The first letter of your name should have nothing to do with the number of letters in your name, therefore the first letter and the number of letters in your name are *independent*.

Statisticians use conditional probability to prove that two events are independent as follows:

$$\text{If } \; P(A \mid B) = P(A), \text{ then the events } A \text{ and } B \text{ are independent.}$$

Read the statement above as follows: If the probability of an event A, given that another event B had taken place, is the same as the direct probability of the event A, then the event B has no effect on the occurrence of A. The probability of event A is independent of whether event B took place or not.

A logical conclusion for two independent events A and B :

$$\begin{aligned} \text{If } \; P(A \mid B) &= P(A), \text{ it implies that} \\ P(B \mid A) &= P(B). \end{aligned}$$

Remember this for independent events!

5.8.8 The union of events

Union of events

The union of events A and B is the event that occurs when either A or B or both occur. It is defined by: A or B. The probability of the event A or B is given by:

$$P\,(A \text{ or } B) = P\,(A) + P\,(B) - P\,(A \text{ and } B)$$

5.8.9 Example

If two events are united, you put them together and the concept is totally different from intersecting two events. Have you heard that marriage is considered the *union between two persons*? They are "put together" and in this union you find all the characteristics of the one and also all the characteristics of the other. If there are any overlapping or common characteristics between the two it is fine, but the union does not depend on any overlapping.

Let us return to the information given in Activity 1, question 2, and use the same events A and B as in the discussion of joint (or intersecting) events. This time consider the *union* of A and B. In ordinary language we simply say "A or B":

A: Mokiti, Jerome, Marius and Denise.

B: **M**okiti, **M**arius and **M**argaret.

For the *union* we write down all the elements in *A and also all the elements* in B, i.e.

A union B = {Mokiti, Jerome, Marius, Denise, **M**okiti, **M**arius and **M**argaret} .

Don't you find it funny that two names are listed twice? Mokiti and Marius are written twice. Common sense tells us that we should not list names twice (and it is a rule in set theory not to list elements in a set twice). So, simply remove one of each double listing, i.e.

A union B = {Mokiti, Jerome, Marius, Denise and **M**argaret} .

Isn't this a nicer presentation? Mokiti and Marius are listed once, even though they belong to both sets. Of course you realise that these "double" elements are the elements of the intersection!

For the curious student:

The mathematical symbol for the word *union* is $\cup$.

The notation $P(A \cup B)$ implies "The probability on the union of the events A and B".

The union of the two events, written as $A \cup B$, implies that all the elements of both sets are put together.

Study this beautiful "flow" between the different concepts and the rules:

Consider events A and B. You have to

○ understand the meaning of *joint probability*	A and B (intersection)
○ be able to define *conditional probability*	$P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$
○ understand the meaning of *independent events* and its effect on *conditional probability*	$P(A \mid B) = P(A)$
○ manipulate conditional probability to formulate the *multiplication rule*	$P(A \text{ and } B) = P(A \mid B) \cdot P(B)$
○ know the influence of independence on the *multiplication rule*	$P(A \text{ and } B) = P(A) \cdot P(B)$

The Addition Rule brings the intersection and union of events together and links onto mutually exclusive events as follows:

For events A and B you have to understand

○ the meaning of the *union of events*	A or B
○ the meaning of the *intersection of events*	A and B
○ the meaning of the *addition rule*	$P(A \text{ or } B) = P(A) + P(A) - P(A \text{ and } B)$
○ the meaning of *mutually exclusive events*	No overlapping of outcomes
○ that *mutually exclusiveness* changes the addition rule into	$P(A \text{ or } B) = P(A) + P(A)$

5.9 Probability rules and trees

5.9.1 Complement rule

The complement rule of an event A is that even A does not occur. It is symbolized by $P\ (A)^C$.

$$P\left(A^C\right) = 1 - P\ (A)$$

5.9.2 Example

If the probability of passing an assignment is 0.7 the probability of falling the exam is $1 - 0.7 = 0.3$

$$P\ (\text{passing}) = 0.7, \quad P\left(\text{passing}^C\right) = 1 - 0.7 = 0.3$$

5.9.3 Multiplication rule

The joint probability of any two events A and B is given by

$$P\ (A \text{ and } B) = P\ (B)\ P\ (A|B) \quad \text{or} \quad \text{P}\ (A \text{ and } B) = P\ (A)\ P\ (B|A)$$

If the events A and B are independent, $P\ (A|B) = P\ (A)$ and $P\ (B|A) = P\ (B)$

5.9.4 Selections with or without replacement

The concept of *replacement or not* of items in a selection, plays a very important role in the probability calculations that follow on the sampling process. Imagine that you want to send out a questionnaire and for that you have to choose 50 names randomly from a list of 200 names. There are two possible ways to select these names randomly. Suppose that the randomly selected name in the first and follow-up draws are crossed out every time and removed from the original list, then the random selection of the next name is from a different list of names every time and this is called sampling *without replacement*. However, should you not cross out the first and follow-up names from the list (simply write them on an extra piece of paper), then the random selections are always made from the original list, this is called sampling *with replacement*. Such a method can result in one name being selected again, and that is a bit silly! Would you prefer to select the 50 names with or without replacement? Of course without replacement! Choosing 50 names with replacement is of little significance in practice as it may result in less than 50 *persons* in the list of names if names were selected more than once! You would not like to send the same questionnaire to one person twice because his/her name was selected more than once. Another application of selection without replacement is the South African lotto. Six numbers are drawn without replacement, so no repetition of numbers occur.

5.9.5 Example

Question 1

Use the events A_2 and B_1 as given in Example 6.1 in Keller and determine if these two events are independent.

Question 2

The joint probabilities shown in a table with two rows, $A_1$and A_2, and two columns, $B_1$and B_2, are as follows:

$P(A_1 \text{ and } B_1) = 0.10,\ P(A_1 \text{ and } B_2) = 0.30,\ P(A_2 \text{ and } B_1) = 0.05$,and $P(A_2 \text{ and } B_2) = 0.55$.

Then $P(A_2) =$

(1) 0.85

(2) $P(A_2 \text{ and } B_1) + P(A_2 \text{ and } B_2)$

(3) $P(A_2 \text{ and } B_1) + P(A_2 \text{ and } B_2) - P(B_2)$

(4) 0.40

(5) $P(A_2 \mid B_1) + P(A_2 \mid B_2)$

(Hint: Make a table of joint and marginal probabilities.)

Question 3

Use the information given in question 2 and compute

(a) $P(B_2 \text{ or } B_1)$ and

(b) $P(A_2 \text{ or } B_1)$.

Solution

Question 1

$P(A_2) = 0.60$ and $P(B_1) = 0.17$ while $P(A_2 \text{ and } B_1) = 0.06$

The test for independence:

$$P(A_2 \mid B_1) = \frac{P(A_2 \text{ and } B_1)}{P(B_1)}$$

$$= \frac{0.06}{0.17}$$

$$= 0.35294$$

$$\neq 0.60 \quad (P(A_2)).$$

This implies that A_2 and B_1 are *not* independent, because if they were, then $P(A_2 \mid B_1) = P(A_2)$.

Question 2

Answer: (2)

Joint and marginal probabilities:

	$\mathbf{A}_1$	$\mathbf{A}_2$	**Totals**
$\mathbf{B}_1$	$P(B_1 \text{ and } A_1) = 0.10$	$P(B_1 \text{ and } A_2) = 0.05$	$P(B_1) = 0.15$
$\mathbf{B}_2$	$P(B_2 \text{ and } A_1) = 0.30$	$P(B_2 \text{ and } A_2) = 0.55$	$P(B_2) = 0.85$
Totals	$P(A_1) = 0.40$	$P(A_2) = 0.60$	1.00

Note that the sum of the marginal totals must always add up to 1.00.

(In this question $0.15 + 0.85 = 1$ and $0.40 + 0.60 = 1$.)

Question 3

$P(B_2 \text{ or } B_1)$. Intuitively you should know that this answer must be 1.00, because either B_2 or B_1 has to occur. If you did not see this, you can calculate it as follows:

$$\begin{aligned} P(B_1 \text{ or } B_2) &= P(B_1) + P(B_2) - P(B_1 \text{and } B_2) \\ &= 0.15 + 0.85 + 0 \text{ (because } B_1 \text{and } B_2 \text{ are mutually exclusive)} \\ &= 1.00. \end{aligned}$$

$P(A_2 \text{ or } B_1)$ can be read from the table, if you reason that either A_2 or B_1 has to take place, which implies that A_2 can occur with either B_1 or with B_2 and B_1 can occur with either A_1 or with A_2. That gives three possible events, i.e. A_2 and B_1, or A_2 and B_2, or A_1 and B_1. Add these probabilities from the table, namely $0.05 + 0.55 + 0.10 = 0.70$.

$P(A_2 \text{ or } B_1)$, using the addition rule:

(The value for $P(A_2)$ is found in the marginal probability of the column under A_2, which is 0.60.)

$$\begin{aligned} P(A_2 \text{ or } B_1) &= P(A_2) + P(B_1) - P(A_2 \text{ and } B_1) \\ &= 0.60 + 0.15 - 0.05 \\ &= 0.70 \end{aligned}$$

5.10 Probability tree

5.10.1 Definition

A probability tree is an effective way of using and applying the rules of probabilities. The tree is used to compute probabilities for several trials or experiment. All the possible outcomes of the experiment are represented by the branches.

5.10.2 Example

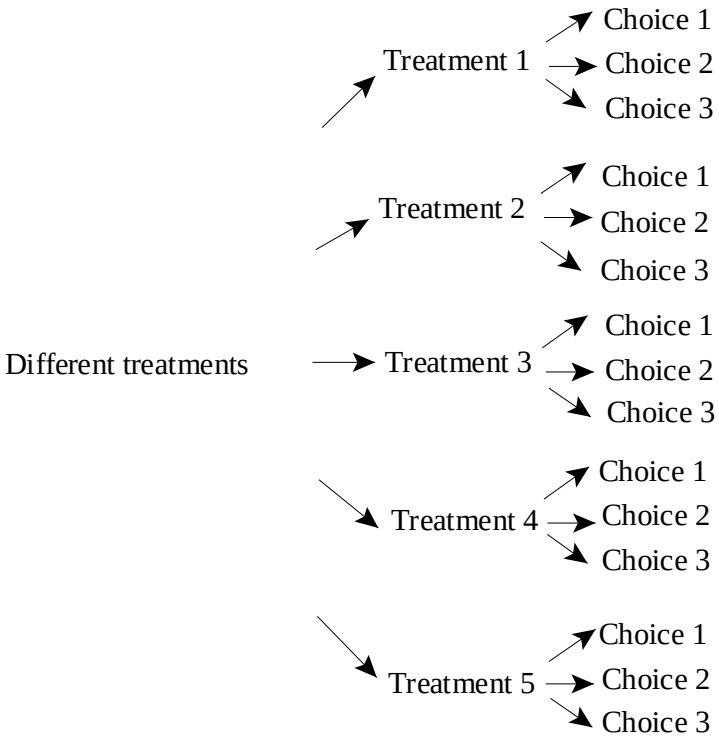
The correct answer is option (4). You cannot give the probability as a percentage. If one student must pass and the other must fail, you have to remember that the order of selections plays a role. It could be that the first student selected failed and the second one passed, or the first one could have passed and the second one failed. If you made a probability tree, the correct answer would have been clear to you.

The answer is the option $P(F \mid P)$ or $P(P \mid F) = 0.16 + 0.16 = 0.32$.

Question 2

(a) False

If you had made a probability tree to illustrate the given information, you would have seen that there were 15 possibilities for treating the patient.



(b) True

Seeing that the decisions "to attend" or "not to attend" are independent, there are 5 independent events taking place and each event has two possible outcomes (attend and not attend). There are therefore $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^5 = 32$ possible compositions of the group.

Question 3

(a) $P(R_1 \text{and } R_2) = P(R_1|R_2)P(R_2) = 0.8 \cdot 0.5 = 0.4$

(b) $P(R_1 \text{or } R_2) = P(R_1) + P(R_2) - P(R_1 \text{and } R_2) = 0.6 + 0.5 - 0.4 = 0.7$

(c) $P(R_1^c|R_2) = 1 - P(R_1|R_2) = 1 - 0.8 = 0.2$

5.11 The rule of Bayes

5.11.1 Bayes's Law formula

$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + ... + P(A_k)\,P(B|A_k)}$$

Do you remember we explained at the beginning of this chapter that a statistician whose main statistical interest lies in probability is called a *probabilist*? Now, Thomas Bayes was a probabilist. He derived a very special application of the multiplicative law that came to be known as the *Law of Bayes*, which is one of the most important advanced rules of probability. In all the conditional problems discussed so far, we wanted to know the probability of an event A at a specified time, given that an event B had occurred at a prior point in time. This process can be approached from another direction. The interest may be in the probability that a particular state exists given that a certain sample is observed. This Bayesian approach has a language of its own, where the word *prior* refers to *before* and *posterior* refers to *after*. The Bayes approach to probability is most applicable in medicine, but just as relevant in many other professions. Most of us have to undergo some form of medical tests at a certain stage of our lives and it is not good to read in Keller that very few of these tests are 100% accurate. Fortunately there is something like evidence-based medicine and we know that physicians are forced to use Bayes' Law to determine the true probabilities associated with screening tests!

For years many statisticians neglected the so-called "Bayesian approach" to inference, which is different and more complex than the classical approach to inference. Fortunately this division between the different theoretical groups of statisticians is gradually disappearing and a broader view of the mathematical and logical foundations of statistics has been initiated. We want to introduce you to Bayes' Law, but we realise the complexity of its applications. The examples in Keller are very interesting and relevant. It is good that you become aware of some of the very important real-life applications of the theories that you learn.

You will not be tested on complex applications of the Law of Bayes.

A few pointers:

- Please remember that conditional probabilities have complements in the same way as $P(A) = 1 - P(A^c)$, e.g.

$$P(B|A) = 1 - P(B^c|A)$$

$$P(B^c|A^c) = 1 - P(B|A^c)$$

- Read through example 6.10 for the sake of general knowledge.

5.11.2 Examples

Question 1

Suppose $P(A) = 0.10$, $P(B|A) = 0.20$, and $P(B|A^c) = 0.40$, find

(a) $P(A \text{ and } B)$

(b) $P(A \text{ and } B^c)$

(c) $P(B^c)$

(d) $P(A \text{ or } B)$

Question 2

Choose the correct option:

A posterior probability value is a prior probability value that has been

(1) changed to a likelihood probability

(2) modified on the basis of new information

(3) multiplied by a conditional probability value

(4) divided by a conditional probability value

(5) added to a conditional probability value

Question 3

A mother with one daughter is expecting her second child. Her doctor has told her that she has a 50 - 50 chance of having another girl. If she has another girl, there is a 90% chance that she will be taller than the first. If she has a boy, however, there is only a 25% chance that he will be taller than the first child. Find the probability that the woman's second child will be taller than the first.

Solutions:

Question 1

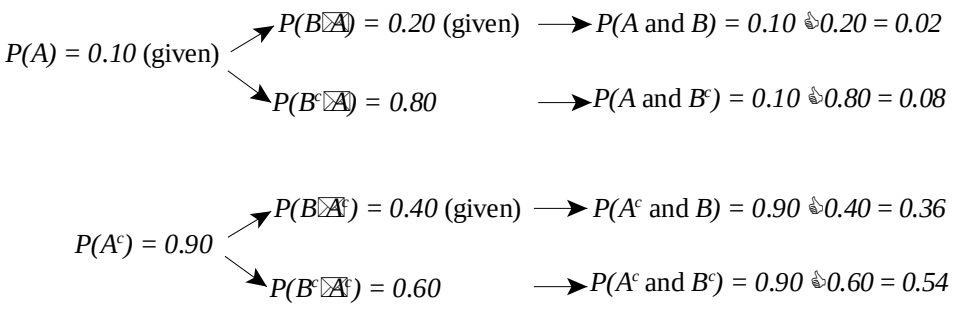
(a) Some mathematical manipulation is necessary if you want to determine this answer:

You know that

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \quad \text{which implies that}$$

$$\begin{aligned} P(A \text{ and } B) &= P(B \mid A) \cdot P(A) \quad \text{(mathematical manipulation)} \\ &= 0.20 \cdot 0.10 \\ &= 0.02 \end{aligned}$$

(b) The mathematical manipulation used above illustrates the principles applied in a probability tree. Use the probability tree below and you can read off the answers:



$P(A \text{ and } B^c) = 0.10 \cdot 0.80 = 0.08$

(c) To determine $P(B^c)$ the principle applied is that B^c occurred either with A or with A^c. Then use the tree once again for the probabilities

$$\begin{aligned} P(B^c) &= P(B^c \text{ and } A) + P(B^c \text{ and } A^c) \\ &= 10 \cdot 0.80 + 90 \cdot 0.60 \quad \text{(from the probability tree)} \\ &= 0.08 + 0.54 \\ &= 0.62 \end{aligned}$$

(d) $$\begin{aligned} P(A \text{ or} B) &= P(A) + P(B) - P(A \text{ and} B) \quad \text{(Recognise this as the addition rule?)} \\ &= 0.10 + (1 - 0.62) - 0.02 \\ &= 0.46 \end{aligned}$$

Wasn't this a beautiful question? To answer a question like this, you have to bring your knowledge of probability together in your mind!

Question 2

Option (2) is the correct one.

Question 3

Define the probabilities of having a boy (B) and a girl (G).

Make a probability tree and fill in the given conditional probabilities. Let T indicate a taller baby the second time and T^c indicate that the second baby will not be taller. Only $P(T \mid B) \cdot P(B)$ and $P(T \mid G) \cdot P(G)$ have to be determined and added.

$$\begin{aligned} P(T \mid B) \cdot P(B) + P(T \mid G) \cdot P(G) &= 0.25 \cdot 0.5 + 0.9 \cdot 0.5 \\ &= 0.125 + 0.45 \\ &= 0.575 \end{aligned}$$

The probability that the second child will be taller than the first baby is 0.575.

The value of the summary, important terms and formulas used in the chapter cannot be emphasised enough. Don't you think that authors are really smart nowadays? They study what the needs of students are and then include it in the book. Feel free to send comments (positive or negative) to the Department about this module and the manner in which it is presented. We must know how you reason and where you have problems. Keep in mind that there is no magic involved in studying statistics – it is all hard work! Once the knowledge and insight is your own – this is when the magic starts!

Key terms

Random experiment
Exhaustive outcomes
Mutually exclusive outcomes
Sample space
Event
Probability of an event
Joint probability
Intersection of events
Marginal probabilities
Conditional probability
Independent events
Union of events
Complement rule
Multiplication rule
Multiplication rule for independent events
With replacement/without replacement
Addition rule
Addition rule for mutually exclusive events
Probability trees
Bayes' Law

5.12 Learning Outcomes

Use the chapter summary as a checklist after you have completed this study unit to evaluate if you have really acquired a good understanding of the work covered.

Can you

- explain what is meant by *a random experiment* and its *outcomes?*
- define the term *sample space* and the concepts of *exhaustive* and *mutually exclusive events*?
- explain the two *requirements of probabilities?*
- discuss three approaches to *assign probabilities?*
- describe the difference between *simple and complex* events?
- explain and distinguish between *joint, marginal* and *conditional probability?*
- define the following rules and differentiate between them?

    - *complement rule*
    - *multiplication rule (in general and for independent events)*
    - *additional rule (in general and for mutually exclusive events)*

- draw a *probability tree* for applications of the different rules?
- understand *Bayes' Law* and the relevant notation?

5.13 Study Unit 5: Summary

I. A *probability* is a value between 0 and 1 inclusive that represents the likelihood that a particular event will happen.

   A. An *experiment* is the observation of some activity or the act of taking some measurement.

   B. An *outcome* is a particular result of an experiment.

   C. An *event* is the collection of one or more outcomes of an experiment.

II. There are three *definitions of probability*.

   A. The classical definition applies when there are n equally likely outcomes to an experiment.

   B. The empirical definition occurs when the number of times an event happens is divided by the number of observations.

   C. A subjective probability is based on whatever information is available.

III. Two events are *mutually exclusive* if by virtue of one event happening the other cannot happen.

IV. Events are *independent* if the occurrence of one event does not affect the occurrence of another event.

V. The *rules of addition* refer to the union of events.

   A. The special rule of addition is used when events are mutually exclusive.

$$P(A \text{ or } B) = P(A) + P(B)$$

   B. The general rule of addition is used when the events are not mutually exclusive.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

   C. The complement rule is used to determine the probability of an event happening by subtracting the probability of the event not happening from 1.

$$P(A) = 1 - P\left(A^C\right)$$

VI. The *rules of multiplication* refer to the product of events.

   A. The special rule of multiplication refers to events that are independent.

$$P(A \text{ and } B) = P(A)\,P(B)$$

   B. The general rule of multiplication refers to events that are not independent.

$$P(A \text{ and } B) = P(A)\,P(B|A)$$

   C. A *joint probability* is the likelihood that two or more events will happen at the same time.

D. A *conditional probability* is the likelihood that an event will happen, given that another event has already happened.

VII. There are three *counting rules* that are useful in determining the number of outcomes in an experiment.

A. The multiplication rule states that if there are m ways one event can happen and n ways another event can happen, then there are mn ways the two events can happen.

$$\text{Number of arrangements } = (m)\,(n)$$

B. A permutation is an arrangement in which the order of the objects selected from a specific pool of objects is important.

$$_nP_r = \frac{n!}{(n-r)!}$$

C. A combination is an arrangement where the order of the objects selected from a specific pool of objects is not important.

$$_nC_r = \frac{n!}{r!\,(n-r)!}$$

Reference

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 6

RANDOM VARIABLES AND DISCRETE PROBABILITY DISTRIBUTIONS

6.1 Learning outcomes

After completing this unit, you should be able to

1. distinguish between a discrete and continuous random variable
2. explain the criteria for a distribution of a discrete random variable
3. describe the population, its probability distribution and the role of the population parameters
4. use the laws of expected value and variance
5. distinguish between univariate and multivariate distributions
6. describe the marginal probabilities, the parameters and covariance of bivariate distribution
7. define a binomial experiment and the binomial probability distribution
8. use the binomial table
9. define the Poisson experiment and the Poisson distribution
10. use the Poisson table

6.2 Introduction

Now that you have a good idea about probability, we have to use your knowledge to introduce you to *probability distributions*. In order to get to probability distributions, the idea of a *random variable* must be understood. Do you still remember what the meaning of *random* is? In the previous chapter you were told that a random experiment involves an "action that leads to one of several possible outcomes". Furthermore, in many examples and questions mention was made of "randomly chosen" names, or numbers, or.... In standard language *random* means that you do not make a choice wittingly (e.g. you make a draw out of a hat, or you allow a number generator to choose a number for you). The word *variable* implies a varying value, so it follows that a *random variable* can assume different values, determined by the outcome of a random action.

Keller defines a *probability distribution* as a table, formula, or graph that describes the values of a random variable and the probability associated with these values. The probability distribution implies that a theoretically determined probability (as discussed in Chapter 6) is used to form a model that describes the nature of the particular event and the likelihood that such an event will occur. In this way better decisions can be made in statistical inference, which is the method used in statistics to generalize about a population, based on a sample from that population. As we proceed through chapters 7 and 8 you will come to realize the value of probability distributions as *representatives of populations*.

Suppose we consider all the possible outcomes of a statistical experiment involving a random variable and for each outcome also give its probability of occurrence. The combination of all these pairs (outcome plus its respective probability) forms the probability distribution of that random variable. If the random variable was discrete, we call it a discrete probability distribution and if the random variable was continuous, we call it a continuous probability distribution.

Chapter 6 is only about discrete probability distributions.

6.3 Random variables and probability distributions

6.3.1 Random variable

A *random variable* is function or rule that assigns a number to each outcome of an experiment. There are two types of random variables, discrete and continuous.

A discrete random variable is a variable that can take on a countable number of values, in other words, a discrete random variable can assume a countable number of possible outcomes.

6.3.2 Example

1. The number of accidents that occur on N1 highway every one hour is a random variable.

2. The delivery time of parcels to clients.

A continuous random variable is random variable which can take on any value over a given intervals of values.

A probability distribution is a table, or a formula that describes the values of a random variable.

6.4 Discrete probability distribution

6.4.1 Definition

A *discrete probability distribution* is a listing of all possible outcomes a discrete random variable can assume.

The word "discrete" in everyday language is mostly used to describe a person who is modest and considerate; one whom you can trust to keep a secret. We suppose the statisticians used the words *discrete random variable* for outcomes that can be counted and listed, which are fixed – you can "trust" these outcomes! Remember that there are "gaps" between the different outcomes of a discrete random variable, whereas the outcomes of a continuous random variable fall within an interval. You cannot pin them down and there are no gaps between the possibilities.

Make sure that you understand the notation $P(X = x)$, which is the same as the shorter $P(x)$. In terms of the example where two balanced coins are flipped and the results observed, we write $P(X = 2)$ to indicate that we consider the probability that, for the random variable X, the result is an outcome with 2 heads occurring. The shorter notation for this particular outcome would be $P(2)$.

