

Introduction to Applied Statistics

Introduction to Applied Statistics

Open Textbook Series in Statistics

WANHUA SU

MACEWAN OPEN BOOKS
EDMONTON



Introduction to Applied Statistics by Wanhua Su is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/), except where otherwise noted.

Cover image is a derivative of an original image designed by ©My Hanh Nguyen, All rights reserved.

Please cite this work as follows:

Su, W. (2024). Introduction to applied statistics: Open textbook series in statistics. MacEwan Open Books. <https://doi.org/10.31542/b.gm.5>

This book was produced with Pressbooks (<https://pressbooks.com>) and rendered with Prince.

Contents

Acknowledgments	xi
---------------------------------	----

[Chapter 1: Statistics, Data, and Data Presentation](#)

1.1 What is Statistics?	2
1.2 Data Collection	5
1.3 Variables and Data	14
1.4 Organizing Data	16
1.5 Shape of a Distribution	29
1.6 Learning Objectives Revisited	33
1.7 Review Questions	34
1.8 Assignment 1	37
Quiz 1	41

[Chapter 2: Descriptive Statistics](#)

2.1 Centre of a Distribution	43
2.2 Quartiles and Percentiles	49
2.3 Spread (Variation) of a Distribution	51
2.4 Five-Number Summary and Boxplot	55
2.5 Descriptive Measures for Population and Sample	62
2.6 Z-Score as a Measure of Relative Standing	63
2.7 Learning Objectives Revisited	66
2.8 Review Questions	67
2.9 Assignment 2	71
Quiz 2	74

Chapter 3: Probability Concepts

<u>3.1 Basic Concepts in Probability</u>	76
<u>3.2 Probability of An Event</u>	78
<u>3.3 Relationship Between Events and Venn Diagrams</u>	82
<u>3.4 Probability Rules</u>	85
<u>3.5 Conditional Probability and Independence</u>	87
<u>3.6 Summary of Probability Rules</u>	90
<u>3.7 Tree Diagrams</u>	92
<u>3.8 Counting Rules: Basic Counting Rule, Combination, and Permutation</u>	95
<u>3.9 Contingency Table: Joint and Marginal Probability</u>	101
<u>3.10 Learning Objectives Revisited</u>	102
<u>3.11 Review Questions</u>	103
<u>3.12 Assignment 3</u>	105
<u>Quiz 3</u>	108

Chapter 4: Discrete Random Variables

<u>4.1 Random Variable</u>	110
<u>4.2 Probability Distribution of a Discrete Variable</u>	112
<u>4.3 Defining Events Using Random Variable Notation</u>	115
<u>4.4 Mean and Standard Deviation of a Discrete Variable</u>	117
<u>4.5 Binomial Distribution</u>	123
<u>4.6 Learning Objectives Revisited</u>	130
<u>4.7 Review Questions</u>	131
<u>4.8 Assignment 4</u>	134
<u>Quiz 4</u>	136

Chapter 5: The Normal Distribution

<u>5.1 Density Curve</u>	138
<u>5.2 Normal Density Curve</u>	140
<u>5.3 Standard Normal Density Curve</u>	143
<u>5.4 Using the Standard Normal Table</u>	146

5.5 Working With Any Normal Distribution	152
5.6 Assessing Normality: Normal Probability Plot	155
5.7 Learning Objectives Revisited	160
5.8 Review Questions	161
5.9 Assignment 5	163
Quiz 5	166

[Chapter 6: Distribution of the Sample Mean and the Central Limit Theorem](#)

6.1 Parameter and Statistic	168
6.2 Distribution of the Sample Mean	170
6.3 Central Limit Theorem (CLT)	182
6.4 Learning Objectives Revisited	187
6.5 Review Questions	188
6.6 Assignment 6	192
Quiz 6	196

[Chapter 7: Confidence Interval for One Population Mean](#)

7.1 Confidence Interval When σ is Known	198
7.2 Confidence Interval When σ is Unknown	207
7.3 Learning Objectives Revisited	215
7.4 Review Questions	216
7.5 Assignment 7	220
Quiz 7	223

[Chapter 8: Hypothesis Tests for One Population Mean](#)

8.1 Hypotheses	226
8.2 Type I and Type II Errors	229
8.3 Main Idea Behind Hypothesis Tests for μ	232
8.4 Quantify the "Extremeness"	234
8.5 Hypothesis Tests for One Population Mean μ	238

8.6 Relationship Between Confidence Intervals and Hypothesis Tests	247
8.7 Learning Objectives Revisited	250
8.8 Review Questions	251
8.9 Assignment 8	254
Quiz 8	258

[Chapter 9: Inferences for Two Population Means](#)

9.1 Distribution of the Difference between Two Sample Means for Two Independent Samples	260
9.2 Two-Sample t Test and t Interval Based on Two Independent Samples	263
9.3 Paired t Test and Interval Based on Paired Sample	272
9.4 Learning Objectives Revisited	279
9.5 Review Questions	280
9.6 Assignment 9	285
Quiz 9	291

[Chapter 10: Inferences for Population Proportions](#)

10.1 Population Proportion and the Sample Proportion	293
10.2 Distribution of the Sample Proportion	294
10.3 One-Proportion z Interval	298
10.4 Margin of Error and Sample Size Calculation for Proportion	300
10.5 One-Proportion z Test for p	302
10.6 Inferences for Two Population Proportions	304
10.7 Learning Objectives Revisited	311
10.8 Review Questions	312
10.9 Assignment 10	315
Quiz 10	319

[Chapter 11: Chi-Square Procedures](#)

11.1 Introduction	321
11.2 Chi-Square Distribution	322

11.3 Chi-Square Goodness-of-Fit Test	325
11.4 Chi-Square Independence Test	330
11.5 Chi-Square Homogeneity Test	338
11.6 Learning Objectives Revisited	339
11.7 Review Questions	340
11.8 Assignment 11	344

[Chapter 12: One-way ANOVA](#)

12.1 Opening Example	348
12.2 Main Idea Behind One-Way ANOVA	350
12.3 F Distribution	355
12.4 One-Way ANOVA F Test	358
12.5 Learning Objectives Revisited	364
12.6 Review Questions	365
12.7 Assignment 12	369
Quiz 11	372

[Chapter 13: Descriptive and Inferential Methods in Simple Linear Regression](#)

13.1 Introduction	374
13.2 Least-Squares Straight Line	376
13.3 Prediction and Extrapolation	380
13.4 Outliers and Influential Observations	382
13.5 Correlation Coefficient r	384
13.6 The Coefficient of Determination	387
13.7 Simple Linear Regression Model (SLRM)	390
13.8 Inferences for the Parameters in SLRM	396
13.9 Confidence Intervals and Prediction Intervals	400
13.10 Learning Objectives Revisited	405
13.11 Review Questions	406
13.12 Assignment 13	408
Quiz 12	412

Appendix A: Formula Sheet	413
Appendix B: Statistical Tables	418
Appendix C: Lab Manual	433
Appendix D: Image Descriptions	434
Versioning History	460

Acknowledgments

Many individuals have made great efforts to create this open textbook. Mr. Jianfei Guan, an instructional designer from the eLearning team at MacEwan University, provided me with a lot of instructional guidance and advice when I developed the online version of STAT 151. Our most tremendous thanks go to Mr. John Fedoruk from MacEwan University and Dr. Hugh Chipman from Acadia University. Mr. Fedoruk is our internal reviewer who has spent over a hundred hours revising the course notes. Dr. Chipman, our external reviewer, provided a lot of valuable suggestions. Their volunteer work has significantly improved the quality of the book. The most special thanks go to Dylan Miller, our student assistant who published almost all materials on Pressbooks, and Clarissa Mewhort for her work on the image descriptions. We are grateful to Dr. Kathleen Lawry-Batty from MacEwan University, who revised and enriched the lab manual in R Commander and helped create Part B of the homework assignments. Our thanks are also extended to all faculty members and students at MacEwan University who used the course materials for their feedback. Lastly, we thank the “Open Textbook Pilot Project” and FAS Supplementary FD Fund for financial support; our wonderful librarians, Ms. Robyn Hall, Ms. Ali Foster, Dr. Eva Revitt, and Ms. Lori Walter, for their consistent support; our financial analyst Ms. Christine Zielinski for managing the account of this open textbook project; our Scholarly Communications Technician Ms. Penny Chu for substantially improving the quality of this open textbook; and Mr. Blair Moran and Ms. Carol Woo from the Center of Teaching and Learning for help with formatting equations and copy editing.

CHAPTER 1: STATISTICS, DATA, AND DATA PRESENTATION

Overview

This chapter introduces you to some basic terminology of statistics. Statistics is all about data. We briefly examine three fundamental data collection methods—sampling, designed experiments, and observational studies. We also show you how to classify data by type and choose the proper graphs to summarize different types of data.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Tell what statistics is and how useful it is to solve real-life and work issues.
- Identify the population and sample of a study.
- Identify whether a study is descriptive or inferential.
- Draw a simple random sample using the random-number table.
- Distinguish designed experiments and observational studies.
- Distinguish two types of data: quantitative and qualitative/categorical.
- Describe qualitative/categorical data using pie charts and bar charts.
- Describe quantitative data using stem-and-leaf.
- Describe the shape of a histogram.

1.1 What is Statistics?

Statistics is a branch of science dealing with the collection, analysis, interpretation, and presentation of masses of data. Therefore, statistics is all about data, and data is information organized in variables. Data are evidence and statistics are tools that provide valid and effective ways to collect evidence. Hence, statistics is widely used in many areas to find useful patterns, make predictions, and hence help in decision-making.

STAT 151 covers the three major topics of statistics. For now, we only provide the general idea and more details will be provided as the course progresses.