6.4.2 Requirements for a distribution of discrete random variable

The requirements for the distribution of a discrete random variable are very important. Do they seem familiar to you? Of course; they should remind you of the requirements of Probabilities in 6.1, which were very similar. The similarity lies in the fact that these requirements are also for *probabilities*, namely

Statement in words	Symbols
Probabilities cannot be negative and cannot be larger than one.	$0 \leq P(x) \leq 1$
The sum of the probabilities of all possible outcomes must be equal to one.	$\sum_{all\ x} P(x) = 1$

6.5 Describing the population / probability distribution

Do you remember what a *parameter* is? If you do, what is the difference between a parameter and a *statistic*? Ah, now we think that we have helped you! You were told that a parameter is a descriptive measure of a population, whereas a statistic describes a sample. Did you remember?

Two parameters were discussed, namely *population mean* and *variance*, with which you should be familiar. The new concept is that the population mean is also called the *expected value* of the random variable, with notation $E(X)$.

You know very well how one weighs fruit, vegetables, ... but what on earth is a *weighted average* and a *weighted variance*? The whole principle of weighing is *balance*. Chemists talk about a "chemical balance" when they weigh matter. Balance is reached in a "weighted mean" by considering the *importance of each observation* and according to that, allocating weights to them. An example will illustrate this best.

6.5.1 The population mean

The population mean is the weighted average of all the values. The weights are the probabilities, also called the expected value of X.

The population mean is given by:

$$E(x) = \mu = \sum x P(x)$$

6.5.2 The population variance

$$V(x) = \sigma^2 = \sum (x - \mu)^2 P(x)$$

or $\sum x^2 P(x) - \mu^2$

6.5.3 The standard deviation

$$\sigma = \sqrt{\sigma^2}$$

6.5.4 Example

Suppose that you do certain tasks in order to supplement your monthly allowance: you baby-sit once a week; sell a snack from Monday to Friday during lunch time; drive people around during the weekends. The profit

for each task is as follows:

Description	**Profit per attempt**	**Number of opportunities during February**
Baby-sitting	$R30$	4
Snack selling	$R0.55$	65
Taxi service	$R12$	15

If you want to determine the mean (or average) per activity for the month of February doing these tasks, how do you go about this? In this way?

$$R\left(\frac{30+0.55+12}{3}\right) = R\frac{42.55}{3} = R14.183$$

That is simply the mean profit per task, which does not make a lot of sense! A weighted mean profit would consider the number of times an event took place as a weights indicator. In this example snack selling had the lowest profit, but it also had the largest frequency, so it is given a bigger weight than baby-sitting with the largest profit, but which realised only 4 times during February. The calculation for a more realistic average is

$$\begin{aligned}\text{Weighted mean} &= \frac{\sum w_i x_i}{\sum w_i} \\ &= R\frac{(30\times 4)+(0.55\times 65)+(12\times 15)}{4+65+15} \\ &= R3.997.\end{aligned}$$

In the probability distribution function we use the probabilities of occurrence as the weights. The weighted mean is also called the expected value, i.e.

$$E(X) = \sum_x xP(x).$$

This allows each probability to serve as a weight to be associated with the corresponding outcome to balance the situation. Please study example 7.3 in the book about weighted means and weighted variance. Note that the formula for the weighted variance also uses the probabilities as weights.

6.5.5 Laws of expected value and variance

Expected value	Variance
$E(c) = c$ E.g. 5, is "expected" to be exactly 5.	$V(c) = 0$ A number has no "variability"; its variance should be zero.
$E(X + c) = E(X) + c$ This really makes sense!	$V(X + c) = V(X)$ Previous arguments about having no variance apply.
$E(cX) = cE(X)$ Extension of first argument.	$V(cX) = c^2V(X)$ Very interesting – constant is *squared*! Read note below.

Note: One of the most important facts in statistics is that a *variance can never be negative*. See how beautiful this last law is – assume that c is a negative constant, say -3. In such a case then

$E(cX) = cE(X) \Longrightarrow E(-3X) = -3E(X)$ and

$V(cX) = c^2V(X) \Longrightarrow V(-3X) = (-3)^2\,V(X) = 9V(X).$

6.5.6 Examples

Question 1

The following statements are either true (T) or false (F). Think carefully and classify the following questions:

(a) The length of time for which an apartment in a large complex remains vacant is a discrete random variable.

(b) The number of homeless people in Johannesburg is an example of a discrete random variable.

(c) Given that X is a discrete random variable, the laws of expected value and variance can be applied to show that

$$E(X + 5) = E(X), \text{ and } V(X + 5) = V(X) + 25.$$

(d) A table, formula, or graph that shows all possible values that a random variable can assume, together with their associated probabilities, is referred to as discrete probability distribution.

(e) The number of students that use a computer lab during one day is an example of either a continuous or a discrete random variable, depending on the number of the students.

Question 2

Determine which of the following are not valid probability distributions, and if not, explain.

(a)

x	0	1	2	3
$p(x)$	0.15	0.25	0.35	0.45

(b)

x	2	3	4	5
$p(x)$	-0.10	0.40	0.50	0.20

(c)

x	-2	-1	0	1	2
$p(x)$	0.10	0.20	0.40	0.20	0.10

Question 3

If X is any random variable, some of the following identities are *not true*.

(a) $E(X+2) = E(X) + E(2)$

(b) $V(X+2) = V(X)$

(c) $E(4X-6) = 4E(X) - 6$

(d) $V(4X-6) = 16V(X) + 36$

(e) $V(-4X+6) = -16V(X)$

Choose the correct option:

1. Only statement d
2. Statements b and d
3. Statements d and e
4. Only e
5. Statements b and e

Solutions:

Question 1

(a) Answer: F (time is a continuous variable).

(b) Answer: T (these people can be counted).

(c) Answer: F ($E(X+5) = E(X) + 5$, and $V(X+5) = V(X)$).

(d) Answer: T

(e) Answer: F (the number of students is always a discrete random variable).

Question 2

(a) This is not a valid probability distribution because the probabilities don't sum to one.

(b) This is not a valid probability distribution because it contains a negative probability.

(c) This is a valid probability distribution.

Question 3

The correct option is 3. Statements d and e are both false, because

$V(4X - 6) = 16V(X)$ and

$V(-4X + 6) = 16V(X)$.

6.6 Bivariate distributions

6.6.1 Definitions

Bivariate distribution is a combination of two variables, the joint distribution that two variables will assume the values x and y is given by;

$P(x, y)$. A discrete bivariate distribution assume that:

1. $0 \leq P(x, y) \leq 1$ for all pairs of values (x, y)

2. $\sum\sum P(x, y) = 1$

Why do statisticians have to use words with which we are not familiar? Just look at this. Maybe it is, after all, not that bad!

Prefix	**Reference**	**Example**
uni-	one	*uni*variate (one variable)
bi-	two	*bi*variate (two variables)
tri-	three	*tri*nomial (mathematical expression with three terms)
quad-	four	*quad*rilateral (geometrical figure with four sides)
multi-	many	*multi*variate (many different variables)

All the discussions so far in this chapter referred to the univariate case and you will see that there is a very logical expansion to the bivariate case. Again, there are formulas to study carefully and understand the logic in each one! We have a beautiful discipline in statistics, as accumulation of knowledge can be compared to the construction of a house. One inside wall in a house forms part of two or more different rooms and you will see that certain statistical concepts are also relevant is different sections. Marginal Probabilities, Covariance and the Coefficient of Correlation were all discussed in previous chapters (6 and 4). Please make sure, while studying for this module, that you link concepts to their applications in the different sections. Then you will not feel as if you have a box full of puzzle pieces, but you will have a beautiful completed picture in your mind!

Would you like to see my reasoning when I consider the different formulas given in this section?

Univariate	Bivariate	Logic behind the formula
$P(x)$	$P(x, y)$	The single variable got "married". The pair of variables form a unit, but are still unique within themselves.
$0 \leq P(x) \leq 1$	$0 \leq P(x, y) \leq 1$	Rule: All probabilities must be positive, but not larger than 1. Applies to one or more variables.
$\sum_{all\ x} P(x) = 1$	$\sum_{all\ x}\sum_{all\ y} P(x, y) = 1$	The sum of probabilities of all possible outcomes involving both variables must be added and be equal to one.
$\sum_{all\ x}(x - \mu_x)^2 P(x)$ OR $\sum_{all\ x}(x - \mu_x)(x - \mu_x)P(x)$	$\sum_{all\ x}\sum_{all\ y}(x - \mu_x)(y - \mu_y)P(x, y)$	Variance (one variable x) considers the squared deviations from the mean and in covariance the squared deviation is "split up"; one part for each variable.
	$\rho = \frac{COV(X, Y)}{\sigma_x \sigma_y}$ with x and y *independent* $\rho = 0$ as $COV(X, Y) = 0$	This coefficient of correlation applies to combined variables and measures strength and direction of the relationship.

6.6.2 Examples

Question 1

Study examples 7.6 and 7.7 to ensure that you understand the principles applied. Then do exercises 7.54 and 7.56. Only compare your answers with those given at the back of the book once you have completed your own attempt on both questions.

Question 2

The joint probability distribution of variables X and Y is shown in the table below, where X (the number of tennis racquets) and Y (the number of golf clubs) indicate the daily sales in a small sports store.

$Y \diagup X$	1	2	3
1	0.30	0.18	0.12
2	0.15	0.09	0.06
3	0.05	0.03	0.02

(a) Determine the marginal probability distributions of X and Y.

(b) Are X and Y independent? Explain.

(c) Find $P(Y = 2|X = 1)$.

(d) Calculate the expected values of X and Y.

(e) Calculate the variances of X and Y.

(f) Calculate $COV(X, Y)$. Did you expect this answer? Why?

(g) Find the probability distribution of the random variable $X + Y$.

Solutions:

Question 1

Answers are at the back of the textbook. Please compare your answers.

Question 2

(a)

x	1	2	3
$p(x)$	0.5	0.3	0.2

y	1	2	3
$p(y)$	0.6	0.3	0.1

(b) Determine $p(x, y)$ and $p(x).p(y)$ for all pairs (x, y) and you will find that $p(x, y) = p(x).p(y)$, which makes them independent.

(c) $P(Y = 2|X = 1) = \dfrac{P(X = 1 \text{ and } Y = 2)}{P(X = 1)}$

$= \dfrac{0.15}{0.50}$

$= 0.30.$

(d) $E(X) = 1.7$ and $E(Y) = 1.5.$

(e) $V(X) = 0.61$ and $V(Y) = 0.45.$

(f) $COV(X, Y) = E(XY) - E(X).E(Y) = 2.55 - (1.70)(1.50) = 0.0.$

Yes, since X and Y are independent.

(g)

$x + y$	2	3	4	5	6
$p(x + y)$	0.30	0.33	0.26	0.09	0.02

6.7 Binomial distribution

Experiments or investigations generating two possible outcomes, such as success or failure, accept or decline are referred to as binomial experiment. The binomial distribution is a result of a binomial experiment, which has the following characteristics:

1. A fixed number of trials (sample size), the number of trials is symbolised by n,
2. Each trial has 2 possible outcomes; success or failure,
3. The probability of success is p and the operability of failure is $1 - p$,
4. The trials are independent, meaning that the outcome of one trial does not affect the outcomes of any other trials or the trials are not related.

Note: If only 2, 3, 4 must be satisfied, each trial is called a Bernoulli process.

The probability of x successes in a binomial experiment with n trials and probability of success p is given by :

$$P(x) = \frac{n1}{x!(n - x)!} p^x (1 - p)^{n-x}$$

for $x = 0,\ 1,\ 2,\ ...,$

6.7.1 Example

Pat is a student taking a statistic course. Unfortunately, Pat is not a good student. Pat does not read the textbook before class, does not do homework, and regularly misses class. Pat intends to rely on luck to pass the next quiz. The quiz consists of 10 multiple questions. Each question has five possible answers, only one of which is correct.

Pat plans to guess the answer to each question.

(a) What is the probability that Pat gets no answers correct?

(b) What is the probability that Pat gets two answers correct?

Solution

The experiment consists of 10 identical trials, each with 2 outcomes (correct or wrong). The probability of success is only one out of 5 answer meaning that the probability of success is $\frac{1}{5} = 0.2$

Are the trials independent? Yes, because the outcome of one questions does not affect the outcome of the others questions. The sample size(number of trial), $n = 10$ and the $p = 0.2$

(a) No answer means the value of the random variable X is 0, ($P(x = 0)$, by substituting p, $1 - p$, n and in the formula of the binomial distribution, we get

$$\begin{aligned} P(0) &= \frac{10!}{0!(10-0)!}0.2^0(1-0.2)^{10-0} \\ &= 0.1074 \end{aligned}$$

(b) The value of the random variable is 2, by substituting the value of p, $1 - p$, n and x in the formula, we get

$$\begin{aligned} P(2) &= \frac{10!}{2!(10-2)!}0.2^2(1-0.2)^{10-2} \\ &= 0.3020 \end{aligned}$$

6.8 Cumulative probability

The formula of the binomial distribution allows us to find the probability that the random variable X equals individuals values. In the previous example, the value of the random variables X were 0 and 2 . There are many cases where we wish to determine the probability that the random variable X is less or equal to a value $P(X \leq x)$, where x is that value, such a probability is called cumulative probability.

6.8.1 Example

Find the probability that Pat fails the quiz. A mark is considered a failure if it is less that 50%.

Solution

In this quiz, a mark of less than 5 is a failure. Because the marks must be integers, a mark of 4 or less is a failure. The question is to determine the probability that random variable X is less or equal to 4.

$$P(x \leq 4) = P(0) + P(1) + P(2) + P(3) + P(4)$$

From the previous example, we known that $P(0) = 0.1074$ and $P(2) = 0.3020$. Using the Binomial formula, we find $P(1) = 0.2684$, $P(3) = 0.2013$, and $P(4) = 0.0881$. Thus,

$$P(x \leq 4) = 0.1074 + 0.2684 + 0.3020 + 0.2013 + 0.0881 = 0.9672.$$

6.8.2 Binomial table

The binomial table can be used to determine the probabilities; $P(X \geq x) = 1 - (P(X \leq [x-1]$ the or $P(X = x) = P(X \leq x) - P(X \leq [x-1])$. The following table provides cumulative binomial probabilities for selected values of n and p.

Table 1: Binomial Probabilities

Table of Binomial Probabilities

For a given combination of n and π, entry indicates the probability of obtaining a specified value of X. To locate entry, when $\pi \leq .50$ read π across the top heading and both n and X down the left margin; when $\pi \geq .50$ read π across the bottom heading and both n and X up the right margin.

		π																			
n	X	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	X	n
2	0	0.9801	0.9604	0.9409	0.9216	0.9025	0.8836	0.8649	0.8464	0.8281	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	2	
	1	0.0198	0.0392	0.0582	0.0768	0.0950	0.1128	0.1302	0.1472	0.1638	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000	1	
	2	0.0001	0.0004	0.0009	0.0016	0.0025	0.0036	0.0049	0.0064	0.0081	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500	0	2
3	0	0.9703	0.9412	0.9127	0.8847	0.8574	0.8306	0.8044	0.7787	0.7536	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250	3	
	1	0.0294	0.0576	0.0847	0.1106	0.1354	0.1590	0.1816	0.2031	0.2236	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750	2	
	2	0.0003	0.0012	0.0026	0.0046	0.0071	0.0102	0.0137	0.0177	0.0221	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750	1	
	3	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0007	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250	0	3
4	0	0.9606	0.9224	0.8853	0.8493	0.8145	0.7807	0.7481	0.7164	0.6857	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625	4	
	1	0.0388	0.0753	0.1095	0.1416	0.1715	0.1993	0.2252	0.2492	0.2713	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500	3	
	2	0.0006	0.0023	0.0051	0.0088	0.0135	0.0191	0.0254	0.0325	0.0402	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750	2	
	3	0.0000	0.0000	0.0001	0.0002	0.0005	0.0008	0.0013	0.0019	0.0027	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500	1	
	4	—	—	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625	0	4
5	0	0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312	5	
	1	0.0480	0.0922	0.1328	0.1699	0.2036	0.2342	0.2618	0.2866	0.3086	0.3280	0.3915	0.4096	0.3955	0.3601	0.3124	0.2592	0.2059	0.1562	4	
	2	0.0010	0.0038	0.0082	0.0142	0.0214	0.0299	0.0394	0.0498	0.0610	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125	3	
	3	0.0000	0.0001	0.0003	0.0006	0.0011	0.0019	0.0030	0.0043	0.0060	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125	2	
	4	—	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003	0.0004	0.0022	0.0064	0.0146	0.0283	0.0488	0.0768	0.1128	0.1562	1	
	5	—	—	—	—	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0312	0	5
6	0	0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156	6	
	1	0.0571	0.1085	0.1546	0.1957	0.2321	0.2642	0.2922	0.3164	0.3370	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0937	5	
	2	0.0014	0.0055	0.0120	0.0204	0.0305	0.0422	0.0550	0.0688	0.0833	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344	4	
	3	0.0000	0.0002	0.0005	0.0011	0.0021	0.0036	0.0055	0.0080	0.0110	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125	3	
	4	—	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1372	0.1861	0.2344	2	
	5	—	—	—	—	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0937	1	
	6	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156	0	6

7	0	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078	7	
	1	0.0659	0.1240	0.1749	0.2192	0.2573	0.2897	0.3170	0.3396	0.3578	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547	6	
	2	0.0020	0.0076	0.0162	0.0274	0.0406	0.0555	0.0716	0.0886	0.1061	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	5	
	3	0.0000	0.0003	0.0008	0.0019	0.0036	0.0059	0.0090	0.0128	0.0175	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734	4	
	4	—	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0011	0.0017	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734	3	
	5	—	—	—	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641	2	
	6	—	—	—	—	—	—	—	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547	1	
	7	—	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078	0	7
8	0	0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039	8	
	1	0.0746	0.1389	0.1939	0.2405	0.2793	0.3113	0.3370	0.3570	0.3721	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0312	7	
	2	0.0026	0.0099	0.0210	0.0351	0.0515	0.0695	0.0888	0.1087	0.1288	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094	6	
	3	0.0001	0.0004	0.0013	0.0029	0.0054	0.0089	0.0134	0.0189	0.0255	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2187	5	
	4	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0013	0.0021	0.0031	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734	4	
	5	—	—	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2187	3	
	6	—	—	—	—	—	—	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094	2	
	7	—	—	—	—	—	—	—	—	—	—	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0312	1	
	8	—	—	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039	0	8
9	0	0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020	9	
	1	0.0830	0.1531	0.2116	0.2597	0.2985	0.3292	0.3525	0.3695	0.3809	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176	8	
	2	0.0034	0.0125	0.0262	0.0433	0.0629	0.0840	0.1061	0.1285	0.1507	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703	7	
	3	0.0001	0.0006	0.0019	0.0042	0.0077	0.0125	0.0186	0.0261	0.0348	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641	6	
	4	0.0000	0.0000	0.0001	0.0003	0.0006	0.0012	0.0021	0.0034	0.0052	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461	5	
	5	—	—	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008	0.0050	0.0165	0.0390	0.0735	0.1181	0.1672	0.2128	0.2461	4	
	6	—	—	—	—	—	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641	3	
	7	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703	2	
	8	—	—	—	—	—	—	—	—	—	—	—	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176	1	
	9	—	—	—	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020	0	9
10	0	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010	10	
	1	0.0914	0.1667	0.2281	0.2770	0.3151	0.3438	0.3643	0.3777	0.3851	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098	9	
	2	0.0042	0.0153	0.0317	0.0519	0.0746	0.0988	0.1234	0.1478	0.1714	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439	8	
	3	0.0001	0.0008	0.0026	0.0058	0.0105	0.0168	0.0248	0.0343	0.0452	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172	7	
	4	0.0000	0.0000	0.0001	0.0004	0.0010	0.0019	0.0033	0.0052	0.0078	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051	6	
	5	—	—	0.0000	0.0000	0.0001	0.0001	0.0003	0.0005	0.0009	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461	5	
	6	—	—	—	—	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051	4	
	7	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172	3	
	8	—	—	—	—	—	—	—	—	—	—	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439	2	
	9	—	—	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098	1	
	10	—	—	—	—	—	—	—	—	—	—	—	—	—	0.0000	0.0000	0.0001	0.0003	0.0010	0	10
n	***X***	**0.99**	**0.98**	**0.97**	**0.96**	**0.95**	**0.94**	**0.93**	**0.92**	**0.91**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**	***X***	***n***

continued

TABLE E.6

Table of Binomial Probabilities (*Continued*)

n	X	π 0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	X	n
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	—	20	
	1	0.1652	0.2725	0.3364	0.3683	0.3774	0.3703	0.3526	0.3282	0.3000	0.2702	0.1368	0.0576	0.0211	0.0068	0.0020	0.0005	0.0001	0.0000	19	
	2	0.0159	0.0528	0.0988	0.1458	0.1887	0.2246	0.2521	0.2711	0.2818	0.2852	0.2293	0.1369	0.0699	0.0278	0.0100	0.0031	0.0008	0.0002	18	
	3	0.0010	0.0065	0.0183	0.0364	0.0596	0.0860	0.1139	0.1414	0.1672	0.1901	0.2428	0.2054	0.1339	0.0716	0.0323	0.0123	0.0040	0.0011	17	
	4	0.0000	0.0006	0.0024	0.0065	0.0133	0.0233	0.0364	0.0523	0.0703	0.0898	0.1821	0.2182	0.1897	0.1304	0.0738	0.0350	0.0139	0.0046	16	
	5		0.0000	0.0002	0.0009	0.0022	0.0048	0.0088	0.0145	0.0222	0.0319	0.1028	0.1746	0.2023	0.1789	0.1272	0.0746	0.0365	0.0148	15	
	6	—	—	0.0000	0.0001	0.0003	0.0008	0.0017	0.0032	0.0055	0.0089	0.0454	0.1091	0.1686	0.1916	0.1712	0.1244	0.0746	0.0370	14	
	7			—	0.0000	0.0000	0.0001	0.0002	0.0005	0.0011	0.0020	0.0160	0.0545	0.1124	0.1643	0.1844	0.1659	0.1221	0.0739	13	
	8						0.0000	0.0000	0.0001	0.0002	0.0004	0.0046	0.0222	0.0609	0.1144	0.1614	0.1797	0.1623	0.1201	12	
	9								0.0000	0.0000	0.0001	0.0011	0.0074	0.0271	0.0654	0.1158	0.1597	0.1771	0.1602	11	
	10										0.0000	0.0002	0.0020	0.0099	0.0308	0.0686	0.1171	0.1593	0.1762	10	
	11											0.0000	0.0005	0.0030	0.0120	0.0336	0.0710	0.1185	0.1602	9	
	12												0.0001	0.0008	0.0039	0.0136	0.0355	0.0727	0.1201	8	
	13												0.0000	0.0002	0.0010	0.0045	0.0146	0.0366	0.0739	7	
	14													0.0000	0.0002	0.0012	0.0049	0.0150	0.0370	6	
	15														0.0000	0.0003	0.0013	0.0049	0.0148	5	
	16															0.0000	0.0003	0.0013	0.0046	4	
	17																0.0000	0.0002	0.0011	3	
	18																	0.0000	0.0002	2	
	19																		0.0000	1	
	20																			0	20
n	X	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91	0.90	0.85	0.80	0.75	0.70	0.65	0.60	0.55	0.50	X	n

6.8.3 Mean and variance of a binomial distribution

mean is given by

$$\mu = np$$

the variance is given by :

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

and the standard deviation

$$\sigma = np\sqrt{(1-p)}$$

We are sure that you appreciate the benefits in the availability of a table of calculated values for the rather complex formula of the binomial probability distribution. Just make sure that you know how to use Table 1. If you do not understand the letter "*k*" in the table, it heads the column with the values that the variable can assume. Go to the page with $n = 20$. In the first row you have a p and the values for p are given in the second row. Choose $p = 0.25$. Do you see the value 0.25 in the second row? That 0.25 forms the top value of a column beneath it. The value that you will use for your sum depends on the value of k (first column) that you need for the particular question. Suppose that you had to find $P(X \leq 3)$, you use the value 0.2252, which is where the column under $p = 0.25$ intersects with the row to the right of $k = 3$. If you had to find $P(X \leq 6)$ the answer would have been 0.7858. Hope you understand now that the k-values 1, 2, 3, ..., 19 directs you to the answer according to the particular value of the variable in your question.

To help you, here are some examples:

Statement	Number of successes	Values for Table
P (exactly 3)	$x = 3$	$P(x \leq 3) - P(x \leq 2)$
P (at most 3)	$x = 0, 1, 2, 3$	$P(x \leq 3)$
P(at least 3)	$x = 3, 4, 5...$	$1 - P(x \leq 2)$
P(minimum 5)	$x = 5, 6, 7...$	$1 - P(x \leq 4)$
P(between 2 and 5)	$x = 3, 4$	$P(x \leq 4) - P(x \leq 2)$
P(from 2 to 5)	$x = 2, 3, 4, 5$	$P(x \leq 5) - P(x \leq 1)$
P(greater than 3)	$x = 4, 5, 6.....$	$1 - P(x \leq 3)$
P(less than 3)	$x = 2, 1, 0$	$P(x \leq 2)$
P(maximum 4)	$x = 0, 1, 2, 3, 4$	$P(x \leq 4)$
$P(2 \leq x \leq 6)$	$x = 2, 3, 4, 5, 6$	$P(x \leq 6) - P(x \leq 1)$
$P(2 < x < 6)$	$x = 3, 4, 5$	$P(x \leq 5) - P(x \leq 2)$
$P(2 \leq x < 6)$	$x = 2, 3, 4, 5$	$P(x \leq 5) - P(x \leq 1)$

The formulas for the mean, variance and standard deviation of the Binomial distribution are extremely important and usually very simple in its applications.

6.8.4 Examples

Question 1

Twenty-five percent of the students in an English class of 100 are international students. The standard deviation of this binomial distribution is

(1) 25

(2) 2.24

(3) 10

(4) 18.75

(5) 4.33

Question 2

Which of the following about the binomial distribution is a *true statement*?

(1) The random variable X is continuous.

(2) The probability of success p is stable from trial to trial.

(3) The number of trials n must be at least 30.

(4) The results of one trial are dependent on the results of the other trials.

(5) Sampling in the different trials are done without replacement.

Question 3

If $n = 10$ and $p = 0.60$, then the mean of the binomial distribution is

(1) 0.06

(2) 2.40

(3) 6.00

(4) 5.76

(5) 1.55

Question 4

A multiple-choice test has 25 questions. There are 4 choices for each question. A student who has not studied for the test decides to answer all questions randomly. The probability that this student gets *at least* 5 answers correct in the test is equal to

(1) 0.7863

(2) 0.3783

(3) 0.6167

(4) 0.4207

(5) 0.2137

Solutions:

Question 1

Answer: 5 $\left(\sqrt{100 \cdot 0.25 \cdot 0.75}\right)$

Question 2

Answer: 2

Question 3

Answer: 3 $(10 \cdot 0.60)$

Question 4

Answer: 1 $(n = 25,\ p = 0.25$ and $x = 4)$ because “at least 5" means “5 and more”.

Then $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.2137 = 0.7863.$

6.9 Poisson distribution

A Poisson distribution is a discrete probability distribution used to calculate the probability that a number of success ($X = 0, 1, 2, ..., n$) will occur over a specific period, interval or area. A Poisson random variable is the number of occurrences of events called success. The sample size n is not defined for a Poisson random variable.

Examples of Poisson random variables are:

1. Number of customers arriving at the bank per day. (the interval is one day)
2. Number of accidents per hour in a mine.

6.9.1 Characteristics the Poisson distribution

1. The number of successes that occur in any period of time or interval is independent of the number of successes that occur in any other period or interval.
2. The probability of a success in a period or interval is the same for all equal size intervals.
3. The probability of a success in an interval or period is proportional to the size of the interval.
4. The probability of more than one success in an interval or period approaches 0 as the interval becomes smaller.

The probability that a Poisson random variable can take in a specific interval is :

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!} \quad \text{for} \quad x = 0, 1, 2...$$

where μ

is the mean number of successes in the interval or period and e is the base of the natural logarithm.

6.9.2 Mean

The mean of a Poisson random variable is equal to its variance

$$\mu = \sigma^2$$

6.9.3 Example

Suppose that the number of complaints which the mayor of a city receives per day is a random variable having a Poisson distribution with a mean of 4. What is the probability that the mayor will received no complaints per day?

$$P(X = 4) = \frac{e^{-4}4^{0}}{0!} = 0.0183$$

Poisson Table

TABLE 2 Poisson Probabilities

Tabulated values are $P(X \leq k) = \sum_{x=0}^{k} p(xi)$. (Values are rounded to four decimal places.)

	μ															
k	0.10	0.20	0.30	0.40	0.50	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067	0.0041	0.0025
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404	0.0266	0.0174
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247	0.0884	0.0620
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650	0.2017	0.1512
4		1.0000	1.0000	0.9999	0.9998	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405	0.3575	0.2851
5				1.0000	1.0000	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160	0.5289	0.4457
6						0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622	0.6860	0.6063
7						1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666	0.8095	0.7440
8							1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319	0.8944	0.8472
9								1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682	0.9462	0.9161
10									0.9999	0.9997	0.9990	0.9972	0.9933	0.9863	0.9747	0.9574
11									1.0000	0.9999	0.9997	0.9991	0.9976	0.9945	0.9890	0.9799
12										1.0000	0.9999	0.9997	0.9992	0.9980	0.9955	0.9912
13											1.0000	0.9999	0.9997	0.9993	0.9983	0.9964
14												1.0000	0.9999	0.9998	0.9994	0.9986
15													1.0000	0.9999	0.9998	0.9995
16														1.0000	0.9999	0.9998
17															1.0000	0.9999
18																1.0000
19																
20																

The Poisson table make it easier and simple to compute the Poisson probabilities for individuals value of x as well as cumulative and related probabilities.

Simeon Poisson developed a method for measuring the probability that a certain number of successes will occur over a specified period of time or space. Although time (usually a continuous variable) is considered, it is still a discrete probability distribution, because *the number of successes* within that time period is counted. If the number of trials in a binomial experiment are difficult to count, there is a need for this type of distribution.

Do you understand the similarities and differences between the binomial and the Poisson distributions?

- They are both discrete distributions.
- For application of the Poisson distribution the value of p must be close to zero, say $p \leq 0.05$. If p takes on values close to 0.5, applications of the binomial distribution give good results.
- Independence of events are essential for both distributions.

- The binomial distribution has 2 parameters, n (number of trials) and p (probability of a success), while the Poisson has only one parameter μ $(= np)$, the mean number of successes in the interval, which is also the variance for the Poisson.

  {Only if you are interested: If p is a very small value, $(1 - p)$ will be almost equal to one, e.g. if $p = 0.01 \Longrightarrow 1 - p = 0.99$ and then $np(1 - p) \approxeq np(1) = np$.}

- For large numbers of trials n , say $n \geqslant 30$, it is more practical to use the Poisson model than the binomial in computations, provided that the other characteristics of the Poisson are satisfied. In fact, for the Poisson distribution, the values the discrete random variable x can assume have no upper limit. Possible values of x are: 0, 1, 2, 3, ...

The formula for the Poisson probability distribution may look rather complex, but fortunately Table2 at the back of your book contains the calculated answers. It is, as for the binomial, a table with *cumulative* values compiled for different values of the two entries, μ (the parameter of the distribution) and x (the value of the random variable, again indicated by k in the table). The whole table is spread over 2 pages, which implies that not all possible values are listed.

Suppose we have to find the probability of "not more than 3 occurrences" (i.e. $x \leq 3$), for a Poisson random variable if $\mu = 1.5$. Locate $\mu = 1.5$ as the top element of the column we will use. In the first column, locate the value 3 and move to the right in this row. The intersection of the column under $\mu = 1.5$ and $k = 3$ contains your answer of 0.9344 for $P(X \leq 3)$. Of course, if you are interested, you can find these values using the computer.

For the curious student (only as a matter of interest):

Have you heard of the **hypergeometric distribution?** It is also a discrete distribution function, very similar to the binomial distribution because the interest is also in the number of 'successes' in *n fixed trials*. There is, however, one difference between the binomial and the hypergeometric distributions:
Binomial: The probability of a success must remain constant from one trial to the next and the different outcomes are therefore *independent*.
Hypergeometric: The trials are *dependent* because the outcome of one observation is affected by the outcomes of previous observations. It will correspond to sampling *without replacement.* In such a sampling method, each consecutive draw is made from a smaller population, since the previous 'outcome' is not replaced.

Key terms

random variable

discrete random variable

continuous random variable

probability distribution

expected value

variance

standard deviation

binomial experiment

binomial random variable

binomial probability distribution

cumulative probability

Poisson experiment

Poisson random variable

Poisson probability distribution

6.10 Learning Outcomes

Read through the chapter summary and make sure that you have mastered the knowledge and understand the different concepts.

Can you

- distinguish between a discrete and continuous random variable?
- explain the requirements for a distribution of a discrete random variable?
- describe the population, its probability distribution and the role of the population parameters?
- use the laws of expected value and variance?
- distinguish between univariate and multivariate distributions?
- describe the marginal probabilities, the parameters and covariance of bivariate distributions?
- define a binomial experiment and the binomial probability distribution?
- use the binomial table?
- define the Poisson experiment and the Poisson distribution?
- use the Poisson table?

6.11 Study Unit 6: Summary

I. A *random variable* is a numerical value determined by the outcome of an experiment.

II. A *probability distribution* is a listing of all possible outcomes of an experiment and the probability associated with each outcome.

   A. A *discrete probability distribution* can assume only certain values. The main features are:

      1. The sum of the probabilities is 1.00.
      2. The probability of a particular outcome is between 0.00 and 1.00.
      3. The outcomes are mutually exclusive.

   B. A *continuous distribution* can assume an infinite number of values within a specific range.

III. The *mean* and *variance* of a probability distribution are computed as follows.

   A. The mean is equal to: $\mu = \Sigma [x P(x)]$

   B. The variance is equal to: $\sigma^2 = \Sigma \left[(x - \mu)^2 P(x)\right]$

IV. The *binomial distribution* has the following characteristics:

A. Each outcome is classified into one of two mutually exclusive categories.

B. The distribution results from a count of the number of successes in a fixed number of trials.

C. The probability of a success remains the same from trial to trial.

D. Each trial is independent.

E. A binomial probability is determined as follows:

$$P(x) = {}_nC_x \pi^x (1-\pi)^{n-x}$$

F. The mean is computed as $\mu = n\pi$

G. The variance is $\sigma^2 = n\pi(1-\pi)$

V. The *Poisson distribution* has the following characteristics.

A. It describes the number of times some event occurs during a specific interval.

B. A Poisson probability is determined from the following equation:

$$P(X) = \frac{\mu^x e^{-\mu}}{x!}$$

C. The mean and the variance are the same.

References

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Weiers, RM (2005). *Introduction to Business Statistics.* Thompson Learning, Inc.

Study Unit 7

CONTINUOUS PROBABILITY DISTRIBUTIONS

7.1 Learning outcomes

After completing this unit, you should be able to

1. explain the difference between probability distribution and probability density function
2. calculate the normal probabilities using normal distribution table
3. describe the continuous normal variable and its parameters
4. transform all normal distributions into standard normal distribution
5. use the normal distribution table to determine probabilities as well as to determine values of the normal random variable
6. to understand the exponential distribution and its link to the Poisson distribution

7.2 Introduction

In unit 6 we discussed discrete random variables and their Probability Distributions, with special examples in the form of the binomial and the Poisson distributions. Now, as you can read in the introduction given by Keller, we move to continuous random variables and their Probability Density Functions. Do you notice that the distributions have different names?

Random variable	Probability function: description
Discrete	*Probability Distribution*
Continuous	*Probability Density Function*
Discrete	*Probability Distribution* links population with a sample of *nominal data*
Continuous	*Probability Density Function* links population with a sample of *interval data*
Discrete (distribution function)	P*robabilities expressed by the height of vertical bars in a histogram*
Continuous (density function)	P*robabilities expressed by the area under a smooth curve*

Imagine in your mind a subtle change-over from *very large* countable values (discrete) to uncountable values (continuous). Please keep this comment at the back of your mind while we initially concentrate on the very distinct differences between discrete and continuous random variables. It is only later that you will learn about certain very important links between the two types of random variables and their distributions.

We give most of our attention this year to the normal distribution, as it is the most important continuous density function in the study of statistics and also in its applications to different natural and economic phenomena. As you are introduced to the "normality" of the normal curve, this comment will become clear. Another very important distribution, especially in business applications, is the exponential distribution. The other distributions discussed in this chapter are Student's t-distribution, the chi-square distribution and the F-distribution. In this section we only want you to take note of their existence and their most important features. Student's t and chi-square distributions are discussed in other sections of the module. With the statistical software currently available the use and applications of statistical distributions has become easy, but please do not fall into the trap and be uninformed about the basic characteristics of the different distributions.

7.3 Probability density function

A *continuous random variable* is a random variable that can take an uncountable number of values. Continuous random variables are different from the discrete random variables, we are unable to list the possible values because there is an infinite number of them and the probability of each individually value is virtually 0. As such, we can determine the probability of only a range of values.

In the first part of this chapter the main idea is to highlight the differences between discrete and continuous random variables and their respective distributions. Do you not think that the explanation to move from a dis-

crete histogram to a continuous curve is easy to understand? Smoothing the edges is easy to comprehend and once you let your imagination take flight, you can see a smooth curve in your mind's eye! In the continuous case, probabilities are expressed by the *area under a smooth curve.*

7.4 Requirement of a probability density function

The requirements for a probability density function are very important.

Statements and explanations	
$a \leq x \leq b$	The variable x cannot be isolated, but is somewhere in a specified interval which has the values a and b as borders.
$f(x)$	Substitute for some mathematical formula which describes the movement of the density function of the continuous random variable.
$f(x) \geq 0$	The curve $f(x)$ itself may never assume a negative value. The normal curve is always presented *above a given horizontal line* on which the values of the random variable are indicated. This indicates that all values of $f(x)$ are positive.
$a \leq x \leq b$	Note that there is no specification that these values a and b, which indicate the values of x, must always be positive. Both can be any value from $-\infty$ to $+\infty$.
Area from a to b is 1.0	Remember that a and b are the borders for the x-values; all possible outcomes fall within that interval. The sum of all probabilities must always be 1.0, which is still true, even though the individual probabilities cannot be "separated". They are so close together that they form an area, which must be equal to 1.0.

7.5 Uniform distribution

To explain how to find the area under the curve that describes a probability density function, consider the uniform probability distribution, also called the rectangular probability distribution.

Uniform probability density function

The Uniform distribution is described by the function:

$$f(x) = \frac{1}{b-a} \quad \text{where} \quad a \leq x \leq b$$

The uniform distribution can serve as a "bridge" to carry the concept of a discrete distribution to that of a continuous distribution. Suppose you flip a coin, where your random variable is discrete. There are two

possible outcomes: Heads (H) and Tails (T) and the probability of each outcome is $\frac{1}{2}$. You can think about this discrete distribution as a uniform distribution, because the probability does not change from the one outcome to the other. The sum of probabilities would be $\frac{1}{2} + \frac{1}{2} = 1.0$.

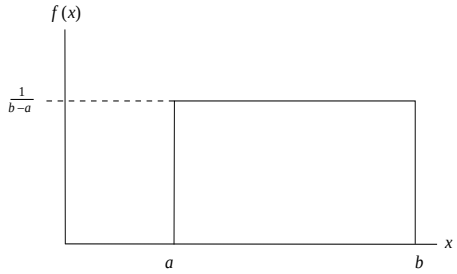


Figure 7.1 Uniform distribution

To calculate the probability of any interval, we need to find the area under the curve. For example, to find the probability that X falls between x_1 and x_2 is to determine the area in the rectangle whose base is $x_2 - x_1$ and whose heights is $\frac{1}{b-a}$.

The figure 8.2 depicts the area we wish to find which is a rectangle and the area of the rectangle is found by multiplying the base times the height.

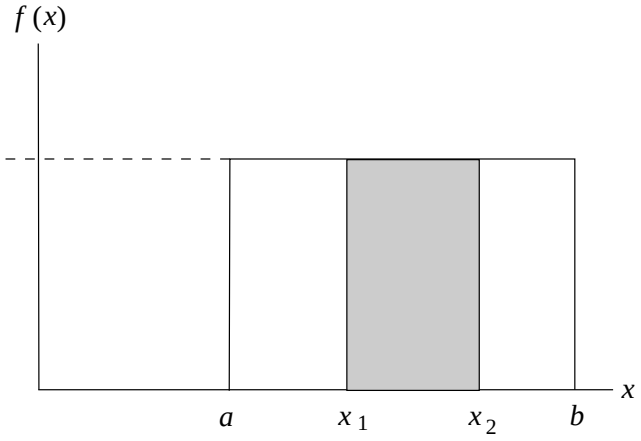


Figure 7.2 $P\ (x_1 < X < x_2)$

Thus, $P(x_1 < X < x_2) = \text{Base} \times \text{Height} = (x_1 - x_2)X\frac{1}{b-a}$

7.5.1 Example

The amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2000 gallons and a minimum of 5000 gallons.

(a) Find the probability that daily sales will fall between 2500 and 3000 gallons.

(b) What is the probability that the service station will sell at least 4000 gallons?

(c) What is the probability that the station will sell exactly 2500 gallons?

Solution:

The probability density function is:

$$f(x) = \frac{1}{5000 - 2000} = \frac{1}{3000} \quad 2000 \leq x \leq 5000$$

(a) The probability that X falls between 2500 and 3000

$$P(2500 \leq x \leq 3000) = (3000 - 2500)\frac{1}{3000} = 0.1667$$

(b) $P(X \geq 4000) = (5000 - 4000)\frac{1}{3000} = 0.3333$

(c) $P(X = 2500) = 0$

Because there is an uncountable infinite number of values X, the probability of each individual value is zero.

Two vertical bars, representing these probabilities, would have the same height. In theory, if one could increase the possible outcomes (but still all with equal probabilities of occurrence), these vertical bars would be so close to one another that they form a block or rectangle. This can now be compared to the situation of a continuous random variable with all possible outcomes within a specified interval $[a, b]$ having the same probability of occurrence. The total of all probabilities then becomes the area of the rectangle. Suppose that all the probabilities are equal to 0.1, the length of the interval $[a, b]$ will have to be 10 so that $0.1 \cdot 10 = 1.0$. (Of course, because the sum of all probabilities must equal 1.0!)

In Example 7.5.1, $P(X = 2,500) = 0$. Why? One single variable value would be indicated by a single vertical line and what is the area of a vertical line? Area is (Base $\cdot$ Height) and for a vertical line, Base $= 0$. Apply your knowledge that (anything) $\cdot\, 0 = 0$, then it is easy to understand!

Have you heard people saying that *power lies in numbers*? In its reference to humans, history has given this comment true meaning in many revolutions where the masses simply used their numbers to enforce change. In statistics numbers also have a lot of power!

- If the number of values of a discrete variable increases enough, it is fine to ignore the influence of sampling without replacement and treat the sampling action as if sampling did take place with replacement. This means that if the number of observations is large enough, a binomial distribution can still be fitted even if the requirement of independence of the trials is relaxed. (Then the binomial and the hypergeometric distributions become the same.)

- If n in a binomial distribution is large and p is a small enough, the Poisson distribution can be used to approximate the binomial distribution. (Both are discrete distributions.)
- If n in a binomial distribution is large and p is not small, the normal distribution can be used to approximate the binomial distribution. This means that, as long as the number of values of a discrete variable increases enough, it is fine to treat the variable as if it is continuous. (Of course it does not become continuous!)

7.6 Normal distribution

The normal distribution is the most important of all probability distributions because of its crucial role in statistical inference.

Normal density function

The probability density function of a normal random variable is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

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

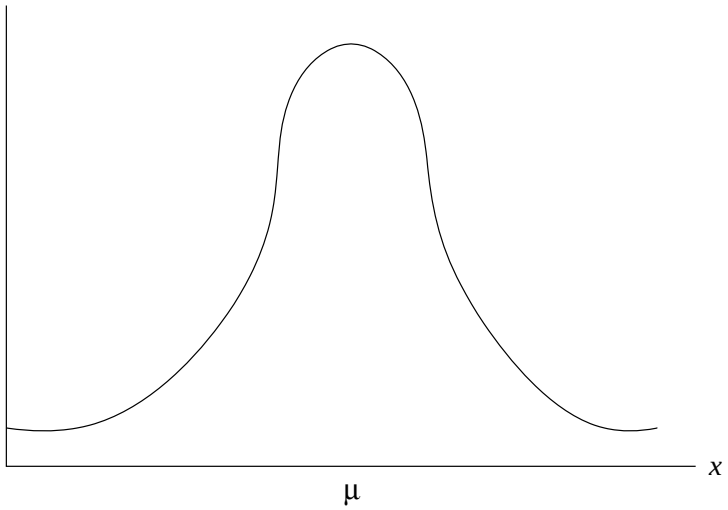


Figure 7.3 Normal Distribution

The normal distribution is the Big Mama of distributions for statistical inference!

- The curve itself is in the form of a bell. The two parameters of the normal distribution are the mean μ and the standard deviation σ.
- On the horizontal axis (the x-axis) the Greek letter μ indicates the value of the mean and this μ seems to be written at the middle of the possible x-values.

- If you *could* cut out this little bell itself (PLEASE DO NOT DO THAT!!!!!), it would be possible to fold the paper along this value of μ and the two halves of the bell will fall exactly on top of each other – that is why we say that it is a *symmetric* curve.
- The actual value of the mean μ is not indicated in this sketch, but keep in mind that it can even be a negative value.
- The comment that the x-values range from $-\infty$ to $+\infty$ is also not indicated in the sketch. The fact that the curve is given in a little "box" formed by the $x-$ and the $f(x)-$axes may confuse you. Rather look at figure 8.10, where the bell-shaped curve is given on only the x-axis. Now you can imagine that the values of x to the left of the mean can be any value down to $-\infty$ and on the right side x they can assume any positive value up to $+\infty$. The "tails" of the curve do not touch the x-axis to indicate that $-\infty \leq x \leq \infty$. (Some books actually have arrows on the tails to illustrate this principle.)
- There is not only one normal curve – the values of the mean μ and the standard deviation σ of each particular normal distribution will determine the shape of the bell and the placement on the number line. This is illustrated very well in Figures 8.8 and 8.9.

Influence of the

mean μ:	causes horizontal movement
standard deviation σ :	changes in the form of the "bell"

If you have a computer, you can play around with Applet 4.

7.6.1 Calculating normal probabilities

The density function of a normal random variable looks complex and when the values of μ and σ are substituted it involves a substantial amount of complex calculations. Fortunately, clever mathematicians realized that all normal distributions can be transformed in terms of an "easy" normal distribution with mean $(\mu) = 0$ and variance $(\sigma^2) = 1$, the so-called *standard normal distributio*n. The next aid was the composition of a table of calculated areas under this special normal curve with mean zero and standard deviation one. Have you given a thought to the fact that two sentences ago we said that *variance must be one* and in the next sentence we said that *standard deviation must be one*! You know that variance is not standard deviation, so how is this possible? It is because the square root of one is one: $\sqrt{1} = 1$. Never forget that standard deviation is the square root of the value of the variance. This exception is only true for the standard normal distribution because "one" is a special number.

Suppose you are given a question involving a certain normal random variable, say X, then you have to transform this variable X to fit into the mould of the standard normal distribution, where it is customary to call the standardized random variable Z. The formula you use for this is

$$Z = \frac{x - \mu}{\sigma}.$$

A few pointers if you use the normal table:

- Make sure to remember that the value in the denominator of this fraction is σ and not σ^2. Read the information carefully and make sure whether you were given standard deviation or variance. If variance was given, first find the square root of that numerical value for the formula.
- If you calculate $P(a < X < b)$, $P(a < Z < b)$, $P(Z < b)$,or $P(a < Z)$, you have to determine a *probability*, so make sure that your answer lies in the interval [0, 1]!
- Recall that the area under the normal curve is considered to be 1.0. So, if the total area from $-\infty$ to $+\infty$ is equal to 1.0, then the area from $-\infty$ to 0 is equal to 0.5 and the area from 0 to $+\infty$ is also equal to 0.5.
- Never hesitate, for each question, to quickly make a small rough sketch of a normal curve (no art work!) and shade the area applicable for that question. That is very helpful and even some lecturers do that!

7.6.2 Examples

Question 1

Given that Z is a standard normal random variable, $P(-1.0 \leq Z \leq 1.5)$ is

(1) 0.7745

(2) 0.8413

(3) 0.0919

(4) 0.9332

(5) 0.0994

Question 2

If Z is a standard normal random variable, then $P(-1.75 < Z < -1.25)$ is

(1) 0.1056

(2) 0.0655

(3) 0.0401

(4) 0.8543

(5) -0.0655

Question 3

Given that X is a normal variable, which of the following statements is/are true?

(1) The variable $X + 5$ is also normally distributed.

(2) The variable $X - 5$ is also normally distributed.

(3) The variable $5X$ is also normally distributed.

(4) None of the above.

(5) All of the above.