1. **Data collection** is about how to collect data. Common data collection methods include sampling surveys, observational studies, and designed experiments. Data collection is not a key focus of STAT 151, so students are only required to have a basic idea about simple random sampling, designed experiments, and observational studies.
2. **Data presentation** is about how to present data. It includes numerical methods such as tables, mean, standard deviation, and five-number summary; and graphical methods such as histograms, boxplots, scatter plots, pie charts, and bar charts. Most data presentation techniques are parts of the subfield of statistics known as descriptive statistics.
3. **Data analysis and interpretation** is the process of analyzing data and summarizing data so that useful information can be uncovered; such information acts as the basis for sound decision-making. Methods covered in STAT 151 include confidence intervals, hypothesis testing, and simple linear regression.

In the first part of the course, we will introduce some essential terminologies in statistics.

- **Population** is the collection of all individuals or items under consideration in a statistical study.
- **Sample** is part of the population from which information is obtained. It is a subset of the population.

Figure 1.1 illustrates the relationship between population and sample. The outside (bigger) ellipse represents the population and the inside (smaller) ellipse represents the sample. In general, we usually handle sampled data, since it is generally impractical or even impossible to collect data from the entire population.

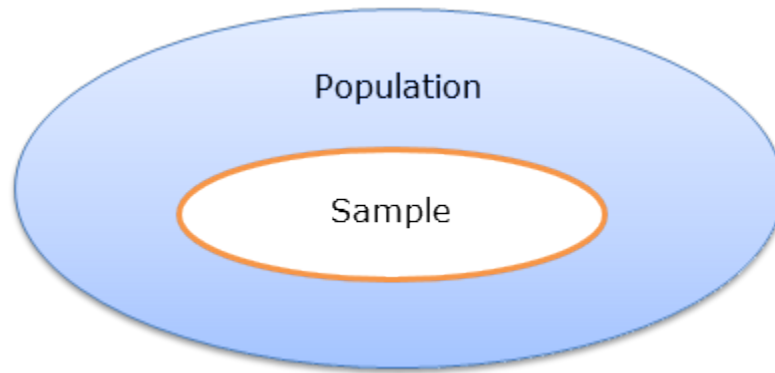


Figure 1.1: Relationship Between Population and Sample. [[Image Description](#)
([See Appendix D Figure 1.1](#))]

Basically, there are two types of statistics:

- **Descriptive statistics** consists of numerical and graphical methods for organizing and summarizing data. Note that descriptive statistics focuses only on data and do not generalize the conclusions from the sample to the population.
- **Inferential statistics** consists of methods for drawing conclusions about a population based on information obtained from sampled data. Inferential statistics includes estimation, decision making, prediction, and other generalizations about a population. For inferential studies, look for the key words such as “**estimate for all**” or “**prediction for all**.”



Activity

Exercise: Basic Concepts

Complete this exercise to see if you understand these basic concepts.

A random sample of 100 students studying at MacEwan University yields 65/35 as an estimate of the ratio of females to males for all students studying at MacEwan.

Answer the following questions related to the above study.

1. What is the population?
2. What is the sample?
3. Is this study descriptive or inferential?

Show/Hide Answer

1. All students studying at MacEwan.
2. Those 100 students randomly selected.
3. Inferential, the ratio 65/35 is based on the sample but used as an estimate of the population ratio.

Note: If you see the key words “**estimate for all**” in the question, it is usually an inferential study.

1.2 Data Collection

Most decision-making is based on facts and evidence, and data collection is an excellent means of obtaining evidence. For example,

- If you wanted to determine whether three cars provide sufficient capacity for the LRT train running between the Health Sciences and MacEwan stations, you could stand at the entrance of the MacEwan Station and count the number of people entering and exiting the station during peak hours on a regular school day. The data would help to determine whether more or less cars are needed.

Another example is in medical science.

- If you want to confirm whether a new drug is more effective than the old one, you would need to obtain a sample of patients with a similar condition and randomly assign them to two groups: subjects in one group receive the new drug, and subjects in the other group receive the old drug. After each drug has taken effect, you would compare the outcomes of the two groups.

Common ways to collect data are sampling surveys, observational studies, and designed experiments. For sampling methods, students in this course are only required to understand simple random sampling, which will be introduced in the next section. The main difference between an observational study and a designed experiment is that a designed experiment involves **manipulations** of the subjects, while an observational study does not.

1.2.1 Sampling Methods

In statistics, a sampling survey describes the process of selecting a sample of individuals/items from a target population in order to conduct a survey. A **census** is a type of survey in which the researcher samples the entire population. Typically, a census requires a population which is reasonably small, otherwise the process of data collection can be expensive, time-consuming and, in some cases, impossible. In fact, it is usually the case that the population is too large for the researcher to survey all of its members. For this reason, it is often the case that a small, carefully chosen sample is used to represent the population.