Question 4

Given that the random variable X is normally distributed with a mean of 80 and a variance of 100, $P(85 < X < 90)$ is

(1) 0.5328

(2) 0.2620

(3) 0.1915

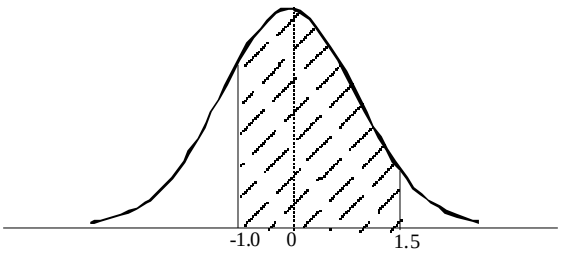
(4) 0.1498

(5) 0.0199

Solutions

Question 1

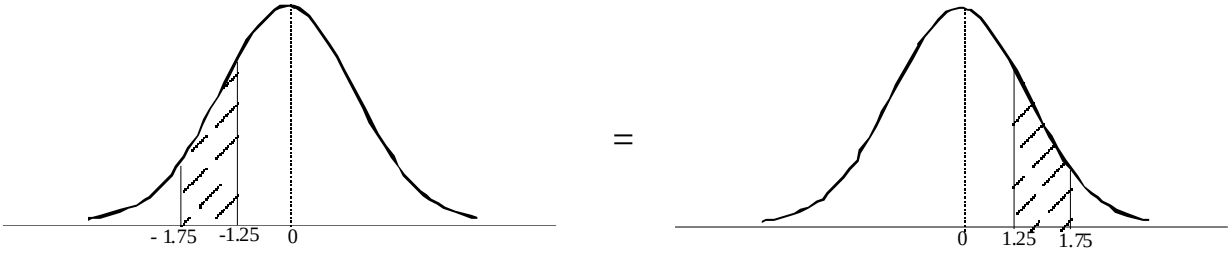
Answer: 1.



$$P(-1.0 \leq Z \leq 1.5) = 0.9332 - 0.1587$$
$$= 0.7745$$

Question 2

Answer: 2.



$$\begin{aligned} P(-1.75 \leq Z \leq -1.25) &= P(1.25 \leq Z \leq 1.75) \\ &= 0.9599 - 0.8944 \\ &= 0.0655 \end{aligned}$$

Question 3

Answer: 5.

Question 4

Answer: 4

$$\begin{aligned} P(85 \leq Z \leq 90) &= P(\frac{85-80}{10} \leq Z \leq \frac{90-80}{10}) \\ &\text{Note that you must use the } \textit{standard deviation} \\ &= P(0.5 \leq Z \leq 1.0) \\ &= 0.8413 - 0.6915 \\ &= 0.1498 \end{aligned}$$

Finding values of Z

In the questions discussed so far, you had to determine a probability for a given value of the random variable, but this can be turned around. A specific probability is given and it is your task to find the corresponding random variable. For this type of question, the same normal Table 3 is used, but in a "reverse" manner.

A few pointers:

- Do not hesitate, for each question, to make a small rough sketch of a normal curve and shade the given area. That gives you an idea of where the variable can be expected to fall.
- Read the question very carefully. The interpretation of $>$ is quite different from $<$. Make sure your little sketch is correct!

- Something interesting – it is not important *for a continuous variable* whether you write $\leq$ or $<$ in the probability (the probability that a continuous variable is equal to one specific value is equal to zero). The same applies to $\geq$ and $>$.
- Both examples 8.4 and 8.5 are about areas in the tail sections of the curve, but there are many different options for questions. See Activity 8.3 below.

7.6.3 Examples

Question 1

Given that Z is a standard normal variable, the value z for which $P(Z \leq z) = 0.6736$ is

(1) 0.94

(2) -0.45

(3) -0.94

(4) 0.45

(5) 0.4591

Question 2

Given that Z is a standard normal variable, the value z for which $P(Z > z) = 0.6736$ is

(1) 0.94

(2) -0.45

(3) -0.94

(4) 0.45

(5) 0.4591

Question 3

Given that Z is a standard normal variable, the value z for which $P(Z \leq z) = 0.2578$ is

(1) 0.700

(2) 0.65

(3) -0.65

(4) 0.242

(5) -0.700

Question 4

If Z is a standard normal random variable, then the value z for which $P(-z < Z < z)$ equals 0.8764 is

(1) 0.35

(2) 1.54

(3) 3.08

(4) 1.16

(5) -1.54

Question 5

If the z-value for a given value x of the random variable X is $z = 1.96$, and the distribution of X is normally distributed with a mean of 60 and a standard deviation of 6, to what x-value does this z-value correspond?

(1) 71.76

(2) 11.96

(3) 10.32

(4) 62.85

(5) 35.88

Question 6

If Z is a standard normal random variable, the area between $z = 0.0$ and $z = 1.50$ compared to the area between $z = 1.5$ and $z = 3.0$ will be

(1) the same

(2) larger

(3) smaller

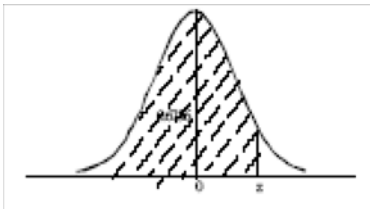
(4) impossible to predict

(5) none of the above

Solutions:

Question 1

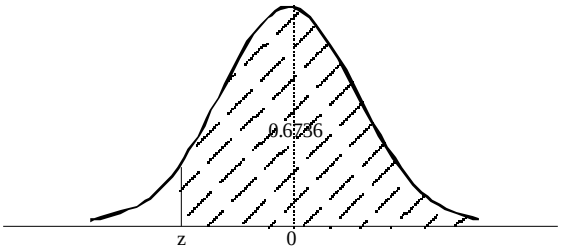
Answer: 4.



The Z-value that corresponds with of 0.6736 is $z = 0.45$. $P(Z \leq 0.45) = 0.6736$.

Question 2

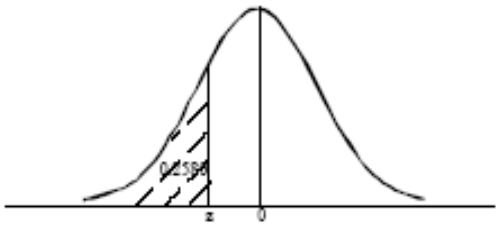
Answer: 2.



Be careful, $P(Z > z) = 0.6736$ implies that the z-value will be negative (look at the $>$ sign). The corresponding Z-value is again $z = 0.45$, but now this value must be negative 0.45, or -0.45.

Question 3

Answer: 3.

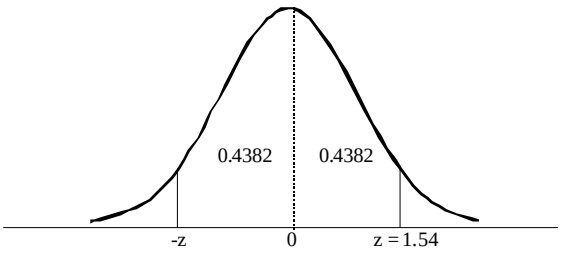


Did you have this one correct? If not, do not feel bad, as this was a rather difficult problem. You have to reason as follows:

The first to note is that the area of 0.2580 is less than 0.500, which implies, because of the $\leq$ -sign with the given probability, that the z-value *has to be negative.*

Question 4

Answer: 2.



This question is not so difficult, but tricky! The magic lies in that $-z \leq Z \leq z$ implies that the z's are numerically the same. They are at the same distance from $z = 0$, just on either side of it. What you have to do is to divide the area into two equal parts.

$$\frac{0.8764}{2} = 0.4382.$$

Add the area on the left side at the graph, then the area is $0.5 + 0.4382 = 0.9382$, which corresponds with $z = 1.54$.

Question 5

Answer: 1.

$$\frac{X - 60}{6} = 1.76$$

$$X - 60 = 11.76$$

$$X = 71.76$$

Question 6

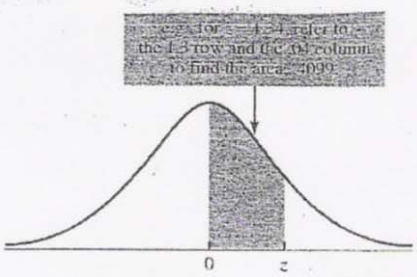
Answer: 2.

Think in terms of the form of the normal distribution. The bell shape makes the area close to the mean much bigger than the area in the tails.

Z_A and Percentiles

As you continue in your studies of statistics you will hear more and more about the normal distribution and its properties and possible applications. We would like to say more about the symmetry of the normal curve. The total area under the curve is equal to 1.00 and this can also be expressed as a percentage, namely 100%. As a result of the symmetry of the distribution we assume that 50% of the area lies on either side of the mean value. Assume a standard normal distribution and consider the following comparison:

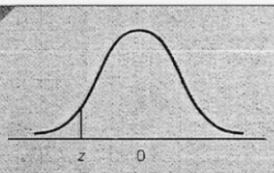
The Standard Normal Distribution



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4964	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Cumulative standard normal probabilites from $z = 0.00$ to $z = 3.09$, generated by Minitab, then rounded to four decimal places.

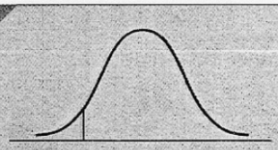
TABLE 3 Cumulative Standardized Normal Probabilities



$P(-\infty < Z < z)$

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

TABLE 3 Cumulative Standardized Normal Probabilities

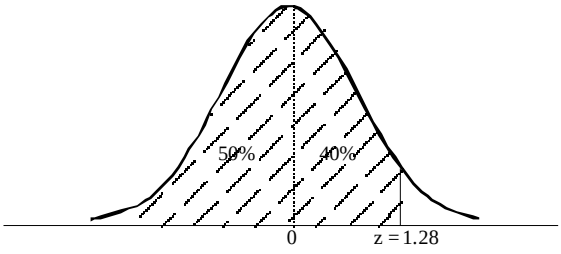


$P(-\infty < Z < z)$

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

7.6.4 Example

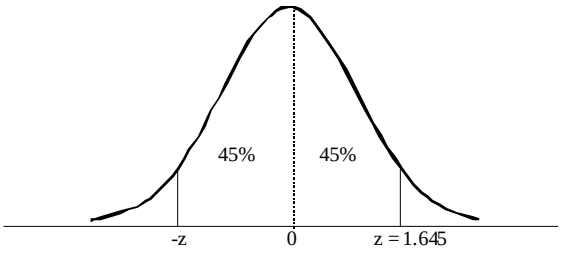
1. *Determine the value of the random variable if 90% of the area under the curve lies to the left of it.*



You use your knowledge that 50% of the area lies to the left of the mean. The other 40% of the area will have to lie to the right of the mean. From Table 3 the relevant z-value is 1.28 (rounded). *Percentiles* divide area in the same manner as in this example above and there are 100 percentiles. Each one can be represented as a Z-value of a standard normal random variable, with notation Z_A. Take note that this notation is such that

- $Z_{.01} = 2.33$ implies that 99% of the total area under the curve lies to the left of it
- $Z_{.05} = 1.645$ implies that 95% of the total area under the curve lies to the left of it
- $Z_{.10} = 1.28$ implies that 90% of the total area under the curve lies to the left of it
- $Z_{.15} = 1.04$ implies that 85% of the total area under the curve lies to the left of it, etc.

2. *Determine the value of the random variable z such that 90% of the area under the curve lies between $-z$ and $+z$.*



This is a different scenario. Because of symmetry the distance from the mean to the value $+z$ must be such that 45% of the 90% area lies to the right of the mean and the distance from the mean to the value $-z$ must be such that 45% of the 90% area lies to the left of the mean. The value in Table 3 which corresponds to an area of 0.9500 is 1.645 (rounded). z will therefore be 1.645 and $-z$ (of course) will be -1.645.

This type of interpretation of area under the normal curve is quite common in interval estimation and hypothesis testing and deserves your attention.

Statisticians realized that

- about 68.35% of the total area under any normal curve lies between $\mu - \sigma$ and $\mu + \sigma$
- about 95.5% of the total area under any normal curve lies between $\mu - 2\sigma$ and $\mu + 2\sigma$

$\multimap$ about 99.7% of the total area under any normal curve (almost all of it) lies between $\mu - 3\sigma$ and $\mu + 3\sigma$

Do you not find the two examples above interesting? Although both refer to 90% of the area, the questions differ and have different answers.

7.7 Other continuous probability distributions

7.7.1 The exponential distribution

The exponential distribution is the continuous counterpart of the discrete Poisson distribution.

Counting of the number of rare events within a period of time represents the discrete random variable for the Poisson distribution, while the continuous random variable of the exponential distribution considers the amount of time or distance between occurrences of these rare events. The mean of the exponential distribution is the same as its standard deviation and both are equal to the inverse of the mean of the Poisson distribution. This is provided that the mean of the Poisson is expressed as a rate.

In simple terms we say: Suppose the mean number of calls received at an emergency service is 10 per hour, then the mean of the corresponding exponential distribution will be $\frac{1}{10}$, or 0.1 hours between calls, which is every 6 minutes.

From the side of the Poisson distribution, it is better to express the 10 calls per hour as 10 calls per 60 minutes. If this is expressed as a ratio we say $\frac{10}{60}$ is the rate of calls per minute. If $\lambda = \frac{10}{60}$ calls per minute, then the mean time between calls for the exponential distribution will be the inverse of this, namely $\frac{1}{\lambda} = \frac{1}{\frac{10}{60}} = 6$, which is an average of 6 minutes between the consecutive calls. This link between the Poisson and exponential distributions is significant, as you will realise in later years when you study queuing theory.

7.7.2 Student's t-distribution

Please read through the little information given in this chapter. Here we do not discuss the use of the tables for the last three distributions. You should find my summary below helpful and interesting.

7.7.3 The Chi-square distribution

This skewed distribution is a sampling distribution, used e.g., in constructing confidence intervals and hypothesis tests where the interest is in the population variance. It is defined only for positive values of the variable and it has, as Student's t-distribution, the number of degrees of freedom υ as parameter.

7.7.4 The F-distribution

Again a skewed distribution used in the sampling distribution of the ratio of variances of samples from normal populations. It also plays an important role in the analysis of variance.

We give you the following comparison (not in depth) of the different distributions, for the sake of interest.

Binomial distribution Number of successes (p) in a set number (n) of trials. Discrete. 2 parameters: { mean (np); StD $[np(1-p)]$ }	►◄	{ **Poisson distribution** Number of successes in an interval of time. Discrete. 1 parameter: mean μ }
Poisson distribution Number of successes in an interval of time. Discrete distribution for rare events. 1 parameter: mean μ	►◄	{ **Exponential distribution** Time between occurrences of events Continuous distribution for rare events *One* parameter: mean (time) $= \frac{1}{\lambda}$ =StD
Standard normal distribution Symmetric, bell-shaped Most important probability distribution Two parameters: { mean $= 0$; $StD = 1$ }	►◄	{ **Student's t-distribution** Symmetric, mound shaped Extensively used in statistical inference One parameter υ { mean $= 0$; $StD = \frac{\upsilon}{\upsilon - 2} > 1$ } For υ large $\Longrightarrow$ standard normal distr. }
Standard Normal distribution Symmetric, bell-shaped Most important probability distribution Two parameters: { mean $= 0$; $StD = 1$ }	►◄	{ **Chi-square distribution** Positive skewed; only positive values Extensively used in statistical inference One parameter υ { mean $= \upsilon$; $StD = \sqrt{2\upsilon}$ } For $\upsilon > 100 \Longrightarrow$ a normal distribution }
Standard Normal distribution Symmetric, bell-shaped Most important probability distribution Two parameters: { mean $= 0$; $StD = 1$ }	►◄	{ **F-distribution** Positive skewed; only positive values Extensively used in statistical inference Two parameters υ_1 and υ_2. Approaches the normal distribution $\upsilon >>$ }

Key terms

probability density function

uniform distribution

rectangular distribution

normal distribution

normal random variable

standard normal random variable

exponential distribution

Student t-distribution

degrees of freedom

chi-squared distribution

F-distribution

7.8 Learning outcomes

Read through the chapter summary and make sure that you have mastered the knowledge and understand the different concepts.

- Can you explain the difference between probability distributions and probability density functions?
- Can you calculate normal probabilities with the normal table?
- Can you describe the continuous normal random variable and its parameters?
- Can you transform all normal distributions into the standard normal distribution?
- Can you use the normal table to determine probabilities as well as to determine values of the normal random variable?
- Do you understand the exponential distribution and its link to the Poisson distribution?
- Do you remember that there are other important continuous distributions that will be treated in detail in later years of statistical studies?

7.9 Study Unit 7: Summary

I. The *uniform distribution* is a continuous probability distribution with the following characteristics.

A. It is rectangular in shape.

B. The mean and the median are equal.

C. It is completely described by its minimum value a and its maximum value b.

D. It is also described by the following equation for the region from a to b:

$$P(x) = \frac{1}{b-a}$$

E. The mean and standard deviation of a uniform distribution are computed as follows:

$$\mu = \frac{(a+b)}{2}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

II. The *normal probability distribution* is a continuous distribution with the following characteristics.

A. It is bell-shaped and has a single peak at the centre of the distribution.

B. The distribution is symmetric.

C. It is asymptotic, meaning the curve approaches but never touches the X-axis.

D. It is completely described by its mean and standard deviation.

E. There is a family of normal probability distributions.

1. Another normal probability distribution is created when either the mean or the standard deviation changes.
2. The normal probability distribution is described by the following formula:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e - \left[\frac{(x-\mu)^2}{2\sigma^2}\right]$$

III. The *standard normal probability distribution* is a particular normal distribution.

A. It has a mean of 0 and a standard deviation of 1.

B. Any normal probability distribution can be converted to the standard normal probability distribution by the following formula:

$$z = \frac{X-\mu}{\sigma}$$

Reference

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 8

SAMPLING DISTRIBUTIONS

8.1 Learning outcomes

After completing this unit, you should be able to

1. formulate the *sampling distribution of the sample mean,* $\overline{X}$
2. define the *standard error of the sample mean,* $\overline{X}$
3. apply the *sampling distribution of the sample mean,* $\overline{X}$
4. compute probabilities for $\overline{X}$
5. formulate the *sampling distribution of the sample proportion* $\widehat{p}$
6. define the *standard error of the sample proportion* $\widehat{p}$
7. apply the *sampling distribution of the sample proportion* $\widehat{p}$
8. compute probabilities for $\widehat{p}$
9. formulate the *sampling distribution of the sample mean difference,* $\overline{X}_1 - \overline{X}_2$
10. define the *standard error of the sample mean difference,* $\overline{X}_1 - \overline{X}_2$
11. apply the *sampling distribution of the sample mean difference,* $\overline{X}_1 - \overline{X}_2$
12. compute probabilities for $\overline{X}_1 - \overline{X}_2$

8.2 Introduction

How will we link the information from a sample to a population? In this study unit you will learn that the **sampling distribution of a statistic** is the vehicle to move between the sample and the population.

- How do we *derive* the sampling distribution?

- How do we *apply* the sampling distribution?

Please note that the two expressions "**sample distribution**" and "**sampling distribution**" are quite different! In the previous units you learned about the **sample distribution**. You had a specific fixed data set called the sample which you had to describe or *organise*. This entailed that you *graphically displayed* the shape of the sample and you *computed measures.* Measures of locality are the mean, mode etc. and measures of dispersion are the range, standard deviation etc. All this was applied to real-life or empirical examples.

However, the **sampling distribution is something theoretical or non-empirical.** Sampling variability leads to sampling distributions which describes the long-run behaviour of a sample statistic when we draw sample after sample after sample. The laws of probability are applied when this theoretical sampling distribution is determined. In this study unit we study the sampling distributions of the sample mean, the sample proportion and the difference between two sample means.

8.3 Central limit theorem

The main goal of the sampling distribution is statistical inference and the sampling distribution is created by sampling. There are two ways to create a sampling distribution. The first is in actually drawing samples of the same size from a population, calculate the statistic such as the mean and then use descriptive techniques to learn more about the sampling distribution. The second approach uses the rules of probability and the laws of expected value and the variance to derive the sampling distribution. As the sample size n gets larger or increases, the sampling distribution of the sample mean becomes normally distributed.

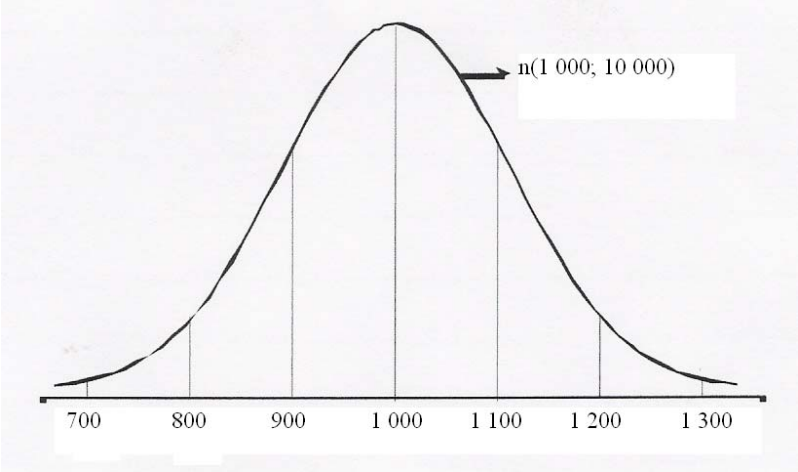
Trying to make intuitive sense of the concept "sampling distribution"

Do you recall that, according to convention, if the theoretical model for a variable X is a normal distribution with mean μ and variance σ^2 we write it as $X \sim n(\mu; \sigma^2)$? [**:-)** If you do not know this, make it part of your statistical vocabulary.]

Suppose that at another gas station the daily demand for regular gasoline is normally distributed with a mean of 1 000 gallons and a standard deviation of 100 gallons.

Thus, if X represents the daily demand for regular gasoline, according to convention, we write this as $X \sim n(1\,000; 10\,000)$. Since the two parameters are known, it means that we have a workable probability

distribution for which we may draw the following bell-shaped normal probability graph:



This is basically the same graph as in figure 8.10 of Keller except that we emphasise the values lying one, two or three standard deviations above and below the mean.

Remember that we want to get an idea of a sampling distribution which implies that as a first step we have to do some sampling!

8.4 Example

Suppose you have to obtain a random sample of size n = 10 from the normal population described above.

What do you expect the sample to be like? In other words write down ten possible daily demands for regular gasoline that could come from a $n(1\,000; 10\,000)$ distribution.

Solution:

We cannot give you immediate, direct feedback because there are many possible answers! However, we can give a lengthy discussion. Remember that anything is possible in sampling and that randomness makes the world interesting. Although we can never be certain how a single sample is going to turn out, we could none the less expect the sample to reflect the properties of a $n(1\,000; 10\,000)$ distribution. What does this mean?

Consider the following samples and decide which one will be the *most likely sample* to *reflect the properties of the population* from which it was drawn. (The sample values are ordered from small to large to make interpretation easier.)

Sample 1

750 810 850 900 910 920 950 980 990 1 000

Sample 2

1 000 1 010 1 020 1 050 1 080 1 090 1 100 1 150 1 190 1 250

Sample 3

750 780 800 820 900 1 100 1 180 1 190 1 210 1 250

Sample 4

750 850 970 990 1 000 1 010 1 030 1 060 1 150 1 250

Sample 5

770 830 920 970 990 1 010 1 030 1 070 1 160 1 230

Sample 1 is extremely unlikely to occur because we would expect 50% of the sample values to be above the mean $\mu = 1\,000$ (this follows from the symmetrical property of the normal distribution) and in this sample *they are all below or equal to the mean.*

Sample 2 is just as unlikely to occur for the same reason as above. Now *they are all above or equal to the mean* and we would expect 50% of the sample values to be below the mean $\mu = 1\,000$.

Although sample 3 has 50% of the sample values above the mean $\mu = 1\,000$ and 50% of the sample values below the mean, it does not reflect the *probability distribution properties* of the normal distribution. The "tail values" are dominating and we would expect most ($\pm 70\%$) of the sample values to lie within one standard deviation below and above the mean, i.e., between 900 and 1 000.

So, in real life, if we draw a random sample of size 10 from a $n(1\,000; 10\,000)$ population, *the most likely outcome will be something similar to either sample 4 or sample 5.*

Now you might wonder what all this has to do with the concept "sampling distribution". The actual important idea we would like to bring across is *the behaviour of the sample mean,* i.e., what happens when we compute the sample mean $\overline{X} = \frac{\Sigma X_i}{10}$ for repeated independent samples?

8.5 Example

Compute the means for samples **4 and 5** given in the feedback above.

Solution:

Please note that it is meaningless to compute the means for samples 1 and 2 since they were absurd examples and given only to make a point! Even for sample 3 (which was not a very likely sample) we find that the

opposite "tail values" *balance out* and that $\overline{X} = \frac{9\,980}{10} = 998$.

Sample 4

750 850 970 990 1 000 1 010 1 030 1 060 1 150 1 250

Sample 5

770 830 920 970 990 1 010 1 030 1 070 1 160 1 230

For sample 4 we find $\overline{X} = \frac{10\,060}{10} = 1\,006$.

For sample 5 we find $\overline{X} = \frac{9\,980}{10} = 998$.

What can we conclude from this?