The logic behind sampling is this: by well-mixing the population, we are able to learn about the entire population by examining a sample. For example, suppose you are cooking a pot of soup and you would like to taste the soup; you do not need to have the whole pot—you just need to taste or sample a spoon of it. How can we “stir” people or items in sampling? We adopt the idea of randomization, which means we select individuals or items randomly.

Simple Random Sampling

Sampling methods are classified as either probability or nonprobability. In probability samples, each member of the population has a known, non-zero probability of being selected. In Stat 151, we only cover one particular sampling procedure, i.e., **simple random sampling**, in which each possible sample of a given size has equal chance of being selected. In simple random sampling, each individual is equally likely to be selected. For example, picking four cards randomly after shuffling the cards well, each card has the same chance of being picked. A sample obtained by simple random sampling is called a **simple random sample (SRS)**.

There are two types of simple random sampling. One is simple random sampling **with replacement**, whereby individuals of the population can be selected more than once; the other is simple random sampling **without replacement**, whereby any individuals of the population can be selected at most once.

There are several ways to obtain a simple random sample:

- Picking slips of paper out of a box
- Generating by computer
- Using a random numbers table

Obtaining a simple random sample by picking slips of paper out of a box is impractical, especially when the population is large. We can either use a computer or the random-number tables (see one below) to generate a simple random sample. The arrows are used to show our example that follows.

Table I: Random Number Table

Line Number	Column Number									
	00-09		10-19		20-29		30-39		40-49	
00	1 6 7 0 4	9 9 7 5 7	0 2 0 8 4	1 3 7 0 1	9 1 0 8 5	2 2 0 1 0	8 4 8 7 0	1 8 4 7 7	6 5 6 1 0	8 1 4 5 3
01	8 7 8 5 8	5 0 3 0 8	4 7 3 4 5	9 1 5 2 8	2 8 4 8 3	0 0 4 6 7	7 2 4 1 9	9 8 1 2 5	5 0 4 3 6	4 0 4 5 4
02	3 0 8 6 8	1 1 9 7 3	7 9 3 6 0	0 1 9 7 6	7 8 2 3 9	7 7 8 5 4	7 0 0 5 3	0 2 5 0 3	0 6 7 7 4	7 4 5 8 4
03	3 6 0 7 9	3 4 9 9 7	5 3 9 1 3	1 9 9 6 4	2 2 8 8 9	0 4 3 4 8	1 9 6 4 0	8 4 2 2 3	8 6 6 4 1	8 7 1 8 7
04	3 1 9 7 2	1 8 3 1 3	8 7 7 9 8	6 0 4 5 5	8 2 9 9 4	2 7 9 0 4	5 4 2 0 7	2 4 8 4 3	6 4 6 6 0	7 0 8 2 2
05	8 2 5 2 8	6 7 3 1 4	1 2 7 0 0	9 2 3 2 3	1 9 5 1 3	2 3 9 2 2	4 7 1 3 3	4 5 4 2 0	4 5 6 2 2	6 5 8 2 2
06	6 5 6 2 3	5 8 9 0 8	5 1 8 1 3	0 2 3 8 5	3 2 9 8 9	5 0 5 9 7	9 3 0 0 6	1 8 2 6 2	5 2 9 7 8	5 7 2 4 3
07	5 7 6 7 8	0 8 5 6 9	1 5 1 9 8	8 8 2 1 6	6 6 4 3 8	2 9 0 0 8	8 4 1 6 1	5 0 1 2 0	6 3 1 5 3	7 8 9 8 2
08	6 7 3 5 7	8 7 7 6 3	7 2 5 4 8	6 3 5 7 7	4 7 5 6 2	7 4 4 9 5	0 7 7 5 2	3 0 6 4 8	4 1 0 3 4	1 0 8 2 3
09	6 8 2 2 9	8 4 1 3 4	9 1 0 2 2	3 6 0 7 8	7 3 4 4 1	8 5 3 3 3	9 5 7 2 3	9 0 4 4 5	3 4 3 6 4	8 9 7 4 6
10	1 8 4 2 9	8 8 0 4 1	4 2 3 6 3	8 0 2 9 9	0 5 2 4 1	8 3 5 2 0	4 8 7 8 6	7 7 4 2 8	4 7 5 2 8	1 5 8 1 8
11	9 1 6 7 7	0 4 9 2 0	2 4 2 2 0	7 6 0 2 5	8 8 2 9 6	8 8 2 3 7	9 9 1 4 7	7 1 6 9 1	0 8 4 9 8	9 1 9 9 0
12	9 7 3 2 0	4 4 1 6 4	3 6 0 8 7	0 2 9 7 4	7 8 6 4 6	1 0 8 4 5	6 4 4 5 0	3 9 2 0 5	4 9 4 4 6	8 9 8 3 9
13	7 1 9 5 2	7 1 9 9 0	9 6 9 5 2	8 9 6 4 5	2 1 9 5 3	9 5 6 7 4	0 3 3 0 7	9 4 5 8 0	6 9 3 9 5	5 3 1 6 6
14	7 4 4 3 4	2 3 2 5 1	6 9 3 8 7	1 1 4 3 5	8 5 2 9 7	2 9 3 6 0	9 1 1 6 7	3 1 9 9 9	9 1 9 5 2	9 9 9 7 6
15	7 5 6 2 5	5 4 0 7 2	1 5 5 9 6	1 9 7 6 0	8 2 2 0 5	3 8 6 0 2	8 1 5 7 1	4 2 9 0 5	7 3 0 7 2	2 1 4 9 8
16	7 8 0 6 8	6 1 7 9 9	0 4 1 4 9	6 0 1 8 2	2 9 8 8 6	9 8 3 3 1	1 6 5 2 2	0 2 8 7 7	8 8 4 3 1	8 9 7 8 0
17	1 6 9 0 4	4 5 9 9 8	3 2 4 7 6	6 6 1 9 3	6 1 1 8 8	5 8 1 7 7	1 3 3 7 7	9 3 9 5 4	9 2 1 4 0	9 7 7 1 3
18	0 2 4 2 5	0 0 5 4 8	5 9 4 9 0	2 7 8 6 8	3 6 7 0 0	4 1 3 9 0	5 7 1 5 3	7 7 3 6 1	2 4 6 2 8	5 7 5 3 0
19	7 4 8 1 3	1 9 2 1 5	4 4 6 0 5	2 1 4 6 7	6 2 7 7 6	2 2 8 9 2	0 0 3 9 4	4 2 2 4 9	9 6 6 9 7	0 2 2 4 1

Table 1.1: Random Number Table. [\[Image Description \(See Appendix D Table 1.1\)\]](#)

Example: Simple Random Sampling Using the Random Number Table

Let us use the table above to show how to obtain a simple random sample without replacement.

Suppose our section consists of 60 students, from which we would like to obtain a random sample of size 10. First, we can number the students from 1 to 60. To select 10 random numbers between 1 and 60, we first pick a random starting point. We can close our eyes and randomly point our finger into the table and use the first two digits of the number we point on as the starting point. For example, if the number you pick is 82 (Line number 05 and column number 00-01), then go down the column, the second number is 65 which is greater than 60 and hence is discarded, the third one is 57 which is good to keep, the next one is 67 which is greater than 60 and hence is discarded, the next ones are 68 (discarded), 18 (good to keep), 91 (discarded), 97 (discarded), 71 (discarded), 74 (discarded), 75 (discarded), 78 (discarded), 16 (good to keep), 02 (good to keep), 74 (discarded). Now we are at the end of columns 00-01, move to columns 02-03 at the bottom then move up, the numbers are 81 (discarded), 42 (keep), 90 (discarded), 06 (keep), 62 (discarded), 43 (keep), 95 (discarded), 32 (keep), 67 (discarded), 42 (already in the list, discarded), 22 (keep), 35 (keep). As a result, the 10 random numbers are: 57, 18, 16, 02, 42, 06, 43, 32, 22, 35.

In this course, a simple random sample means a simple random sample without replacement by default.

In practice, the most popular way to obtain a simple random sample is using computer.

Example: Simple Random Sampling Using R Commander

Suppose our section consists of 60 students, use R Commander to generate a simple random sample of size $n = 10$ without replacement. In order to obtain the same sample, use “4061” as the random seed.

Similar to picking a random starting point, we first need to set a random seed using the function “**set.seed()**”. A simple random sample can be obtained using R Commander in two steps:

1. Type “**set.seed(4061)**” in the R Script window, then press “**Submit**”. By doing this, we will obtain the same sample if we rerun the command line.
2. Type “**sample(1:60, 10)**” in the R Script window, and then press “**Submit**”. The function “**sample()**” is used to generate a simple random sample. The first input argument indicates from what the sample is taken from. In this example, it is “1:60” which means 1, 2, 3, \dots , 59, 60. The second input argument specifies the sample size. The command line “**sample(1:60, 10)**” means we could like to randomly take 10 different numbers between 1 and 60.

The resulting sample is 53, 14, 57, 13, 8, 45, 11, 50, 59, 25 (see the snapshot below).