The sample mean is not a fixed value but will *vary from sample to sample* and our "gut feeling" is that the values will vary in the vicinity of $\mu = 1\,000$. If we really intend to derive the distribution of $\overline{X}$ empirically, we would have to repeat this process over and over, i.e., draw an extremely large number of random samples (say $N = 500$) each of size $n = 10$ and compute the value of $\overline{X}$ for each sample. This means that we treat $\overline{X}$ as a brand new random variable and consider $N = 500$ as a sample of observations of $\overline{X}$-values. Applying the descriptive techniques of the earlier study units we could compute the mean and variance for this sample and also draw a histogram or relative frequency histogram to give us an idea (approximation) of the theoretical probability distribution of $\overline{X}$. This approach involves horrendous work and effort and it can never give the exact distribution! The second method to find the sampling distribution of $\overline{X}$ relies on the rules of probability and the laws of expected value and variance to derive expressions for these parameters. Keller gives us a feeling of this method by deriving the **Sampling Distribution of the Mean for Two Dice** but this derivation is still not a proper theoretical derivation which falls beyond the scope of this course.

Back to our intuitive approach, where you almost have to imagine an extremely large number (say $N = 500$) of random samples. Taking the average of all possible means (over an infinite number of samples) we would expect this value to be equal to $\mu = 1\,000$.

What will happen if we increase the sample size? Imagine a large number of random samples, each of size $n = 1\,000$. The sample means will still not be a fixed value and they will still *vary from sample to sample* but our gut feeling is that these values will lie closer to each other (in the vicinity of $\mu = 1\,000$) and vary less than in the case where the sample size is $n = 10$. So, the larger the sample size the closer $\overline{X}$ will vary around $\mu = 1\,000$.

In general, the variability of $\overline{X}$ is a function of the variability of the original variable X and of n, the sample size.

Are you comfortable with the following conclusion?

The mean of all possible sample means, denoted by $\mu_{\overline{X}}$, is exactly equal to the population mean μ and the variance of all possible sample means, denoted by $\sigma^2_{\overline{X}}$, is equal to the population variance divided by the sample size n.

What happens if we **cannot assume** that the population has a **normal** distribution? We employ the Central Limit Theorem.

8.6 Sampling distribution of the mean

The sampling distribution of the mean of a random sample drawn from any population is approximately normal for a sufficiently large sample size. The larger the sample size, the more closely the sampling distribution of X will resemble a normal distribution.

Sampling Distribution of the Sample Mean

1. $\mu_{\overline{x}} = \mu$

2. $\sigma^2_{\overline{x}} = \dfrac{\sigma^2}{n}$ and $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

3. If X is normal, $\overline{X}$ is normal. If X is non-normal, $\overline{X}$ is approximately normal for sufficiently large sample sizes. The definition of "sufficiently large" depends on the extent of non-normality of X.

8.6.1 Examples

Question 1

Suppose that the actual size of computer chips is normally distributed with a mean of 1 cm and a standard deviation of 0.1 cm. A random sample of 15 computer chips is taken. What is the standard error for the sample mean?

Solution:

$$\begin{aligned} \sigma_{\overline{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{0.1}{\sqrt{15}} \\ &\approx 0.025\,8 \end{aligned}$$

Question 2

A population that consists of 500 observations has a mean of 40 and a standard deviation of 15. A sample of size 100 is taken at random from this population. What is the standard error for the sample mean?

Solution:

$$\begin{aligned}\sigma_{\overline{x}} &= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \\ &= \frac{15}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} \\ &\approx 1.343\end{aligned}$$

Question 3

An infinite population has a mean of 70 and a standard deviation of 6. A sample of 50 observations will be taken at random from this population. What is the probability that the sample mean will be between 68.5 and 72?

Solution:

We first need to find

1. $\mu_{\overline{x}} = \mu = 70$
2. $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.848\,53$

$$\begin{aligned}P(68.5 < \overline{X} < 72) &= P\left(\frac{68.5-70}{0.848\,53} < \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{x}}} < \frac{72-70}{0.848\,53}\right) \\ &= P(-1.767\,76 < Z_{\overline{X}} < 2.357\,02) \\ &= P(-1.77 < Z_{\overline{X}} < 2.36) \\ &= 0.990\,9 - 0.038\,4 \qquad \text{(using Table 3, Keller)} \\ &= 0.952\,5\end{aligned}$$

The problem with an example such as activity 9.4 is that it does not reflect real-life values. For the ease of computation and with the purpose of mastering the technique we stated: “An infinite population has a mean of 70 and a standard deviation of 6.” In real-life situations these parameter values are usually not nice integers, or even worse, the information is unknown.

8.6.2 Examples

Reading Books

Assume that academic personnel read an average of 3.12 books in their recess period, with a standard deviation of 2.15 books. A researcher conducted a survey on this university campus for a sample of 64 academic personnel.

Solution:

We first need to find:

1. $\mu_{\overline{x}} = \mu = 3.12$
2. $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2.15}{\sqrt{64}} = 0.268\,75$

Question 1

Determine the probability that the sample mean is above 3.45 books.

Solution:

$$
\begin{aligned}
P(\overline{X} > 3.45) &= P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{x}}} > \frac{3.45 - 3.12}{0.268\,75}\right) \\
&= P(Z_{\overline{X}} > 1.227\,91) \\
&= P(Z_{\overline{X}} > 1.23) \\
&= 1 - P(Z_{\overline{X}} < 1.23) \\
&= 1 - 0.8907 \qquad \text{(using Table 3, Keller)} \\
&= 0.1093
\end{aligned}
$$

Question 2

Determine the probability that the sample mean is between 3.38 and 3.58 books.

Solution:

$$\begin{aligned} P(3.38 < \overline{X} < 3.58) &= P\left(\frac{3.38-3.12}{0.268\,75} < \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{x}}} < \frac{3.58-3.12}{0.268\,75}\right) \\ &= P(0.967\,44 < Z_{\overline{X}} < 1.711\,63) \\ &= P(0.97 < Z_{\overline{X}} < 1.71) \\ &= 0.956\,4 - 0.834\,0 \qquad \text{(using Table 3, Keller)} \\ &= 0.1224 \end{aligned}$$

Question 3

Determine the probability that the sample mean is below 2.94 books.

Solutions:

$$\begin{aligned} P(\overline{X} < 2.94) &= P\left(\frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{x}}} < \frac{2.94-3.12}{0.268\,75}\right) \\ &= P(Z_{\overline{X}} < -0.669\,77) \\ &= P(Z_{\overline{X}} < -0.67) \\ &= 0.2514 \qquad \text{(using Table 3, Keller)} \end{aligned}$$

8.7 Sampling distribution of a proportion

Remember that a binomial distribution has 2 parameters, p and $1-p$. To compute the binomial probabilities, we made an assumptions on p, p was given (known). However in real situation p is unknown, requiring that we estimate the value of p from a sample. The estimator of a population proportion of successes is the sample proportion given by:

$$\hat{p} = \frac{X}{n}$$

using the laws of expected value and variance, we derived the parameters of sampling distribution of a sample proportion as follows:

If X represents a binomial variable, in other words X is the number of successes in n independent Bernoulli trials with p the probability of a success, according to convention, we write this as $X \sim b(n;\ p)$.

The shorthand notation for the *normal approximation to the binomial distribution* is

$X \sim b(n;p) \approx Y \sim n(np;np(1-p))$.

8.7.1 Example

First revise section 7.4 (BINOMIAL DISTRIBUTION) to refresh what you know about the binomial distribution, before you work through **Normal Approximation to the Binomial Distribution** and reflect upon the following questions:

(a) Define *the sampling distribution* of $\widehat{p}$.

(b) How could you check to ensure that the approximation will be accurate?

(c) What is the correction for continuity and how will you apply it?

(d) What is the relationship between the sampling distribution of x and the sampling distribution of $\hat{p} = \frac{x}{n}$ and how will it affect the correction for continuity?

Solution:

(a) If p is the proportion of a population that has a certain attribute and we draw a random sample of size n from this population then $\widehat{p} = \frac{x}{n}$ (= proportion in the sample having this certain attribute) has an approximate normal distribution with mean, $E(\widehat{p}) = p$ and variance, $var(\widehat{p}) = \dfrac{p(1-p)}{n}$. We call this *the sampling distribution* of $\widehat{p}$.

(b) If we want to use the normal approximation to the binomial distribution to calculate probabilities associated with x, the number of successes in n trials, **the rule of thumb** is usually that $np > 5$ and $n(1-p) > 5$ in order to ensure the accuracy of the approximation.

(c) A correction for continuity can be used to make the approximation more accurate. Hence, the standard normal random variable used in section 9.2, Keller, had the approximation

$$z = \frac{\left(x \pm \frac{1}{2}\right) - np}{\sqrt{np(1-p)}}.$$

This means the expression $P(X = x)$ is approximated by $P(x - 0.5 \leq Y \leq x + 0.5)$ which is standardised to a z-distribution.

(d) The sampling distributions of x and $\hat{p}$ are equivalent. If we divide both numerator and denominator of the fraction above by n, we see the relationship between the sampling distribution of x and $\hat{p} = \frac{x}{n}$

$$z = \frac{\left(\frac{x}{n} \pm \frac{1}{2n}\right) - p}{\sqrt{\frac{pq}{n}}} = \frac{\left(\hat{p} \pm \frac{1}{2n}\right) - p}{\sqrt{\frac{pq}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}.$$

The quantity $\pm\frac{1}{2n}$ will be ignored for large values of n because the value of z changes very little.

Please note that the equivalent check for the "appropriateness" of the normal approximation for the sampling distribution of $\hat{p}$ will still be when $np > 5$ and $n(1-p) > 5$.

8.7.2 Sampling distribution of sample proportion

1. $\hat{P}$ is approximately normally distributed provided that np and $n(1-p)$ are greater than equal to 5.
2. The expected value: $E\left(\hat{P}\right) = p$.
3. The variance: $V(\ ) = \dfrac{p(1-p)}{n}$
4. The standard deviation $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$

The standard deviation $\hat{P}$ is a called the standard error of the proportion.

8.7.3 Example

In the following multiple-choice questions, **please select the correct answer**.

Question 1

A sample of 250 observations will be selected at random from an infinite population. Given that the population proportion is 0.25, the standard error of the sampling distribution of the sample proportion is

(a) 0.027 4

(b) 0.500 0

(c) 0.031 6

(d) 0.054 8

Question 2

As a general rule, the normal distribution is used to approximate the sampling distribution of the sample proportion only if

(a) the sample size n is greater than 30

(b) the population proportion p is close to 0.50

(c) the underlying population is normal

(d) np and $n(1-p)$ are both greater than or equal to 5

Solutions:

Question 1

$\upsilon ar(\widehat{p}) = \dfrac{p(1-p)}{n} = \frac{0.25(1-0.25)}{250} = 0.000\,75$

$\therefore \sqrt{\upsilon ar(\widehat{p})} = \sqrt{0.000\,75} = 0.027\,39 =$ the standard error.

ANSWER: (a)

Question 2

As a general rule, the normal distribution is used to approximate the sampling distribution of the sample proportion only if np and $n(1-p)$ are both greater than or equal to 5.

ANSWER: (d)

8.8 Sampling distribution of the difference between two means

The sampling plan calls for independent random samples drawn from each of two normal populations. The samples are referred to as independent if the selection of the members of one sample is independent of the selection of the members of the second sample. Using the laws of expected value and variance we derive the expected value and the variance of the sampling distribution of the difference between two means $\left(\bar{X}_1 - \bar{X}_2\right)$:

$$\mu_1 - \mu_2$$

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

Thus, it follows that in repeat independent sampling from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively, the sampling distribution of $X_1 - X_2$ is normal with mean $\mu_1 - \mu_2$.

In this study unit you are going to do inference for a single population. However you will be introduced to the sampling distribution for the difference between two population means. This is in preparation for STA1502. What we are really interested in is the *sampling distribution* of $(\overline{X}_1 - \overline{X}_2)$. In other words we would like the *two-sample extension* of the sampling distribution of the sample mean (where we previously had a single sample). This means we have a **population 1** from which we draw a sample of size n_1 and we have a **population 2** from which we draw a sample of size n_2.

It makes intuitive sense that

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

but what many students find confusing is that

$$var(\overline{X}_1 - \overline{X}_2) = var(\overline{X}_1) + var(\overline{X}_2).$$

Yes, there is a *plus sign* even though you might expect a minus sign!

In other words if we create a new variable by **subtracting two variables**, the variance of this new variable will be the **sum of the variances** of the two original variables.

Please Note: Strictly speaking, when we derive this variance mathematically, there is a third term that takes care of the *dependency or connectedness* between the two variables but this term falls away if we assume that **population 1 and population 2 are independent**. You will find that many textbooks emphasize this condition that the two samples need to be from independent populations.

8.8.1 Example

In the following multiple-choice questions, **please select the correct answer**.

Question 1

If two random samples of sizes n_1 and n_2 are selected independently from two populations with means μ_1 and μ_2 respectively, then the mean of the sampling distribution of the sample mean difference, $\overline{X}_1 - \overline{X}_2$, is equal to

(a) $\mu_1 + \mu_2$

(b) $\mu_1 - \mu_2$

(c) $\dfrac{\mu_1}{\mu_2}$

(d) $(\mu_1)(\mu_2)$

Question 2

If two populations are normally distributed, the sampling distribution of the sample mean difference will be

(a) approximately normally distributed

(b) normally distributed only if both sample sizes are greater than 30

(c) normally distributed

(d) normally distributed only if both population sizes are greater than 30

Solution:

Question 1

If two random samples of sizes n_1 and n_2 are selected independently from two populations with means μ_1 and μ_2 respectively, then the mean of the sampling distribution of the sample mean difference, $\overline{X}_1 - \overline{X}_2$ is equal to $\mu_1 - \mu_2$.

ANSWER: (b)

Question 2

If two populations are normally distributed, the sampling distribution of the sample mean difference will be normally distributed.

ANSWER: (c)

8.9 Study Unit 8: Summary

I. There are many *reasons for sampling a population*.

   A. The results of a sample may adequately estimate the value of the population parameter, thus saving time and money.

   B. It may be too time consuming to contact all members of the population.

   C. It may be impossible to check or locate all the members of the population.

   D. The cost of studying all the items in the population may be prohibitive.

   E. Often testing destroys the sampled item and it cannot be returned to the population.

II. The *sampling error* is the difference between a population parameter and a sample statistic.

III. The *sampling distribution* of the sample mean is a probability distribution of all possible sample means of the same sample size.

A. For a given sample size, the mean of all possible sample means selected from a population is equal to the population mean.

B. There is less variation in the distribution of the sample mean than in the population distribution.

C. The standard error of the mean measures the variation in the sampling distribution of the sample mean. The standard error is found by

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

D. If the population follows a normal distribution, the sampling distribution of the sample mean will also follow the normal distribution for samples of any size. Assume the population standard deviation is known. To determine the probability that a sample mean falls in a particular region, use the following formula:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \text{ otherwise use } t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

IV. In comparing two population means we wish to know whether they could be equal.

A. We are investigating whether the distribution of the difference between the means could have a mean of 0.

B. The test statistic follows the standard normal distribution if the population standard deviations are known.

1. No assumption about the shape of either population is required.
2. The samples are from independent populations.
3. The formula to compute the value of z is

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

V. The test statistics to compare two means is the t-distribution if the population standard deviations are not known.

A. Both populations must follow the normal distribution.

B. The populations must have equal standard deviations.

C. The samples are independent.

D. Finding the value of t requires two steps.

1. The first step is to pool the standard deviations according to the following formula:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

2. The value of t is computed from the following formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. The degrees of freedom for the test are $n_1 + n_2 - 2$.

Study Unit 9

INTRODUCTION TO ESTIMATION

9.1 Learning outcomes

After completing this unit, you should be able to:

1. describe the term point estimation and interval estimation
2. describe the desirable quality of estimator : Unbiasedness, consistency and relative efficiency
3. apply the sampling distribution of the sample mean.
4. to compute the confidence interval estimate for the population mean
5. interpret a confidence interval
6. compute and interpret a confidence interval estimate of the population mean
7. describe the effect of the width of the confidence interval if we change the confidence interval, the sample size and the population standard deviation
8. compute the sample size needed to estimate a population mean when a bound on the error of estimation is specified.

9.2 Introduction

In the previous study unit we derived the sampling distribution of the sample mean, $\overline{X}$, the sample proportion, $\widehat{p}$, and the sample mean difference, $\overline{X}_1 - \overline{X}_2$, which were three distinct sample statistics. We also applied these sampling distributions *in a rather artificial manner* to compute probabilities for $\overline{X}$, for $\widehat{p}$ and for $\overline{X}_1 - \overline{X}_2$.

By saying “artificial manner” we mean “not true to real life” because all our exercises started off with **a statement that specified the values of the population parameters**!

If you understand how a confidence interval is derived for μ using the sampling distribution of $\overline{X}$, you will likely understand how a confidence interval is derived for p using the sampling distribution of $\widehat{p}$, and you will likely understand how a confidence interval is derived for $\mu_1 - \mu_2$ using the sampling distribution of $\overline{X}_1 - \overline{X}_2$, which you will do in the study guide STA1502.

9.3 Concepts of estimation

The purpose of estimation is to find an estimate value of a population parameter on the basis of sample statistic. For example, we compute the sample mean of a data set and used the sample mean to make inference, meaning to estimate the population mean based on the sample mean. The sample mean is called the estimator of the population. The estimate is the computed sample mean. This unit introduce the estimation process called inferential statistics. The estimation process in which a population mean is estimated using a sample data . The value of a population mean μ is estimated using the value of a sample mean X.

The estimation process is conducted in ways; we compute the value of the estimator and consider that value as the estimate of the parameter. Such estimator is called a point estimator.

9.3.1 Point and interval estimator.

Point estimator

A *point estimator* draws inferences about a population by estimating the value of an unknown parameter using a single value or point.

Limitations of point estimator

1. The estimate may be wrong
2. We often need to know how close the estimator is to the parameter
3. Size of the sample may affect the point estimator

To overcome this limitations, a second method of estimation a population parameter called interval estimator is be introduced.

Interval estimator

An *interval estimator* draws inference about a population by estimating the value of an unknown parameter using an interval.

Unbiased Estimator,

An *unbiased estimator* of population parameter is an estimator whose expected value is equal to that parameter.

Consistency

An *unbiased estimator* is said to be consistent if the difference between the estimator and the parameter grows smaller as the sample size increases.

Relative Efficiency

If there are two unbiased estimators of a parameter, the one whose variance is smaller is referred to as a *relative efficiency*.

9.4 Point and interval estimator

Suppose that we have a population with a mean μ and the standard deviation σ. The population mean is assumed to be unknown, its value need to be estimated.

Using the algebraic manipulation , we can express the probability in a slightly different form:

$$P(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$1 - \alpha$ is called the confidence level

The confidence interval estimator of μ is given by:

$$\bar{x} \pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

where $\bar{x} - Z_{\frac{\alpha}{2}}$ is the lower confidence limit and $\bar{x} + Z_{\frac{\alpha}{2}}$ is the upper confidence limit.

9.5 Estimating the population mean when the population standard deviation is known

[Please note that you will not be examined on the last subsection: **(Optional) Estimating the Population Mean Using the Sample Median.]**

$$P\left(-z_{\alpha/2} \leq \frac{(\overline{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

He points out that it was "manipulated" in chapter 8 to be used as **a probability expression concerning** $\overline{X}$.

$$P(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

Sample of size $n = 15$ from a normal population where $\mu = 1$ and $\sigma = 0.1$ which is yet again *a rather artificial situation* as we pointed out in the introduction! Nevertheless, we have all the information to apply the probability expression above. More specific, for a 95% probability expression concerning $\overline{X}$, we substitute the values of $z_{\alpha/2} = 1.96$; $\mu = 1$ and $\sigma = 0.1$ and compute the interval limits as

$$\begin{aligned} P(1 - 1.96\frac{0.1}{\sqrt{15}} \leq \overline{X} \leq 1 + 1.96\frac{0.1}{\sqrt{15}}) &= 1 - 0.05 \\ P(0.949\,39 \leq \overline{X} \leq 1.050\,61) &= 0.95 \end{aligned}$$

From this we conclude that we are 95% sure that the sample mean will be a value between 0.949 39 and 1.050 61 if we draw a sample of size $n = 15$ from a normal population where $\mu = 1$ and $\sigma = 0.1$ **Big deal! It helps us nothing with estimating an unknown parameter!**

How can we use this expression "in reverse" and what will it then imply?

Suppose we start off with the probability statement

$$P(-1.96 \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq 1.96) = 0.95$$

but now we manipulate it "the other way around" to end up with

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Do you think it is valid as a probability statement? **No, it is not!** Although it is a valid algebraic manipulation we cannot call it a *probability interval* because the ends are not fixed but depend on the variable outcome of $\overline{X}$.

We have something of a *paradoxical situation*! For some samples the interval estimate of μ will be correct but for other samples it could be incorrect! If we continue to compute such intervals for many more observed

sample values for $\overline{X}$ we will find that 95% of all sample means will produce an interval that includes the true population mean. Thus we say that we are *95% confident* that the population mean will be contained in the interval $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ if we compute the interval for a specific selected sample.

Please note that this is a very subtle use of language. We are strictly speaking not allowed to say we are 95 % *sure* (which implies there is a probability of 0.95) that μ will fall between $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ but we are willing to say that we are *95% confident* that the population mean will be contained in the interval

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}};\ \overline{X} + 1.95\frac{\sigma}{\sqrt{n}}).$$

Moral of the story?

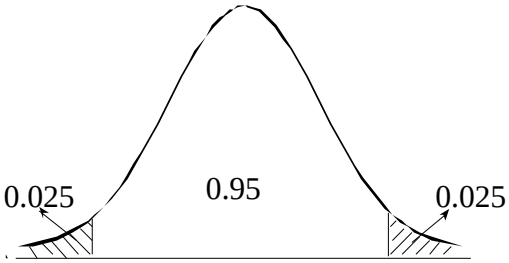
You need to *know the formula* and you need to know it is called a *confidence* interval!

Show that

$$\begin{aligned} P(-1.645 \leq Z \leq 1.645) &= 0.90 \\ P(-1.96 \leq Z \leq 1.96) &= 0.95 \\ P(-2.33 \leq Z \leq 2.33) &= 0.98 \\ P(-2.575 \leq Z \leq 2.575) &= 0.99 \end{aligned}$$

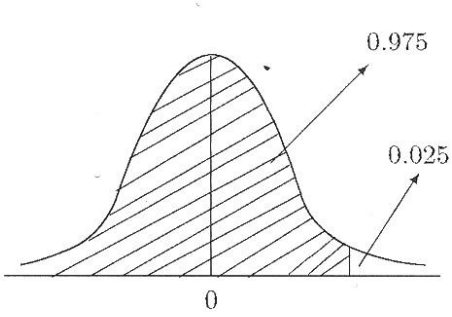
Solution:

Suppose $1 - \alpha = 0.95$



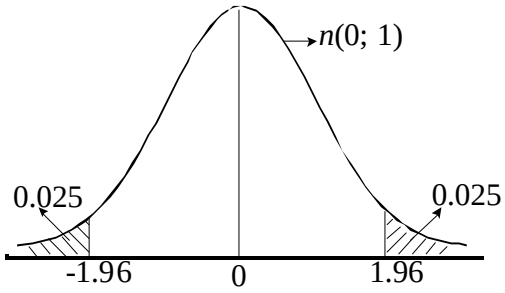
Step 1: Draw a sketch showing

$(1 - \alpha) = 0.95$ and $\frac{\alpha}{2} = 0.025$



Step 2: Convert this to an area compatible Normal Distribution Table

Step 3: Use Normal Distribution Table to find the z-value corresponding with 0.9750



Step 4: Final sketch showing the associated $z_{\alpha/2}$-value on the horizontal axis where $\frac{\alpha}{2} = 0.025$

In a similar fashion we find the other three values.

$z_{0.05} = 1.645$ where $\alpha = 0.10 \Longrightarrow \frac{\alpha}{2} = 0.05$

$z_{0.01} = 2.33$ where $\alpha = 0.02 \Longrightarrow \frac{\alpha}{2} = 0.01$

$z_{0.005} = 2.575$ where $\alpha = 0.01 \Longrightarrow \frac{\alpha}{2} = 0.005$

:-) Please note: In many textbooks the values 1.645, 1.96, 2.33 and 2.575 are called **critical values**. For most students the biggest hurdle in computing a confidence interval estimate is to find the appropriate $z_{\alpha/2}$-value! Did you find the activity above easy to do? Then the rest of the problems will be even easier. Also

note that question 3 of the next activity is “the most typical” and almost like a standard blue print type of question on this section!

9.5.1 Example

Question 1

Suppose that in developing an interval estimate for a population mean, the population standard deviation σ was assumed to be 20. The interval estimate was 39.43 ± 1.07. Show that if σ had equalled 40, the interval estimate would be 39.43 ± 2.14.

Solution:

An interval estimate for a population mean is computed as $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Thus we know that $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 39.43 \pm 1.07$

$$\Longrightarrow z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.07$$

Doubling the value of σ will result in $z_{\alpha/2}\frac{(2)\sigma}{\sqrt{n}} = (2)1.07 = 2.14$

Hence the interval estimate becomes 39.43 ± 2.14

Question 2

Suppose that in developing an interval estimate for a population mean, a sample of 50 observations was used. The interval estimate was 19.76 ± 1.32. Show that if the sample size had been 200 instead of 50, the interval estimate would have been 19.76 ± 0.66.