Snapshot 1.1: Generate a simple random sample of size 10 using R commander. [\[Image Description \(See Appendix D Snapshot 1.1\)\]](#)



Activity

Exercise: Generate Simple Random Sample Using R Commander

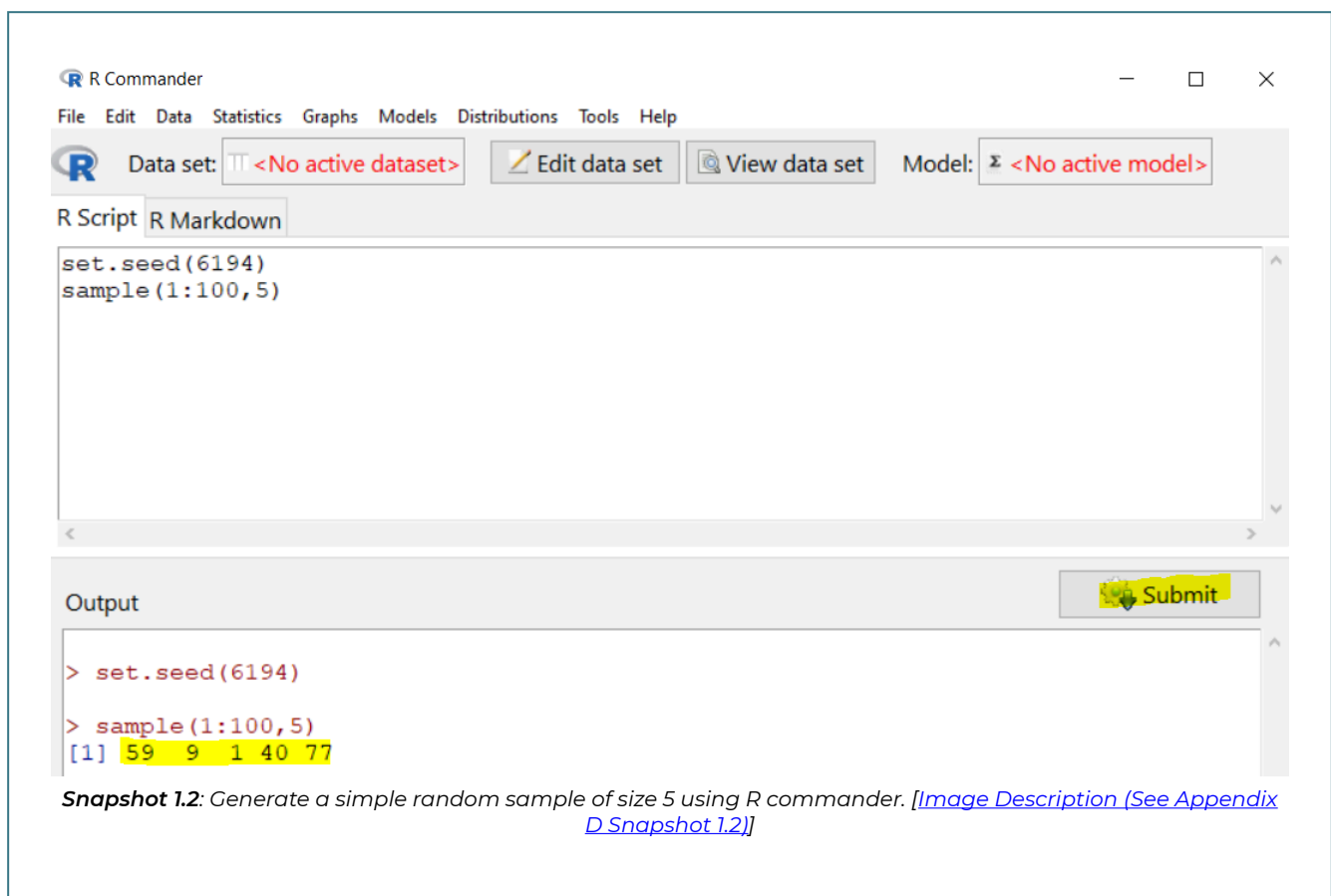
Suppose that a class consists of 100 students, use R Commander to generate a simple random sample of size $n = 5$ without replacement. Use "6194" as the random seed.

Use R Commander to generate a simple random sample of size $n = 5$ without replacement. Use "6194" as the random seed.

Show/Hide Answer

1. Type "**set.seed(6194)**" in the R Script window, then press "**Submit**".
2. Type "**sample(1:100, 5)**" in the R Script window, and then press "**Submit**".

The resulting sample is 59, 9, 1, 40, 77.



Some Other Sampling Methods (not required, extra reading)

Besides the simple random sampling, some other good sampling methods include:

- **Stratified sampling:** Population is divided into homogeneous groups called strata and then simple random sampling is applied within each stratum. For example, 65% of students at MacEwan are female students. A stratified sample of 100 students can be obtained by drawing a simple random sample of 65 female students and a simple random sample of 35 male students.
- **Cluster sampling:** Split the population into clusters and select one or several clusters at random. And then conduct a census within each cluster. Each cluster should represent the full population fairly. For example, Edmonton is geographically divided into 375 neighborhoods. A cluster sample residents of Edmonton can be obtained by taking a simple random sample of 20 neighborhoods and then taking all residents in those 20 selected neighborhoods.
- **Systematic sampling:** Select every k th individual from the sampling frame, e.g.,

choose every 5th person on an alphabetical list of students. If we start from the 4th individual, and we choose every 5th person, the resulting list will be $4, 4 + 5 = 9, 9 + 5 = 14, 14 + 5 = 19$, and etc. Therefore the 4th, 9th, 14th, 19th, and so on on the list will be in the sample.

Some convenient but relatively not that desirable sample methods are:

- **Voluntary response sampling:** A large group of individuals is invited to respond and all who do respond are counted, e.g., online survey.
- **Convenience sampling:** This includes individuals who are convenient to sample. For example, stop people in a mall and ask questions.



Exercise: Sampling Methods

If I want to know the percentage of residents in Edmonton who have taken at least one statistics course like STAT 151, identify and explain the advantages and disadvantages of the following sampling methods.

1. Stop 100 people at the entrance of MacEwan and ask for their response.
2. Get the phone numbers of all residents in Edmonton from the census data, randomly pick 100 people, call them and ask for their response.
3. Send invitations to fill out an online survey.

Show/Hide Answer

1. This is a convenience sample. Advantage is convenient to take the sample. Disadvantage is the estimate is biased and might overestimate the percentage.
2. This is a simple random sample. However, we will miss those whose phone numbers are not listed and those who do not answer the phone.
3. This is a voluntary sample. It is cost effective to send invitations. However, the response rate might be low and we will miss those who have no access to computer.

Note: In practice, there might be no sampling method that will correspond exactly to simple random sampling; but some will be better than others.

1.2.2 Designed Experiments

Design experiments are another method of data collection. In a designed experiment, investigators randomly assign subjects to different experimental groups (called treatments), observe the outcomes, and test whether treatment differences are statistically significant. A designed experiment is ideal for investigating a cause-effect relationship.

Example: A Designed Experiment

Researchers in a pharmaceutical company want to test whether their new pain killer is effective or not in reducing pain. Forty females with similar conditions (e.g., age, diet) are randomly assigned to two groups: 20 take a vitamin and another 20 take the pain killer. The subjects are then asked to report their pain score in a scale of 0 to 10, four hours after taking the pill. Here is how the experiment is designed:

- The 40 female participants should be in similar conditions to minimize the effect of other factors, such as age, diet, and etc.
- It is also important to ensure that the vitamin and the pain killer look similar, so that participants do not know which group they are in. Here, the vitamin is called the **placebo**. It is well-known that people tend to feel better after they receive some kind of treatment, even if the treatment does not have any physical effect; this is called the **placebo effect**. By having both a treatment group and a placebo group, the researchers are able to minimize the placebo effect and hence more accurately measure the effectiveness of the pain killer.
- Randomly assign the individuals to the two different groups by collecting a simple random sample of 20 out of the 40 individuals, assigning them to the vitamin (placebo) group, and assigning the remaining 20 individuals to the group receiving the pain killer.

1.2.3 Observational Studies

In some studies, it may not be possible to randomly assign the subjects to different treatment groups due to ethical or practical reasons. For example, we cannot randomly assign people into smoker and non-smoker groups. In those cases, we might have to conduct observational studies.

In an observational study, the investigator observes the characteristics of individuals in samples from a population of interest to discover trends and possible relationships between variables.



Activity

Exercise: An Observational Study

In order to study the association between breast cancer and smoking, can we randomly assign the participants to the smoker or non-smoker group? Which of the following two studies do you think better?

1. Follow 20 smokers and 20 non-smokers who have similar conditions, compare the occurrences of breast cancer in the two groups at the end of study.
2. From one hospital, sample 20 breast cancer patients and another 20 patients without breast cancer by matching, i.e., we try to make the two groups as similar as possible except for the cancer status. Determine the smoking status for each subject and compare the percentages of smokers in both groups. This is called a case-control study. A breast cancer patient is a case, while a patient without breast cancer is a control.

Show/Hide Answer

The second study plan is better. Not everyone will develop breast cancer at the end of the first study; we might not observe any cases of cancer in either groups. The first study plan could potentially end up with no useful results. In the second study, however, we are always able to compare the percentage of smokers between groups and therefore, establish the association between smoking and breast cancer.



Instructor's Note

1. For rare disease, it is better to first recruit participants with the condition (called cases) and then recruit participants without the condition (controls) by matching other characteristics.
2. It is easier to establish a causal relationship with experimental studies than observational studies. Whenever possible, experimental studies are preferred.

1.3 Variables and Data

Data are more than just numbers; they are information about a group of individuals organized in variables. The values of the variables are called **data**. For example, I have the following information: Kate, car, 1, John, bicycle, 2, Mary, public transportation, 5, Adam, walk, 7. It is very hard to interpret this information without context. If I arrange these information in variables, it is easy to understand the data.

Table 1.2: Table of Variables

Name	Transportation	Number of hours/day on Internet
Kate	car	1
John	bicycle	2
Mary	public transportation	5
Adam	walk	7

Given the three variables, i.e., Name, Transportation, and Number of hours/day on Internet, the data tell us that Kate comes to school by car, she spends 1 hour on average per day on surfing internet, while John comes to school by bike and he spends 2 hours a day on average surfing the Internet.