Solution:

An interval estimate for a population mean is computed as $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Thus we know that $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 19.76 \pm 1.32$

$$\Longrightarrow z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.32$$

Changing the sample size from 50 to 200 will result in $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}(\frac{\sqrt{50}}{\sqrt{200}}) = 1.32\frac{\sqrt{50}}{\sqrt{200}} = 0.66$

Hence the interval estimate becomes 19.76 ± 0.66

Question 3

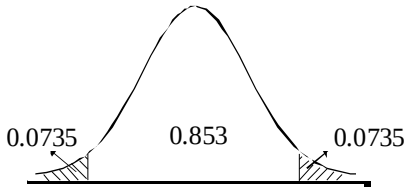
A random sample of 64 observations has a mean of 30. The population variance is assumed to be 9. Compute the 85.3% confidence interval estimate for the population mean (to the third decimal place).

Solution:

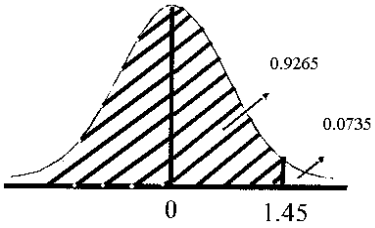
An interval estimate for a population mean is computed as $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. [**:-)** I suppose that by now you realise that this is a formula you must know by heart!]

From the given information we know that

$\overline{X} = 30;\ n = 64$ and $\sigma^2 = 9$. We need to find the critical value associated with a 85.3% confidence level.



$$\begin{aligned} \text{If } (1-\alpha) &= 0.853 \\ \Longrightarrow \frac{\alpha}{2} &= (1-0.853)/2 \\ &= 0.0735 \end{aligned}$$



Thus $z_{\alpha/2} = 1.45$

$$\begin{aligned} \text{and } \overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} &= 30 \pm z_{\alpha/2}\frac{\sqrt{9}}{\sqrt{64}} \\ &= 30 \pm (1.45)\frac{3}{8} \\ &= 30 \pm 0.543\,75 \\ &= (29.4563;\ \ 30.5438). \end{aligned}$$

9.5.2 Information of the width of the interval

A large confidence interval is more desirable because it contains a larger proportion of the informations of the estimate that may be corrected in the future. The width of the interval and the confidence level are connected. We need to widen the interval so that we will be more confident in the estimate.

9.5.3 Examples

1. For $\bar{X} = 370.16$, $s = 25$ and $n = 150$

   Construct a 95% confidence interval: $370.16 \pm 1.96\frac{150}{\sqrt{25}} = 370.16 \pm 58.80$

2. For $\bar{X} = 370.16$, $s = 25$ and $n = 75$

   Construct a 90% confidence interval: $370.16 \pm 1.645\frac{75}{\sqrt{25}} = 370.16 \pm 24.68$

3. For $\bar{X} = 370.16$, $s = 25$ and $n = 75$

   Construct a 99% confidence interval: $370.16 \pm 2.575\frac{75}{\sqrt{25}} = 370.16 \pm 38.63$

This example showed that when decreasing the confidence, the intervals becomes narrow.

9.6 Selecting the sample size

An interval estimate for a population mean is computed as $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. [**:-)** Yes, this is once again our know-it-by-heart formula!]

$(z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ contains *three different components* that could cause the interval to narrow or widen.

For a *fixed confidence level*, implying that $z_{\alpha/2}$ is fixed, as well as a *fixed population standard deviation*, σ, the only other possible way we can shrink the sampling error is to increase $\sqrt{n}$. Taking care of the square root sign and equating it to the bound on the error of estimation, simplifies to the formula

$n = \left[\frac{(z_{\alpha/2})(\sigma)}{B}\right]^2$ where B is the bound on the error of estimation.

9.6.1 Examples

Question 1

What sample size should be used to estimate the mean of a normal population with 99% confidence, if the population standard deviation is assumed to be 6? The bound on the error of estimation must not exceed 1.2.

An interval estimate for a population mean is computed as $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Suppose this is the only formula you know, then you can figure the value of n out for yourself!

So, for this problem we know that $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.2$

$$\Longrightarrow n = \left[\frac{(z_{\alpha/2})(\sigma)}{1.2}\right]^2.$$

[**:-)** Of course this checks with the formula "Sample Size to Estimate μ" which Keller gives in the **Chapter Summary**, as the formula $n = \left[\frac{(z_{\alpha/2})(\sigma)}{B}\right]^2$ where B is the bound on the error of estimation.]

From activity 10.2 we have that

$z_{0.005} = 2.575$ where $\alpha = 0.01$

$$\Longrightarrow n = \left[\frac{(2.575)(6)}{1.2}\right]^2$$

$$= 165.765\,63.$$

$n = 166$ should be used.

Question 2

A statistician wants to estimate the mean weekly family expenditure on clothes for a specific defined group. He believes that the standard deviation of the weekly expenditure for this group is R125. Determine with 95% confidence the number of families that must be sampled to estimate the mean weekly family expenditure on clothes to within R14 of the true value.

An interval estimate for a population mean is computed as $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Thus we know that $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq 14$

$$\Longrightarrow n \geq \left[\frac{(z_{\alpha/2})(\sigma)}{14}\right]^2.$$

From activity 10.2 we have that

$z_{0.025} = 1.96$ where $\alpha = 0.05$

$$\Longrightarrow n \geq \left[\frac{(1.96)(125)}{14}\right]^2$$

$$= 306.25$$

$n = 307$ should be used.

Please note: n should always be rounded up to the next integer in order to achieve our precision.

Key Terms/Symbols

point estimation

interval estimation

unbiasedness

consistency

relative efficiency

confidence interval estimate

width of the confidence interval

confidence level

bound on the error of estimation

9.7 Study Unit 9: Summary

I. A *point estimate* is a single value (statistic) used to estimate a population value (parameter).

II. A *confidence interval* is a range of values within which the population parameter is expected to occur.

   A. The factors that determine the width of a confidence interval for a mean are the following:

      1. The number of observations in the sample, n.
      2. The variability in the population, usually estimated by the sample standard deviation s.
      3. The level of confidence.
         a. To determine the confidence limits when the population standard deviation is known, we use the z–distribution. The formula is $\overline{X} \pm z\frac{\sigma}{\sqrt{n}}$
         b. To determine the confidence limits when the population standard deviation is unknown, we use the t–distribution. The formula is $\overline{X} \pm t\frac{2}{\sqrt{n}}$

III. The major characteristics of the *t-distribution* are the following:

   A. It is a continuous distribution.

   B. It is mound-shaped and symmetrical.

   C. It is flatter, or more spread out, than the standard normal distribution.

   D. There is a family of t-distributions, depending on the number of degrees of freedom.

IV. A *proportion* is a ratio, fraction, or percent that indicates the part of the sample or population that has a particular characteristic.

A. A sample proportion is found by X, the number of successes, divided by n, the number of observations.

B. We construct a confidence interval for a sample proportion from the following formula: $p \pm z\sqrt{\frac{p(1-p)}{n}}$

V. For a finite population, the standard error is adjusted by the factor $\sqrt{\frac{N-n}{N-1}}$.

VI. We can determine an appropriate sample size for estimating both means and proportions.

A. There are three factors that determine the sample size when we wish to estimate the mean.

1. The desired level of confidence, which is usually expressed by z.
2. The maximum allowable error, B.
3. The variation in the population, expressed by s.
4. The formula to determine the sample size for the mean is $n = \left(\frac{z\sigma}{B}\right)^2$.

References

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 10

INTRODUCTION TO HYPOTHESIS TESTING

10.1 Learning outcomes

After completion of this study unit, you should be able to do the following:

1. Define the following concepts of hypothesis testing:
   - *null hypothesis*
   - *alternative hypothesis*
   - *significance level*
   - *test statistic*
   - *rejection region*
   - *p-value*
2. Describe the types of errors in hypothesis testing
   - *Type I error*
   - *Type II error*
3. Explain the difference between *one-sided* and *two-sided* testing
3. Apply the *sampling distribution of the sample mean,* $\overline{X}$, to perform a hypothesis test for the population mean μ
4. Explain what is meant by the *power* of a hypothesis test
5. Obtain the *rejection region* for any given *significance level*

6. Perform a complete hypothesis test for the population mean

7. Describe the connection between confidence interval estimation and hypothesis testing

10.2 Introduction

In the previous study unit we derived an interval estimate that consisted of two points defining an interval which was used to estimate the unknown population parameter. Most importantly, we attached a probability or a degree of certainty to our result. The starting point was to have a *known sampling distribution* of a *sample statistic* and to manipulate a probability expression whereby we create a lower and an upper bound for the interval.

Hypothesis testing and the computation of a confidence interval for an unknown population parameter usually go hand in hand. Most textbooks prefer to organise study contents in such a manner that the derivation of a confidence interval is followed directly by the derivation of the hypothesis test for the same, specific sampling distribution. Keller organised it slightly differently by first emphasizing the concept of interval estimation and deriving a confidence interval for a population mean only. (Confidence intervals for other population parameters will follow in other chapters.)

In this chapter he derives the corresponding hypothesis tests for a population mean. (Hypothesis tests for other population parameters will follow in other chapters.)

We have already mentioned that a hypothesis test and the computation of a confidence interval for an unknown population parameter usually go hand in hand. With hypothesis testing we will make a statement – usually a supposition about the value of an unknown population parameter – which we call the *null hypothesis* and which is represented with the notation H_0. This null hypothesis uniquely specifies a value for the relevant population parameter. A so-called *alternative hypothesis* (represented with the notation H_1) gives us an alternative supposition about the population parameter and we have to *choose one of these two statements*.

10.3 Concepts of hypothesis testing

A hypothesis test is an statement or an assumption about a population, which may be true or not. The purpose of hypothesis testing is to determine when it is reasonable to make a decision or a conclusion on a given population based on a sample.

Steps on how to conduct an hypothesis testing

Step 1

Formulate the two hypotheses:

1. The null hypothesis symbolised by H_0 (always carries the equality sign $=$, $\geq$ or $\leq$)
2. The alternative hypothesis or research symbolised by H_1 (carries the sign $\neq$, $>$ and $<$)

Step 2

Testing procedure begins with the assumption that the null hypothesis is true.

Step 3

The Goal of the process is to determine whether there is enough evidence to conclude that the alternative is true

Step 4

Determine the level of significance

Step 5

Calculate the test Statistics

Step 6

Make a decision :

1. Conclude that there is enough evidence to support the alternative hypothesis
2. Conclude that there is not enough evidence to support that the alternative hypothesis

   Two possible errors can be made in a test:

   1. A type one error : reject a true null hypothesis
      A type two error : we don't reject false null hypothesis.

   $P(\text{Type 1 error}) = \alpha$

   $P(\text{Type 2 error}) = \beta$

The following multiple-choice questions are aimed at doing this explaining the concept of Hypothesis testing.

In the following multiple-choice questions, **select the correct answer**.

Question 1

In a criminal trial, a Type I error is

(a) made when a guilty defendant is acquitted

(b) made when an innocent person is convicted

(c) made when a guilty defendant is convicted

(d) made when an innocent person is acquitted

(e) not applicable

Question 2

In a criminal trial, a Type II error is

(a) made when a guilty defendant is acquitted

(b) made when an innocent person is convicted

(c) made when a guilty defendant is convicted

(d) made when an innocent person is acquitted

(e) not applicable

Question 3

The probability of a Type I error is denoted by

(a) β

(b) $1-\beta$

(c) α

(d) $1-\alpha$

(e) $P(\alpha)$

Question 4

A Type I error is committed if we make

(a) a correct decision when the null hypothesis is false

(b) a correct decision when the null hypothesis is true

(c) incorrect decision when the null hypothesis is false

(d) incorrect decision when the null hypothesis is true

(e) none of the above

Question 5

A Type II error is committed if we make

(a) a correct decision when the null hypothesis is false

(b) a correct decision when the null hypothesis is true

(c) incorrect decision when the null hypothesis is false

(d) incorrect decision when the null hypothesis is true

(e) none of the above

Question 6

A professor of statistics wants to disprove the claim that the average student spends 3 hours studying for the statistics exam. Which hypothesis is used to test the claim?

(a) $H_0: \mu \neq 3$ vs. $H_1: \mu > 3$

(b) $H_0: \mu = 3$ vs. $H_1: \mu \neq 3$

(c) $H_0: \mu \neq 3$ vs. $H_1: \mu = 3$

(d) $H_0: \mu = 3$ vs. $H_1: \mu < 3$

(e) none of the above

Question 7

In hypothesis testing, whatever we are investigating or researching is specified as

(a) the null hypothesis

(b) the alternative hypothesis

(c) either the null or alternative hypothesis

(d) the p-value

(e) the α-value

Question 8

A wife complains to her husband that the average amount of money spent on Christmas gifts for immediate family members is above R1 200. The correct set of hypotheses is

(a) $H_0: \mu = 1\,200$ vs. $H_1: \mu < 1\,200$

(b) $H_0: \mu > 1\,200$ vs. $H_1: \mu = 1\,200$

(c) $H_0: \mu = 1\,200$ vs. $H_1: \mu > 1\,200$

(d) $H_0: \mu < 1\,200$ vs. $H_1: \mu = 1\,200$

(e) none of the above

Solutions:

Question 1

ANSWER: (b)

Question 2

ANSWER: (a)

Question 3

ANSWER: (c)

Question 4

ANSWER: (d)

Question 5

ANSWER: (c)

Question 6

ANSWER: (b)

Question 7

ANSWER: (b)

Question 8

ANSWER: (c)

10.4 Testing the population mean when the population standard deviation is known

The essence of a statistical test of H_0 is a comparison of actual sample data with what might be expected when H_0 is true. Such a comparison is based on the outcome of a relevant sample statistic, which in a hypothesis-testing setup, is called the **test statistic**.

Similar to confidence interval estimation the starting point is to have a *known sampling distribution* of a *test statistic* and to use this information (but in a slightly different manner, as for a confidence interval) in order **to specify values for the test statistic that are extreme under** H_0. Of course this will be in the direction of H_1 , suggesting that H_1 provides a better explanation of the data than H_0.

What does "to specify values for the test statistic that are extreme under H_0" mean? Keller defines it as "**to obtain a rejection region**".

Testing the population mean when the population standard deviation is known

Back to this specific application of hypothesis testing, our starting point is the sampling distribution of $\overline{X}$. We know that $\overline{X}$ has either a normal or approximately normal distribution, i.e. $\overline{X} \sim n(\mu; \frac{\sigma^2}{n})$.

From chapter 8 we also know that we can switch between any normal distribution and the standardised version of the normal distribution, which implies that

$$Z_{\bar{X}} = \frac{\bar{X} - E\left(\bar{X}\right)}{\sqrt{var\left(\bar{X}\right)}} = \frac{\sqrt{n}\left(\bar{X}\right) - \mu}{\sigma} \sim n\left(0; 1\right).$$

In general, for any hypothesis test *the first step* is to **specify the hypotheses** to be tested. Obviously in this application our null hypothesis will be a statement about μ.

10.4.1 Example

The manager of a department store is thinking about establishing a new billing system for the store's credit customers. After a through financial analysis, she determine that the new system will be cost-effective only if the mean monthly account is more than R170. A random sample of 400 monthly accounts is draw, for which the sample mean is R178. the knows that the account are approximately normally distributed with a standard deviation of R65. Can the manager conclude form this that the new system will be cost effective?

$$\begin{aligned} H_0: &\quad \mu &=& \quad 170 \\ H_1: &\quad \mu &>& \quad 170 \end{aligned}$$

[**:-) Do you know?** The $ is the well-known American currency and the value of the dollar depends on the exchange rate, which changes on a daily basis.]

In general, *the second step* is to decide on a **test statistic** that contains information about an unknown population parameter and that would make a statement about the unknown population parameter possible. Now in this application our test statistic is $Z_{\overline{X}} = \dfrac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$.

Did you notice that the test statistic contains both $\overline{X}$ and μ? So whatever statement we make about μ (in the null hypothesis) as well as the observed value of the sample statistic $\overline{X}$, will be substituted in this expression of the test statistic, to finally result in a numerical value.

In general, *the third step* is to find a **decision rule** that would enable you to make the correct decision, and there are two routes:

Route A is only possible when you have a computer program to compute the *p-value*.

Route B is compulsory when you perform a hypothesis tests manually, and you have "**to obtain a rejection region**".

10.4.2 Rejection region

The rejection region is a range of values such that if the test statistics falls into that range, the null hypothesis is rejected in favour of the alternative hypothesis.

Let $\bar{x_L}$ be the value of the sample mean that is just large enough to reject the null hypothesis, then the rejection region is defined by:

$$\bar{x} > \bar{x_L}$$

Set the rejection region in term of the sample mean by computing the following

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

We reject the null hypothesis in favour of the alternative if the value of the sample mean is large relative to 170. If the calculated sample mean was 500, it would be quite apparent that the null hypothesis is false and rejected. If the value of the sample mean is close to 170, do not allow us to reject the null hypothesis because it is entirely possible to observe a sample mean of 171 from a population whose mean is 170. Unfortunately, the decision is not always so obvious. In this example, the calculated sample mean is 178, a value which is neither very close to 170. To make an informed decision, we need determine the rejection region.

Using the example 11.3 , the rejection region is $Z > Z_\alpha = Z_{0.05} = 1.645$

The value of the statistic is given by $Z = \frac{178-170}{65/\sqrt{400}} = 2.46$

Because 2.46 is greater that 1.645, we reject the null hypothesis and conclude that there is enough evidence to infer that mean monthly account is greater that R170.

10.4.3 *P*–value

The rejection region gives only a yes or a no to the question, is there enough evidence to suggest that the alternative hypothesis is true? To take full advantage of the available information from the test result and make better and informed decisions, we need to compute the value called the p-value. To determine the p-value, we need the level of significance.

The p-value of a test is the probability of observing a test statistic at least as extreme as the one computed knowing that the null hypothesis is true.

Using example, the p-value is given by:

$$\begin{aligned} P\left(\bar{X} > 178\right) &= P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{n}} > \frac{178-170}{65/\sqrt{400}}\right) = P\left(Z > 2.46\right) \\ &= 1 - P(Z < 2.46) = 1 - 0.9931 = 0.0069 \end{aligned}$$

Describing the p-value

If the p-value is less than 0.01, there is overwhelming evidence to infer that the alternative hypothesis is true: the test is highly significant.

If the p values lies between 0.01 and 0.05, there is strong evidence to infer that the alternative hypothesis is true, the test is significant.

It the p-values is 0.05 and 0.10, there is a weak evidence to indicate that the alternative hypothesis is true. When p is > 0.05 , the result is statistically significant if the p value is > 0.10, there is no evidence to infer that the alternative hypothesis is true.

The p-value rejection method

If the p value is less than α, (level of significances) we judge the p-value to be small enough to reject the null hypothesis if the p value is greater than α, we do not reject the null hypothesis

When do you conduct one and two tail tests?

A two tail test is conducted whenever the alternative hypothesis specifies that the mean is not equal (not the same) to the value stated in the null hypothesis. For a two tail test hypotheses are as follows:

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

For a one tail test that focuses on the right tail of the sampling distribution, the hypotheses are as follows:

$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

For a one tail test that focuses on the left tail of the sampling distribution, the hypotheses are as follows:

$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

This rejection region will enable you to make the correct decision. Unfortunately this manual third step is usually the most difficult one where you have to understand what goes for what! There isn't a clear "recipe" as only understanding the principles will guarantee a correct result! (You will see that the more hypothesis tests we perform the easier this step will become!) However, it is important to remember that the rejection region is a function of the significance level α, as well as the direction of the alternative hypothesis.

In this application (i.e. the Keller example of the "Billing System") it was decided to use $\alpha = 0.05$. Since we have a one-sided alternative hypothesis we will also have a one-sided rejection region, and more specifically a *right-sided rejection region* as depicted in figure 11.1.

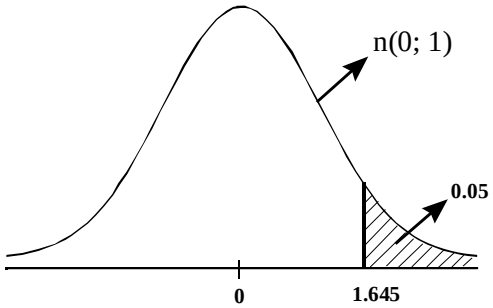


Figure 11.1 Right-sided rejection region for a 5% significance level

Please note:

[1] For $H_1: \mu < 170$ we would have used α on *the left side* and obtained a *left-sided rejection region.*

[2] You must keep in mind that in testing a hypothesis, statements for the null and alternative hypotheses as well as the selection of the level of significance *should precede* the collection and examination of the data.

To round off this specific hypothesis test, all we have to do is to compute $Z_{\overline{X}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$ for the given sample/research project, and make the final conclusion.

With the given information we substitute and compute

$$\begin{aligned} Z_{\overline{X}} &= \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \\ &= \frac{\sqrt{400}(178 - 170)}{65} \\ &= 2.46 \end{aligned}$$

Conclusion: Since the computed value of $Z_{\overline{X}}$ falls in the rejection region ($2.46 > 1.645$), we **reject the null hypothesis**, and conclude that there is enough statistical evidence to infer that the alternative hypothesis is true. The mean monthly account is more than \$170.

10.4.4 Example

Question 1

Refer to 9.5, Unit 9, we obtained

$$\begin{aligned} P(-1.645 \leq Z \leq 1.645) &= 0.90 \\ P(-1.96 \leq Z \leq 1.96) &= 0.95 \\ P(-2.33 \leq Z \leq 2.33) &= 0.98 \\ P(-2.575 \leq Z \leq 2.575) &= 0.99 \end{aligned}$$

How can these **critical values** we obtained for the derivation of confidence intervals, be viewed or utilised as rejection regions?

Draw diagrams showing the *critical values* and shade the areas representing the *rejection regions*.

Solution:

It is important to note that the information for a *confidence interval* can only be translated to a *two-sided hypothesis test*. This implies that, for the population mean, we will have a significance level of α and we will test

$$\begin{aligned} H_0: \quad \mu &= \mu_0 \\ H_1: \quad \mu &\neq \mu_0 \end{aligned}$$

$\mu = \mu_0$ is our notation meaning "μ is a specified value under H_0" and for a real-life application it will be a numerical value.

If

$$\begin{aligned} \alpha = 0.10 &\Longrightarrow \quad \tfrac{\alpha}{2} = 0.05 \quad \Longrightarrow \quad z_{0.05} = 1.645 \\ \alpha = 0.05 &\Longrightarrow \quad \tfrac{\alpha}{2} = 0.025 \quad \Longrightarrow \quad z_{0.025} = 1.96 \\ \alpha = 0.02 &\Longrightarrow \quad \tfrac{\alpha}{2} = 0.01 \quad \Longrightarrow \quad z_{0.01} = 2.33 \\ \alpha = 0.01 &\Longrightarrow \quad \tfrac{\alpha}{2} = 0.005 \quad \Longrightarrow \quad z_{0.005} = 2.575 \end{aligned}$$

So, what is the *difference in focus* between a confidence interval and a hypothesis test?

With a confidence interval our focus is on the *inside* of the probability statement and with a hypothesis test our focus is on the *outside* of the probability statement. For example, for a 90% confidence interval

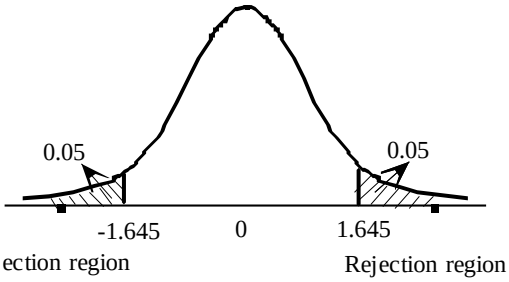
$$P(-1.645 \leq Z \leq 1.645) = 0.90$$

which implies that

$$P(Z \leq -1.645) + P(Z \geq 1.645) = \alpha.$$

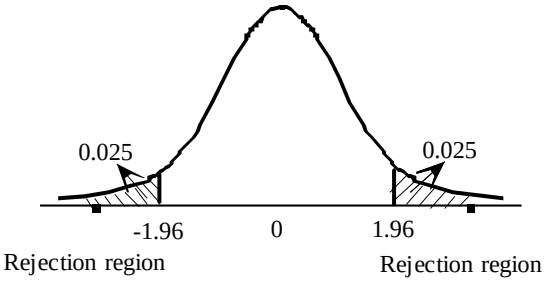
In the following four sketches we show the "origin" of the confidence interval at the top and the significance level for the corresponding two-sided hypothesis test at the bottom. If you truly grasp this, the rest of the chapters will be easy as pie! **:-)**

$P(-1.645 \leq Z \leq 1.645) = 0.90$



Two-sided hypothesis test using $\alpha = 0.10$

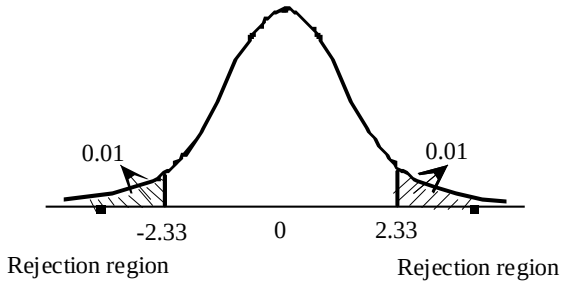
$P(-1.96 \leq Z \leq 1.96) = 0.95$



Two-sided hypothesis test using $\alpha = 0.05$

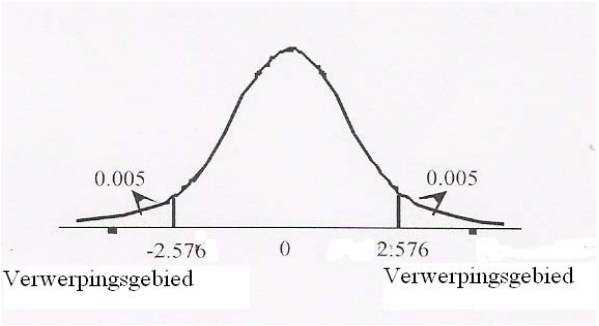
In a similar fashion we find

$$P(-2.33 \leq Z \leq 2.33) = 0.98$$



Two-sided hypothesis test using $\alpha = 0.02$

$$P(-2.575 \leq Z \leq 2.575) = 0.99$$



Two-sided hypothesis test using $\alpha = 0.01$

Question 2

If a hypothesis is not rejected at the 0.10 level of significance,

(a) it must be rejected at the 0.05 level

(b) it may be rejected at the 0.05 level

(c) will not be rejected at the 0.05 level

(d) it must be rejected at the 0.025 level

Solution:

ANSWER: (c)

Question 3

If a test of hypothesis has a Type I error probability of 0.05, this means that if the null hypothesis is

(a) true, we don't reject it 5% of the time

(b) true, we reject it 5% of the time

(c) false, we don't reject it 5% of the time

(d) false, we reject it 5% of the time

Solution:

ANSWER: (b)

Question 4

For a two-sided test (sometimes also called two-tailed test), the null hypothesis will be rejected at the 0.05 level of significance if the value of the standardized test statistic z is:

(a) smaller than 1.96 or greater than – 1.96

(b) greater than –1.96 or smaller than 1.96

(c) smaller than –1.96 or greater than 1.96

(d) smaller than 1.645 or greater than – 1.645

Solution:

ANSWER: (c)

:-) As stated before, *the third step* in any hypothesis test is to find a *decision rule* that would enable you to make the correct decision, and there are two routes: Route A is only possible when you have a computer program and it would be to compute the *p-value*.

Route B is compulsory when you perform a hypothesis tests manually, and you have to obtain a *rejection region.* (For this module access to a computer is not compulsory, but you can still be tested on both routes in the examination.)

Although you spent so much time to laboriously do all the activities above, you must not come to the conclusion that we put more emphasis on route B than on route A! Understanding how a *p-value* is obtained is just as important as understanding how and where you must find the rejection region and interpreting a *p-value* correctly, is extremely important.

10.4.5 Example

Question 1

Suppose that when testing the hypothesis $H_0 : \mu = 75$ vs. $H_1 : \mu < 75$, the Z-value of the test statistic equals -2.42

(a) Show how you can compute the p-value by making use of Table 3 (i.e., without the use of a computer!).

(b) How can you describe the p-value?

Question 2

Suppose that when testing the hypothesis $H_0 : \mu = 75$ vs. $H_1 : \mu \neq 75$, the Z-value of the test statistic equals -2.42

(a) Show how you can compute the p-value by making use of Table 3 (i.e. without the use of a computer!).

(b) How can you describe the p-value?

Question 3

In order to determine the p-value, which of the following is not needed?

(a) The level of significance

(b) Whether the test is one- or two-sided

(c) The value of the test statistic

(d) All of the above

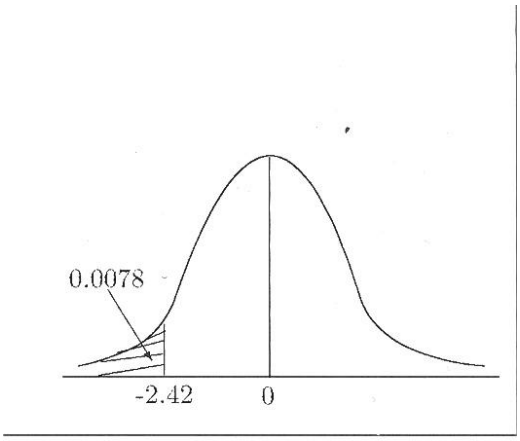
Solutions:

Question 1

(a) The only reason you need to know the two hypotheses is to have information about whether the *alternative is one- or two-sided*. So, instead of specifying $H_0 : \mu = 75$ vs. $H_1 : \mu < 75$, we could simply have stated that the hypothesis test has a left-sided alternative. The crucial information is that $Z = -2.42$ and you need to figure out what probability needs to be computed.

Remember that the reasoning behind the computation of a p-value is to obtain **a probability for the test statistic that is extreme under** H_0 (of course this will be in the direction of H_1 , suggesting that H_1 provides a better explanation of the data than H_0).

We have to find $p = P(Z \leq -2.42)$.



Thus $P(Z \leq -2.42) = 0.0078$

$p = P(Z \leq -2.42) = 0.0078$

(b) Since $p < 0.01$ the test is **highly significant** or there is overwhelming evidence to infer that the alternative hypothesis is true.

Question 2

(a) From the long explanation given in question 1 it is clear that we now have to find the p-value for a *two-sided test.*

We have to find $p = P(Z \leq -2.42) + P(Z \geq 2.42) = 0.0078 + 0.0078 = 0.0156$, since the standard normal curve is symmetric around 0.

(b) Since $0.015\,6 < p < 0.05$ the test is **significant** or there is strong evidence to infer that the alternative hypothesis is true.

Question 3

ANSWER: Option (a).

10.5 Calculating the probability of a type II error

Please note that you will not be examined in this section on how to **manually compute the probability of a Type II error,** but you should know and understand the following concepts:

Definition of a Type II error

Definition of the power of a test

The effect on a Type II error if α is changed

The effect on a Type II error if the sample size is changed

You should also be able to interpret computer output regarding the power of a test and a Type II error.

10.5.1 Example

Question 1

The power of a test is the probability of making

(a) a correct decision when the null hypothesis is false

(b) a correct decision when the null hypothesis is true

(c) incorrect decision when the null hypothesis is false

(d) incorrect decision when the null hypothesis is true

Question 2

The power of a test is the probability that it will lead us to

(a) reject the null hypothesis when it is true

(b) reject the null hypothesis when it is false

(c) fail to reject the null hypothesis when it is true

(d) fail to reject the null hypothesis when it is false

Question 3

For a given level of significance, if the sample size increases, the probability of a Type II error will

(a) remain the same

(b) increase

(c) decrease

(d) be equal to 1.0 regardless of α

Question 4

The power of a test is measured by its capability of

(a) rejecting a null hypothesis that is true

(b) not rejecting a null hypothesis that is true

(c) rejecting a null hypothesis that is false

(d) not rejecting a null hypothesis that is false

Question 5

If the probability of committing a Type I error for a given test is to be decreased, then for a fixed sample size n

(a) the probability of committing a Type II error will also decrease

(b) the probability of committing a Type II error will increase

(c) the power of the test will increase

(d) a one-sided test must be utilized

Question 6

The power of a test is denoted by

(a) α

(b) β

(c) $1-\alpha$

(d) $1-\beta$

Solutions:

Question 1

ANSWER: Option (a)

Question 2

ANSWER: Option (b)

Question 3

ANSWER: Option (c)

Question 4

ANSWER: Option (c)

Question 5

ANSWER: Option (b)

Question 6

ANSWER Option (d)

10.6 The road ahead

The topics of Table 11.3 hopefully whetted your appetite for "Techniques To Follow". [**:-)** Very similar to the "Forthcoming Attractions" of the movies!] It is a smaller snapshot of the "mother of all summaries", given on the inside cover of your textbook. You will be amazed to see how everything falls into place as we proceed through the rest of the chapters.

The last sub-section of this section has the heading **“Derivations”**. It is an excellent summary in a nutshell of what confidence intervals and hypothesis testing is all about.

Key Terms/Symbols

null hypothesis

alternative hypothesis

significance level

test statistic

rejection region

p-value

Type I error

Type II error

one-sided testing

two-sided testing

power of a hypothesis test

10.7 Study Unit 10: Summary

I. The *objective of hypothesis testing* is to check the validity of a statement about a population parameter.

II. The *steps in conducting a test of hypothesis* are as follows:

A. State the null hypothesis (H_0) and the alternative hypothesis (H_1) .

B. Select the level of significance.

1. The level of significance is the likelihood of rejecting a true null hypothesis.
2. The most frequently used significance levels are 0.01, 0.05, and 0.10, but any value between 0 and 1.00 is possible.

C. Select the test statistic.

1. A test statistic is a value calculated from sample information used to determine whether to reject the null hypothesis.
2. Two test statistics were considered in this chapter.
   a. The standard normal distribution (the z-distribution) is used when the population follows the normal distribution and the population standard deviation is known.
   b. The t-distribution is used when the population follows the normal distribution and the population standard deviation is unknown.

D. State the decision rule.

1. The decision rule indicates the condition or conditions when the null hypothesis is rejected.
2. In a two-tailed test, all of the rejection region is evenly split between the upper and lower tails.
3. In a one-tailed test, all of the rejection region is in either the upper or the lower tail.

E. Select a sample, compute the value of the test statistic, make a decision regarding the null hypothesis, and interpret the results.

III. A *p-value* is the probability that the value of the test statistic is as extreme as the value computed, when the null hypothesis is true.

IV. When testing a hypothesis about a population mean, if the population standard deviation, σ , is known, the test statistic is the standard normal distribution and is determined from

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}.$$

References

Keller, Gerald et al. (2005) *Instructor's Suite CD for the Student Edition of Statistics for Management and Economics*, Belmont, CA USA: Duxbury, Thomson.

Study Unit 11

INFERENCE ABOUT A POPULATION

11.1 Learning outcomes

After completing this unit you should be able to

1. explain the difference between the normal distribution and the t-distribution.
2. obtain the rejection region for any given significance level for the t-test.
3. derive a confidence interval estimation for the population mean when the population variance is unknown.
4. perform a complete hypothesis test for the population mean when the population variance is unknown.
5. derive a confidence interval for the total amount when the size N of a finite population is known.
6. define the finite population correction factor and do you know how to apply it?
7. obtain the rejection region for any given significance level for the χ^2-test.
8. interpret computer output regarding inferences about a population variance.
9. derive a confidence interval estimation for p, the population proportion.
10. perform a complete hypothesis test for the population proportion.
11. select the correct sample size to estimate a population proportion for an allowable sampling error and confidence level.

11.2 Introduction

You have learned in the previous three study units that statistical inference could be either in the form of hypothesis testing or the computation of a confidence interval. In both scenarios the starting point was to have

a *known sampling distribution* of a *sample statistic* and to manipulate a probability expression to either create a statistical test or a lower and an upper bound for the interval. We also pointed out to you that hypothesis testing and the computation of a confidence interval for an unknown population parameter usually go hand in hand. Keller first derived a confidence interval for a *population mean* and followed it up by deriving the corresponding hypothesis tests for a population mean.

Are you now anticipating that we are going to perform hypothesis tests (and derive their corresponding confidence intervals) for **other population parameters**?

Well, yes and no!! In the first part of this study unit you will find that Keller still dwells on the *population mean.* Why? Because he explains a slightly different application (almost repeating everything!) necessitated by the more realistic setup where we assume that the **population variance is unknown**.

It is only in the last part of this study unit that you will encounter two "new set-ups"; we discuss inference about a population variance and inference about a population proportion.

11.3 Inference about a population mean when the standard deviation is unknown

If you page back to the previous units you will notice that our "workhorse" was the standardised $z_{\overline{x}}$ variable.

The problem with $z_{\overline{x}} = \dfrac{(\overline{x} - \mu_{\overline{x}})}{\dfrac{\sigma}{\sqrt{n}}}$ to be a "workable" variable is that we need to **know the value of** σ. In real life this is usually not the case! The alternative is to replace σ with it's unbiased estimator s, the sample standard deviation. **This, however, changes the nature of the theoretical distribution.**

So, for inference about a population mean, $z_{\overline{x}} = \dfrac{(\overline{x} - \mu_{\overline{x}})}{\dfrac{\sigma}{\sqrt{n}}}$ **is replaced by** $t_{\overline{x}} = \dfrac{(\overline{x} - \mu_{\overline{x}})}{\dfrac{s}{\sqrt{n}}}$ which has a t-distribution with $\upsilon = n - 1$ degrees of freedom. Please note: The t-distribution has a parameter called the **degrees of freedom** (sometimes indicated as "df" but mostly indicated as υ, the Greek letter *nu*).

You learned about "Other Continuous Distributions", where **Student's t-distribution** (which we usually simply just call the *t-distribution*) was explained. You need not know the formula of the density function, but you must feel very comfortable with the table of critical values of t.

Test statistic for μ when σ is unknown

When the population standard deviation is unknown and the population is normal , the test statistic for the testing hypotheses about μ is:

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

which is the student t-distributed with $\upsilon = n - 1$ degrees of freedom

Confidence Interval Estimator of μ when σ is unknown is given by

$$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$
$$\upsilon = n - 1$$

11.4 Quick revision

The t-distribution has a different probability density graph for each different value of υ and forms a family of similar probability distributions that are symmetrically distributed about zero but have thicker (or wider) tails than the standard normal distribution. The difference between the t-distribution and the $n(0; 1)$ distribution becomes smaller and smaller as the number of the degrees of freedom increases.

Figure 11.1

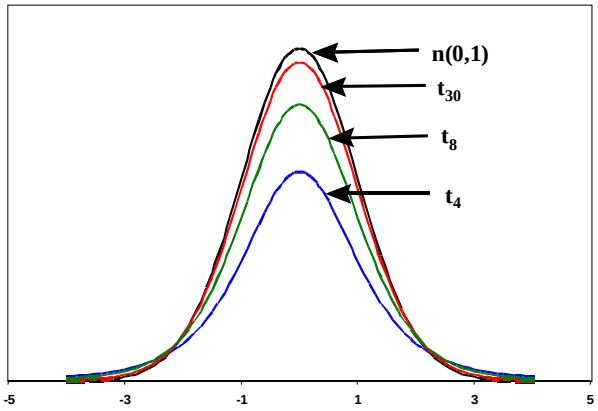


Figure 11.1 compares the $n(0; 1)$ distribution with three graphs of the t-distribution having $\upsilon = 30$, $\upsilon = 8$ and $\upsilon = 4$ degrees of freedom respectively. Please note that the degrees of freedom is indicated as a subscript, i.e., t_{30} means a t-distribution having 30 degrees of freedom.

Do you recall from study unit 8 that, according to convention, if the theoretical model for a variable x is a normal distribution with mean μ and variance σ^2 we write it as $x \sim n(\mu; \sigma^2)$? This would imply that we write $z_{\overline{x}} \sim n(0; 1)$. Similarly if t has a t-distribution with $\upsilon = n - 1$ degrees of freedom we write $t \sim t_{n-1}$.

The point of drawing graphs such as figure 12.1 is that we usually associate areas (which represent probabilities) with density functions.

11.5 Example

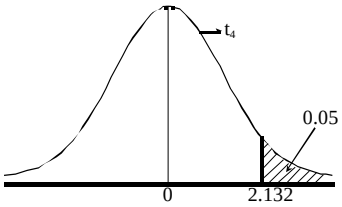
Find the following t-values:

(a) $P(t_4 \geq t) = 0.05$

Solution:

$P(t_4 \geq t) = 0.05 \implies$ we have to do no "manipulations" and can look up $t_{0.05,\,4}$ directly.

$P(t_4 > t_{0.05,\,4}) = 0.05 \implies t = 2.132$

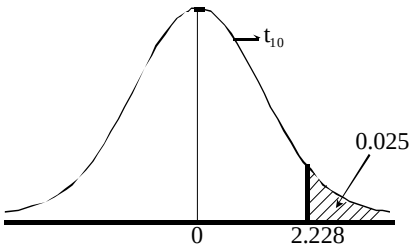


(b) $P(t_{10} \geq t) = 0.025$

Solution:

$P(t_{10} \geq t) = 0.025 \implies$ we have to do no "manipulations" and can look up $t_{0.025,\,10}$ directly.

$P(t_{10} > t_{0.025,\,10}) = 0.025 \implies t = 2.228$

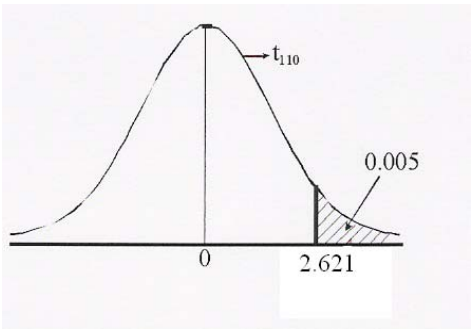


(c) $P(t_{110} \geq t) = 0.005$

Solution:

$P(t_{110} \geq t) = 0.005 \implies$ we have to do no "manipulations" and can look up $t_{0.005,\,110}$ directly.

$P(t_{110} > t_{0.005,\ 110}) = 0.005 \implies t = 2.621$



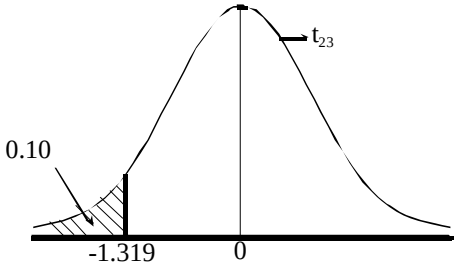
(d) $P(t_{23} \leq -t) = 0.10$

Solution:

$P(t_{23} \leq -t) = 0.10$ is not in the standard format and we have to do some "manipulations".

If $P(t_{23} \leq -t) = 0.10 \implies P(t_{23} \geq +t) = 0.10$ and now we can look up $t_{0.10,\ 23}$.

$P(t_{23} > t_{0.10,\ 23}) = 0.10 \implies t = 1.319$

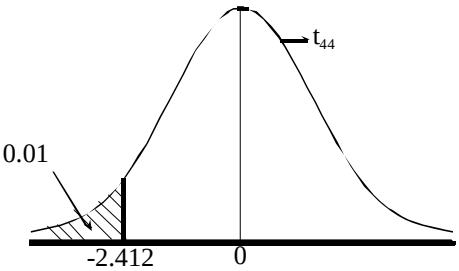


(e) $P(t_{44} \geq t) = 0.01$

Solution:

$P(t_{44} \geq t) = 0.99$ is not in the standard format and we have to do some "manipulations".

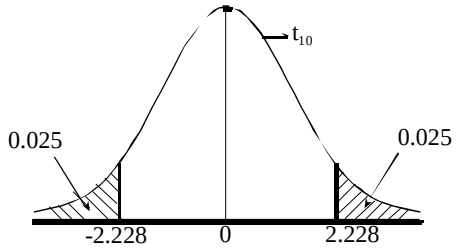
If $P(t_{44} \leq t) = 0.01 \implies t$ **must be negative** and we have to look up $t_{0.01,\ 44} = \pm 2.412$



(f) $P(-t \leq t_{10} \leq t) = 0.95$

Solution:

If $P(-t \leq t_{10} \leq t) = 0.95 \implies P(t_{10} \geq t) = 0.025$ and we have the same t-value as in question (b).

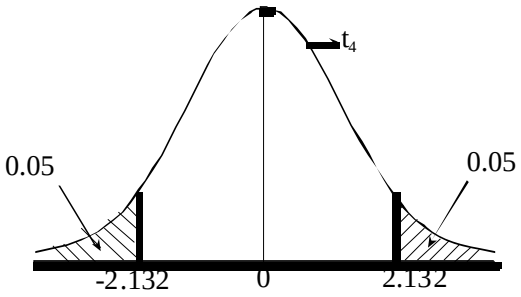


(g) $P(-t \leq t_4 \leq t) = 0.90$

Solution:

If $P(-t \leq t_4 \leq t) = 0.90 \implies P(t_4 \geq t) = 0.05$ and we have the same t-value as in question (a).

$t = 2.132$



11.6 Example

Question 1

Suppose we draw the following random sample from a population with a normal distribution:

18.2 19.6 24.4 16.3 17.3 20.5 21.7 19.2 22.5 20.1 23.5

(a) Compute a 99% confidence interval estimate for the unknown mean μ.

Solution:

You must try to have at least a *scientific pocket calculator* with *statistical functions* that will enable you to compute the following **sample statistics**:

$\overline{x} = \dfrac{\Sigma x_i}{n} = \dfrac{223.3}{11} = 20.3$ and

$s = \sqrt{\dfrac{\Sigma (x_i - \overline{x})^2}{(n-1)}} = \sqrt{\dfrac{65.04}{10}} = \sqrt{6.504} = 2.550\,29.$

For a 99% confidence interval we need $t_{\frac{\alpha}{2};(n-1)} = t_{\frac{0.01}{2};\,(11-1)} = t_{0.005\,;10} = 3.169$. The interval is computed as

$$\begin{aligned}\overline{x} \pm (t_{\frac{\alpha}{2};(n-1)})\frac{s}{\sqrt{n}} &= 20.3 \pm (3.169)\frac{2.550\,29}{\sqrt{11}} \\ &= 20.3 \pm 2.436\,8 \\ &= (17.863\,2;\ 22.736\,8).\end{aligned}$$

We are 99% confident that the unknown mean will be between 17.863 2 and 22.736 8.

(b) Test $H_0: \mu = 20$ against
$H_1: \mu \neq 20$ at the 1% level of significance.

Solution:

To perform a hypothesis test, we will employ our normal "three steps" as discussed in section 11.3 of study unit 11.

Step 1: **Specify the hypotheses** to be tested:. $H_0: \mu = 20$ against
$H_1: \mu \neq 20.$

Step 2: Decide on a **test statistic** and compute the value:

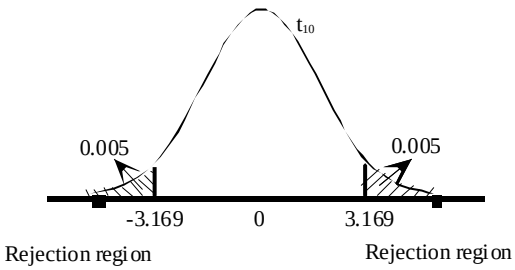
Now in this application our test statistic is $t_{\overline{x}} = \dfrac{(\overline{x} - \mu_{\overline{x}})}{\frac{s}{\sqrt{n}}} = \dfrac{20.3 - 20}{\frac{2.550\,29}{\sqrt{11}}} = 0.390\,1$

Step 3: Find a **decision rule** that would enable you to make the correct decision, and there are two routes:

Route A is only possible when you have a computer program to compute the *p*-value for a *t*-distribution.

Route B is compulsory when you perform a hypothesis test manually, and you have **to obtain a rejection region**.

At the 1% level of significance, using the critical values of the *t*-distribution, we obtain $t_{0.005\,;10} = 3.169$.



Conclusion: Since the computed value of $t_{\overline{x}}$ falls outside the rejection region, $(-3.169 < 0.3902 < 3.169)$ we **cannot reject the null hypothesis,** and conclude that there is not enough statistical evidence to infer that the null hypothesis is not true.

Question 2

(*Refer to question 3 of the self-correcting exercises for study unit 10.*)

Assume that the number of cars sold annually by used cars salespeople, is normally distributed. (We do not know what standard deviation of the population is.) A random sample of 21 used cars salespeople was drawn and the number of cars that each person sold are listed below.

79	43	58	66	101	63	79
33	58	71	60	100	74	55
88	70	69	82	61	76	63

(a) Compute the 95% confidence interval estimate of the population mean.

Solution:

We now have to **assume that σ is unknown** and cannot use the value $\sigma = 15$ as we previously did in question 3 of the self-correcting exercises for study unit 10.

We need to compute $\overline{x} = \dfrac{\sum x_i}{21} = \dfrac{1\,449}{21} = 69$ and $s = \sqrt{\dfrac{\Sigma (x_i - \overline{x})^2}{(n-1)}} = 16.477\,26$

An interval estimate for a population mean is computed as $\overline{x} \pm t_{\alpha/2}\dfrac{s}{\sqrt{n}}$. From table 8.2 we find the critical value associated with a 95% confidence level as $t_{\alpha/2\,,\,n-1} = t_{0.025\,,\,20} = 2.086$.

$$\begin{aligned} \overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} &= 69 \pm (2.086)\frac{16.477\,26}{\sqrt{21}} \\ &= 69 \pm 7.500\,5 \\ &= (61.499\,5;\quad 76.500\,5). \end{aligned}$$

(b) How do you interpret this interval?

Solution:

The confidence interval estimate for μ is $(61.499\,5;\quad 76.500\,5)$. If we would repeatedly draw samples of size 21 from this population, 95% of the values of $\overline{x}$ will produce an interval estimate that would include μ. Only 5% of the values of $\overline{x}$ will produce an interval estimate that would not include μ.

(c) Test $H_0: \mu = 70$ against
$H_1: \mu < 70$ at the 5% level of significance.

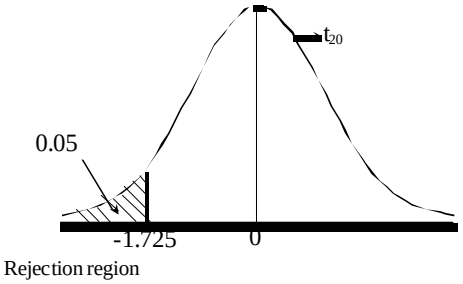
Solution:

Step 1: **Hypotheses** to be tested: $H_0: \mu = 70$ against
$H_1: \mu < 70.$

Step 2: **Test statistic**: $t_{\overline{x}} = \dfrac{(\overline{x} - \mu_{\overline{x}})}{\dfrac{s}{\sqrt{n}}} = \dfrac{69 - 70}{\dfrac{16.477\,26}{\sqrt{21}}} = -0.278\,1$

Step 3: **Decision rule:** We need to have a one-sided rejection region.

At the 5% level of significance, using the critical values of the t-distribution, we obtain $t_{0.05;20} = 1.725$. For left-sided testing we have the following rejection region:



Conclusion: Since the computed value of $t_{\bar{x}}$ falls outside the rejection region, $(-0.278\,1 > -1.725)$ we **cannot reject the null hypothesis,** and conclude that there is not enough statistical evidence to infer that the null hypothesis is not true.

:-) The following activity is optional but we are urging you to try your best to have at least one session on a computer where you perform a t-test and a t-estimate for the mean. If it is not possible, at least read through my comments and the output and compare it to your manual computations.

11.7 Example

Question 1

Use either *EXCEL* or *MINITAB* and repeat question 1 of Example 11.6.

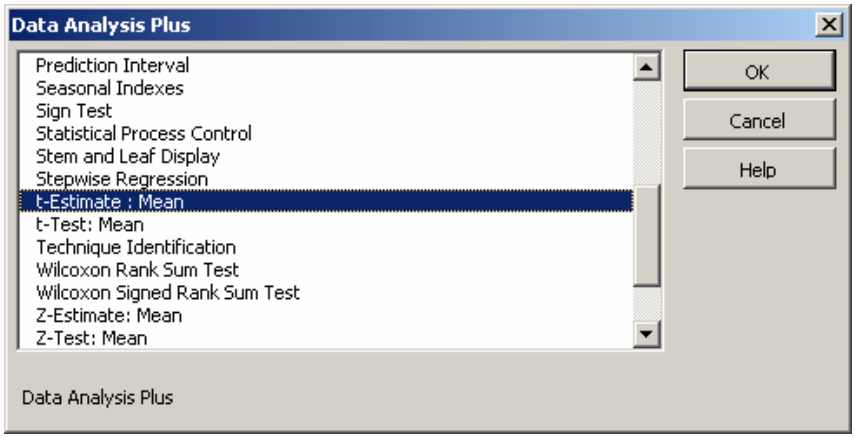
Question 2

Use either *EXCEL* or *MINITAB* and repeat (a) and (c) of question 2 of Example 11.6.

Solutions:

Question 1

(a) This is what the Menu looks like to select a confidence interval:



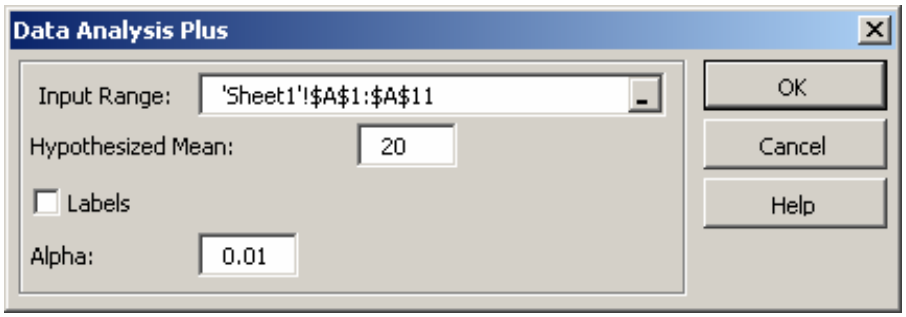
In the dialog-box you will choose *Alpha = 0.01* to obtain a 99% confidence interval.

t-Estimate: Mean	
	Column 1
Mean	20.3
Standard Deviation	2.5503
LCL	17.86315
UCL	22.73685

This output confirms our manually computed confidence interval of (17.86315 ; 22.73685).

(b) In the dialog-box you specify μ_0 (in other words the value of the mean under the null hypothesis) as well as the significance level *Alpha = 0.01*.

Input:



One cannot specify the direction of the test and hence the **computer output always gives both the one-sided and the two-sided test**. You have to select the appropriate output. Please note that all the

sample statistics check with our manually computed values (where "**t Stat**" is the computed value of $t_{\overline{x}} = \frac{(\overline{x} - \mu_{\overline{x}})}{\frac{s}{\sqrt{n}}} = 0.3901$). **All you have to do is to make the final conclusion. The computer does not do this!**

Output:

t-Test: Mean	
	Column 1
Mean	20.3
Standard Deviation	2.5503
Hypothesized Mean	20
df	10
t Stat	0.3901
P(T<=t) one-tail	0.3523
t Critical one-tail	2.7638
P(T<=t) two-tail	0.7046
t Critical two-tail	3.1693

The computer output always includes a p-value and to make the final conclusion you may either use the p-value or the traditional "rejection region" as we do manually. "**P(T<=t) one-tail"** is the one-sided p-value and it is not necessarily a correct mathematical statement! In fact, the p-value (which is always a probability) is $P(T \geq 0.3901) = 0.3523$. [**:-)** Do you understand why it is vital to understand what you are doing before you leap into computer support?]

The main thing is to realise that this probability is not sufficiently small to reject H_0.

"**P(T<=t) two-tail"** is the two-sided p-value and is once again not necessarily a correct mathematical statement! In fact, $P(T \geq 0.3901) + P(T \leq -0.3901) = 0.3523 + 0.3523 = 0.7046$.

Question 2

(a)

t-Estimate: Mean

	Column 1
Mean	69
Standard Deviation	16.4773
LCL	61.49971
UCL	76.50029

(b) The same interpretation as for the manual solution.

(c)

t-Test: Mean

	Column 1
Mean	69
Standard Deviation	16.4773
Hypothesized Mean	70
df	20
t Stat	-0.2781
P(T<=t) one-tail	0.3919
t Critical one-tail	1.7247
P(T<=t) two-tail	0.7838
t Critical two-tail	2.086

"**P(T<=t) one-tail"** is the one-sided p-value and for this test it is $P(T \leq -0.2781) = 0.3919$. This probability is not sufficiently small to reject H_0.

Please note that EXCEL always gives the critical value for a **right-sided test,** even though we need to test left-sided. So, strictly speaking, the output should have stated "**t Critical one-tail**" $= -1.7247$.

11.8 Inference about the mean: What else need you keep in mind?

Checking the Required Conditions

It might seem innocent and unimportant at this stage but for each and every statistical test you are going to perform you must always keep in mind that the test was developed on the assumption of certain theoretical conditions. So, strictly speaking, you should first check whether these conditions are met before you proceed with any test. The conditions for the t-test (and of course the forerunner Z-test) is that the sample must come from a **normal** population. If a test is sensitive to the condition of normality, there exist additional "preliminary tests" where we can formally test for normality, but since the t-test is robust to the condition of normality it is sufficient to draw a histogram and inspect the shape. The shape must be more or less like a bell, in other words symmetrical and with one peak to accept normality.

Please note: The underlying assumption of most of the sampling distributions derived in this module is that the data follow a *normal distribution.* What happens if we cannot assume normality? There is a special branch in Statistics, called *distribution-free methods*, where tests have been developed or designed when we cannot assume normality. In other words, tests were designed to apply to data that is not based on a specific distribution – hence the name distribution-free methods. An alternative name, which many statistical applications or packages use, is the term *non-parametric tests*. Most of these test are designed for use in nominal and/or ordinal data but they can also be applied to quantitative data. Keller treats this in chapter 19 of the textbook.

11.9 Estimating the totals of finite populations

In many practical situations we are more interested in a **total amount** than in the average. For example, the municipality of Johannesburg would be more interested in the total amount of electricity used than in the average amount of electricity used. Based on a sample (of size n) where they compute a confidence interval for *the mean*, they simply have to multiply by N (the finite population size) to obtain a good interval estimate for the *total amount*.

$$\text{Confidence interval for the } \textit{total amount:} \quad N\left[\overline{x} \pm (t_{\frac{\alpha}{2};(n-1)})\frac{s}{\sqrt{n}}\right].$$

Finite Population Correction Factor

When estimating the mean when the population is small, we need to multiply the standard error of the mean with the "Finite Population Correction Factor" (FPCF) $= \sqrt{\frac{N-n}{N-1}}$.

The confidence interval for the mean is then computed as $\overline{x} \pm (t_{\frac{\alpha}{2};(n-1)})\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

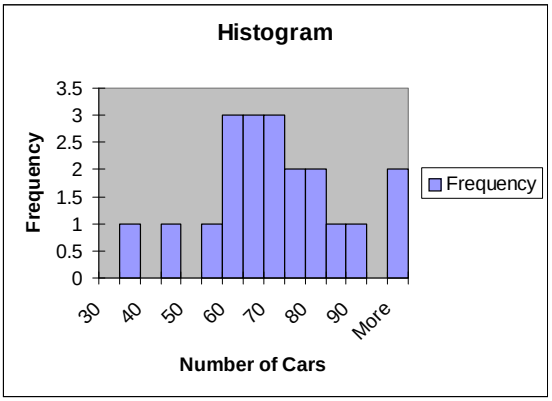
This would imply that if we are interested in estimating *the total* of a finite population (when N is small), we need to incorporate FPCF, hence a good interval estimate for the *total amount* is

$$N\left[\overline{x} \pm (t_{\frac{\alpha}{2};(n-1)})\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right].$$

11.9.1 Example

Use either *EXCEL* or *MINITAB* to **check for the normality assumption** of question 2 of activity 12.2 (and by implication question 2 of activity 12.3).

Solution:



Even though we have a very small sample, the distribution could pass as normal [**:-)** with a little bit of extra imagination...].

11.10 Inference about a population variance

Statistic and sampling distribution

The statistic is given by : $\chi^2 = \dfrac{(n-1)\,s^2}{\sigma^2}$ is called the chi-squared statistic with degree of freedom $\nu = n-1$

Confidence interval Estimator of σ^2

Lower confidence limit (LCL) $= \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}$

Lower confidence limit (UCL) $= \frac{(n-1)s^2}{\chi^2_{\frac{1-\alpha}{2}}}$

The **Chi-Squared Distribution** was explained. You need not know the formula of the density function, but need to understand the applications which means that you must feel very comfortable with the table of critical values of χ^2. You must be able to obtain the *rejection region* for any given significance level for the χ^2*-test* and understand the difference between one-sided and two-sided testing. [**:-)** It helps to draw little sketches.]

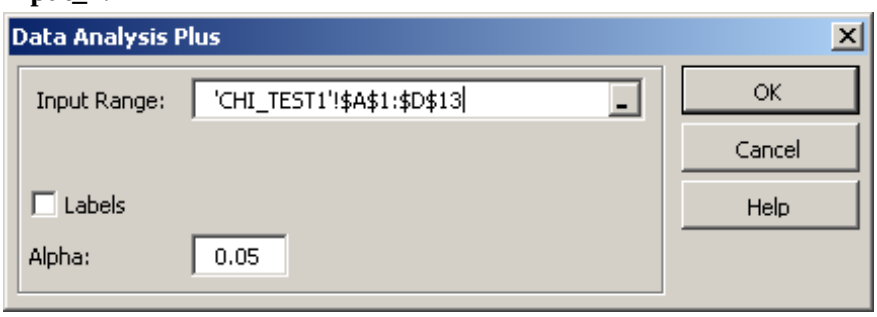
We are not expecting of you, i.e. you will not be examined on "how to" **manually compute a confidence interval for an unknown population variance or how to perform a hypothesis test,** but you should understand the steps and be able to **interpret computer output** regarding inference about a population variance.

11.10.1 Example

Refer to the "*used cars salespeople*" example (in other words question 2 of activity 4.2).

Consider the following computer input and output regarding this sample:

Input_1:

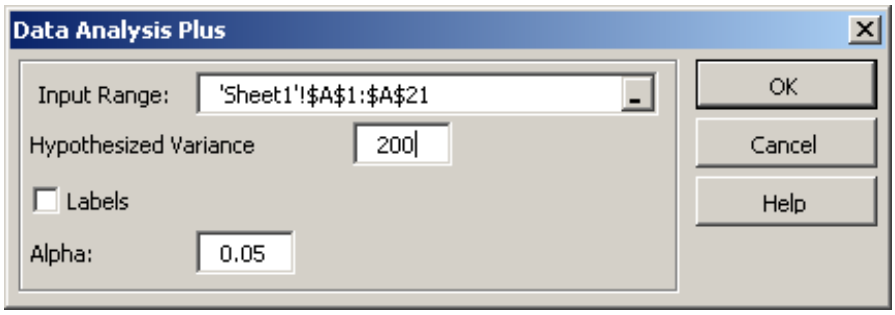


Output_1:

Chi Squared Estimate: Variance

	Column 1
Sample Variance	271.5000
df	20
LCL	158.9132
UCL	566.1689

Input_2:



Output_2:

Chi Squared Test: Variance

		Column 1
Sample Variance		271.5000
Hypothesized Variance		200
df		20
chi-squared Stat		27.1500
P (CHI<=chi) one-tail		0.1311
chi-squared Critical one tail	Left-tail	10.8508
	Right-tail	31.4104
P (CHI<=chi) two-tail		0.2622
chi-squared Critical two tail	Left-tail	9.5908
	Right-tail	34.1696

Question 1

Which one of the following statements is **false**?

(a) The computed value for the sample variance is 271.500 which is an unbiased point estimate for the unknown population variance.

(b) We are 95% sure that unknown population variance lies in the interval (158.9132 ; 566.1689).

(c) With the EXCEL input for the chi-squared estimate of the variance, we have an option to specify alpha.

(d) If we want to compute the interval manually we would need to look up the critical values of χ^2 in Table 5 (Appendix B) of Keller.

(e) If we specify alpha as $\alpha = 0.05$ it implies we have a 90% two-sided interval.

Question 2

Which one of the following statements is **false**?

(a) With the EXCEL input for the chi-squared Test of the Variance, we have an option to specify the value of σ^2 to be tested.

(b) For a two-sided hypothesis test we are testing $H_0: \sigma^2 = 200$
$H_1: \sigma^2 \neq 200.$

(c) The chi-square test statistic for this hypothesis test has 20 degrees of freedom.

(d) The computed value of the chi-square test statistic for this hypothesis test is 27.1500.

(e) EXCEL gives two critical values: Left-tail and Right-tail which will differ from the Critical Values of χ^2 in Table 5 of Keller.

Question 3

Which one of the following statements is **false**?

(a) The p-value for the hypothesis test is 0.1311.

(b) Since $27.1500 < 34.1696$ we cannot reject H_0 in favour of $H_1 : \sigma^2 \neq 200$.

(c) Since $0.1311 > 0.05$ we cannot reject H_0 in favour of $H_1 : \sigma^2 > 200$.

(d) For a two-tailed test, the right-sided critical value of χ^2 is 34.1696 which means that $P(\chi^2 > 34.1696) = 0.05$.

(e) For a two-tailed test, the left-sided critical value of χ^2 is 9.5908 which means that $P(\chi^2 < 9.5908) = 0.025$.

Solutions:

Question 1

ANSWER: Option (e).

If we specify alpha as $\alpha = 0.05$ it implies we have a 95% two-sided interval.

Question 2

ANSWER: Option (e).

Each and every critical values that EXCEL gives as part of the output must (and will!) be exactly the same as the Critical Values of χ^2 in Table 5 of Keller.

For two-sided testing at the **5% level** of significance, the critical values are $\chi^2_{0.975\,;\,20} = 9.5908$ and $\chi^2_{0.025\,;\,20} = 34.1696$.

For left-sided testing at the **5% level** of significance the critical value is $\chi^2_{0.95\,;\,20} = 10.8508$, and

for right-sided testing at the **5% level** of significance the critical value is $\chi^2_{0.05\,;\,20} = 31.4104$.

Question 3

ANSWER: Option (d)

If we have a two-sided test and $\alpha = 0.05$ we need a right-sided critical value such that $P(\chi^2 >$ right-sided critical value$) = \frac{\alpha}{2}$.

Hence $P(\chi^2 > 34.1696) = \frac{0.05}{2} = 0.025$.

11.10.2 Inference about a population proportion

Going through this section, you must have noticed that the author of our textbook picked up speed! He assumes that you understand the *principles of statistical inference* and now he simply applies it to a different set-up.

We are now sampling from a population where a *proportion of the population* has a certain attribute. Usually this proportion, p, is an unknown parameter and we estimate it as $\widehat{p} = \frac{X}{n}$ (i.e. the proportion in the sample having this certain attribute). This value of $\widehat{p}$ could vary from sample to sample but it will on average (over an infinite number of samples) equal the population proportion, p. Thus we can say that $E(\widehat{p}) = p$, or in statistical jargon that $\widehat{p} = \frac{X}{n}$ is an *unbiased estimator* of p. In *sampling theory* we are interested in the *distribution of* $\widehat{p}$, which we have treated in detail in study unit 8.

[**Quick revision:** In chapter 6 you learned about the binomial distribution as a theoretical model for the discrete variable X, where X denoted the number of successes in n independent Bernoulli trials. Our statistical shorthand notation for this is $X \sim b(n;p)$. From chapter 6 we have also seen that the form of the binomial distribution depends on the value of p and that of the sample size n. The closer p gets to 0.50 the more symmetrical the distribution becomes but if the sample size is small the distribution could still be skew. However, for large samples the skewness becomes less of a factor even when p is not so close to 0.50. Furthermore, if X represents the number of successes in n independent Bernoulli trials, this binomial distribution can be **approximated by the normal distribution** if n is large and p is not too close to *zero* or not too close to *one*. Our rule of thumb was that np and $n(1-p)$ must be greater than or equal to 5.

In section study unit 8, we summarised the *sampling distribution* of $\widehat{p} = \dfrac{X}{n}$ as follows:

$$\frac{X}{n} \sim \text{approx. } n(p\ ;\ \frac{p(1-p)}{n}) \implies z_{\widehat{p}} = (\frac{\widehat{p} - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}}) = \frac{\widehat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

end of revision.]

So, the *standardised sampling distribution* of $\frac{X}{n}$ has an approximate $n(0;1)$ distribution, i.e.

$$z_{\widehat{p}} = \frac{\frac{X}{n} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \approx n(0;1).$$

A-ha, and here we have a test statistic!

Deriving a confidence interval estimator of p :

Strictly speaking $var(\widehat{p}) = \dfrac{p(1-p)}{n}$ but when we compute a confidence interval, the value p is unknown and we use the estimated variance $\dfrac{\widehat{p}(1-\widehat{p})}{n}$.

Therefore the confidence interval estimate for p is $\widehat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$.

Performing a hypothesis test for p :

Do you recall that hypothesis testing is where we *make a statement about an unknown parameter*? Thus the null hypothesis will specify a value for the unknown proportion p, indicated in general as p_0. The symbol p_0 implies a **known**, specified value under H_0 of the unknown p.

The difference between the confidence interval and the hypothesis test is that we will use $var(\widehat{p}) = \dfrac{p(1-p)}{n} = \dfrac{p_0(1-p_0)}{n}$ when we compute the value of the test statistic (and not $\dfrac{\widehat{p}(1-\widehat{p})}{n}$).

The test statistic therefore becomes $z = \dfrac{\frac{X}{n} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Selecting the sample size to estimate a proportion

You have seen in study unit 10 (section 10.4) that we can solve for n, the sample size, if the bound on the error of estimation is given. The same happens with this scenario. Once again the width of the interval is determined by the quantity that follows the $\pm$-sign, which in this application is $z_{\frac{\alpha}{2}}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$.

Setting the bound on the error of estimation, B, equal to $z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ and solving for n, we obtain:

$$n = \left(\frac{z_{\frac{\alpha}{2}}\sqrt{\widehat{p}(1-\widehat{p})}}{B}\right)^2.$$

11.10.3 Example

1. The production manager in an automobile plant is concerned with the number of cars that do not pass the final quality control inspection. In the last two hours he noted that only 90 out of 120 cars were acceptable. If these 120 cars can be considered to constitute a random sample, estimate with 90% confidence the proportion of all cars that **would not pass the final quality control inspection**.

   Solution:

   $\widehat{p}$ is the estimated proportion of all cars that would not pass the final quality control $= \frac{120-90}{120} = \frac{30}{120}$.

   $\widehat{p} \pm z_{0.05}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \frac{30}{120} \pm 1.645\sqrt{\frac{\frac{30}{120}(1-\frac{30}{120})}{120}} = 0.25 \pm 0.065 = (0.185; 0.315)$.

   Thus, LCL = 0.185, and UCL = 0.315. We are 90% confident that the proportion of all cars that would not pass the final quality control inspection, lies between 0.185 and 0.315.

2. After a financial analysis, the general manager of a large company decided that **if more than 8% of potential buyers** of a new product purchase that product, the company would show a profit. In a preliminary survey of 500 potential buyers, 56 people say that they will buy the product.

   (a) State the appropriate null and alternative hypotheses.

   Solution:

   Step 1: $H_0 : p = 0.08$ vs $H_1 : p > 0.08$

   (b) Is there sufficient evidence at the 5% significance level that the product will produce a profit?

   Step 2: Test statistic: $z = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{56}{500} - 0.08}{\sqrt{\frac{0.08(1-0.08)}{500}}} = 2.637\,5$

   Step 3: Rejection region: Reject H_0 if $z > z_{0.05} = 1.645$

   Conclusion: Since $2.637\,5 > 1.645 \Rightarrow$ we reject H_0. Yes, there is sufficient evidence at the 5% significance level that the product will produce a profit.

3. The director of the School of Science is interested in estimating the proportion of students entering the College who will choose Statistics as major. A preliminary sample indicates that the proportion will be around 0.25. Therefore, what size sample should the director take if she wants to be 95% confident that the estimate is within 0.10 of the true proportion?

Solution:

The bound on the error of estimation, B, is specified as 0.10. For 95% confidence we use $z_{\alpha/2} = 1.96$ and $\widehat{p} = 0.25$.

$$\text{Thus,} \quad n = \left(\frac{z_{\frac{\alpha}{2}}\sqrt{\widehat{p}(1-\widehat{p})}}{B}\right)^2$$

$$= \left(\frac{1.96\sqrt{(0.25)(0.75)}}{0.10}\right)^2$$

$$= 72.03$$

She must take a sample of size $n = 73$ to estimate the proportion.

Key Terms and Symbols

t-distribution

degrees of freedom

Finite Population Correction Factor (FPCF)

chi-squared distribution

11.11 Study Unit 11: Summary

1. When testing a hypothesis about a *population mean*, if the population standard deviation is not known, s is substituted for σ. The test statistic is the t-distribution, and its value is determined from

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}.$$

The major characteristics of the t-distribution are as follows:

1. It is a continuous distribution.
2. It is mound-shaped and symmetrical.
3. It is flatter, or more spread out, than the standard normal distribution.
4. There is a family of t distributions, depending on the number of degrees of freedom.

2. When testing about a *population proportion*:

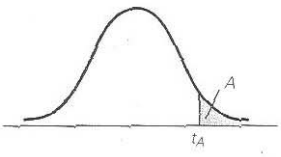
A. The binomial conditions must be met.

B. Both $n\pi$ and $n(1-\pi)$ must be at least 5.

C. The test statistic is

$$z = \frac{\rho - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}.$$

TABLE **4**
Critical Values of the Student t Distribution



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
65	1.295	1.669	1.997	2.385	2.654
70	1.294	1.667	1.994	2.381	2.648
75	1.293	1.665	1.992	2.377	2.643
80	1.292	1.664	1.990	2.374	2.639
85	1.292	1.663	1.988	2.371	2.635
90	1.291	1.662	1.987	2.368	2.632
95	1.291	1.661	1.985	2.366	2.629
100	1.290	1.660	1.984	2.364	2.626
110	1.289	1.659	1.982	2.361	2.621
120	1.289	1.658	1.980	2.358	2.617
130	1.288	1.657	1.978	2.355	2.614
140	1.288	1.656	1.977	2.353	2.611
150	1.287	1.655	1.976	2.351	2.609
160	1.287	1.654	1.975	2.350	2.607
170	1.287	1.654	1.974	2.348	2.605
180	1.286	1.653	1.973	2.347	2.603
190	1.286	1.653	1.973	2.346	2.602
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576