Variable is a characteristic that varies from one individual to another. As shown in the above example, “Name”, “Transportation”, and “Number of hours/day on Internet” are the variables. There are two types of variables: qualitative/categorical and quantitative variable. The quantitative variable can be further classified as either continuous or discrete.

- **Qualitative variable:** A non-numerically valued variable that classifies subjects into different categories, such as “Name” and “Sex”. The values of these two variables are not numbers, so they are also called **categorical variables**. Categorical variable can be further classified as nominal and ordinal.
 - **Nominal variable:** non-numerical variable that cannot be ordered. For example, “Name” and “Sex” (male or female).
 - **Ordinal variable:** non-numerical variable than can be sorted. For example, “how often do you drink” (never, seldom, sometimes, often, everyday). Here, values can be sorted by frequency. Another example is “Size” (small, medium, large).
- **Quantitative variable:** A numerically valued variable, e.g., “Number of hours/day on Internet” is an example of a quantitative variable. There are two types of quantitative

variable—continuous and discrete.

- **Continuous variable:** A quantitative variable whose possible values form some interval of numbers, e.g., height, length, salary, age. Technically speaking, continuous variables have arbitrary number of decimal places.
- **Discrete variable:** A quantitative variable whose possible values can be listed, e.g., number of siblings, number of phone calls within an hour.



Activity

Exercise: Data Types

Classify the following variables as qualitative/categorical nominal, qualitative/categorical ordinal, quantitative continuous, or quantitative discrete.

1. Gender
2. Height
3. Heart rate (beats/minute)
4. Number of siblings
5. Weight
6. Shoe size
7. Length of your left foot
8. Salary
9. Percentage of female students at MacEwan
10. Clothing size

Show/Hide Answer

1. Either female, male, or other sex orientation; it is non-numerical, thus categorical variable. Values cannot be ordered, it is qualitative/categorical nominal.
2. Measurement in length, quantitative continuous variable
3. Beats can be listed, it is quantitative discrete variable
4. Non-negative integers, quantitative discrete variable
5. Quantitative continuous variable
6. All possible values can be listed, quantitative discrete variable
7. Quantitative continuous variable
8. In any currency, technically speaking, all possible values can be listed, should be quantitative discrete. For the sake of convenience, however, it is treated as quantitative continuous in practice.
9. Percentage = $\frac{\text{number of females}}{\text{total number of students}} \times 100\%$, all possible values of both of numerator and denominator can be listed, technically speaking, should be quantitative discrete. For the sake of convenience, however, it is treated as quantitative continuous in practice.
10. The values are XS, S, M, L, XL which can be ordered; therefore, it is qualitative/categorical ordinal.

1.4 Organizing Data

Next, we focus on presenting and summarizing data using different tables and figures.

Given a set of data, how can you present the data? It is essential to plot the data before conducting any data analysis. The definition of descriptive statistics tells us we can use tables and graphs to present the data. Different tables and graphs are used to describe the two different types of data—qualitative and quantitative. Let us start with qualitative variables, then continuous and then discrete variables.

1.4.1 Organizing Qualitative Data

Numerically, we can use frequency or relative frequency tables to summarize qualitative/categorical data. Graphically, we can use pie charts or bar charts.

The distribution of a qualitative variable is given in a **frequency (relative frequency) table**. For example, the students were asked, “How you came to school today?” Fifty-three students answered by car, 136 by public transportation, nine by bicycle, 74 by walking, and one by other means. The results are summarized in the following table.

Table 1.3: Frequency and Relative Frequency Table of “Transportation”

Transportation	Frequency	Relative Frequency	Percentage
Car	53	$53/273 = 0.1941$	19.41%
Public	136	$136/273 = 0.4982$	49.82%
Bicycle	9	$9/273 = 0.0330$	3.30%
Walking	74	$74/273 = 0.2710$	27.10%
Other	1	$1/273 = 0.0037$	0.36%
Total	273	1.000	100%

- The first column gives all possible outcomes of the variable, which are called categories.
- The second column gives the number of observations falling into each category; we call this number the **frequency** of that category. For example, the frequency of taking public transit is 136.

- The third column gives the **relative frequency**, which is calculated as:

$$\text{relative frequency} = \frac{\text{frequency}}{\text{total}}.$$

For example, the relative frequency of taking public transit is 0.4982, which means 49.82% of the students came to school by public transit. Note that the relative frequencies always add up to 1 across all the categories.

- The fourth column gives the percentage, which is calculated as
percentage=relative frequency \times 100.

Based on the (relative) frequency table, we can draw a bar chart or a pie chart to summarize the data.

- A **bar chart** represents each category with a bar whose height equals each category's relative frequency or frequency. The bars are plotted next to each other without touching each other. One bar for one category.
- A **pie chart** is a disc divided into wedge-shaped pieces whose areas are proportional to the relative frequencies. One slice for one category, the angle of each slice = relative frequency \times 360°.

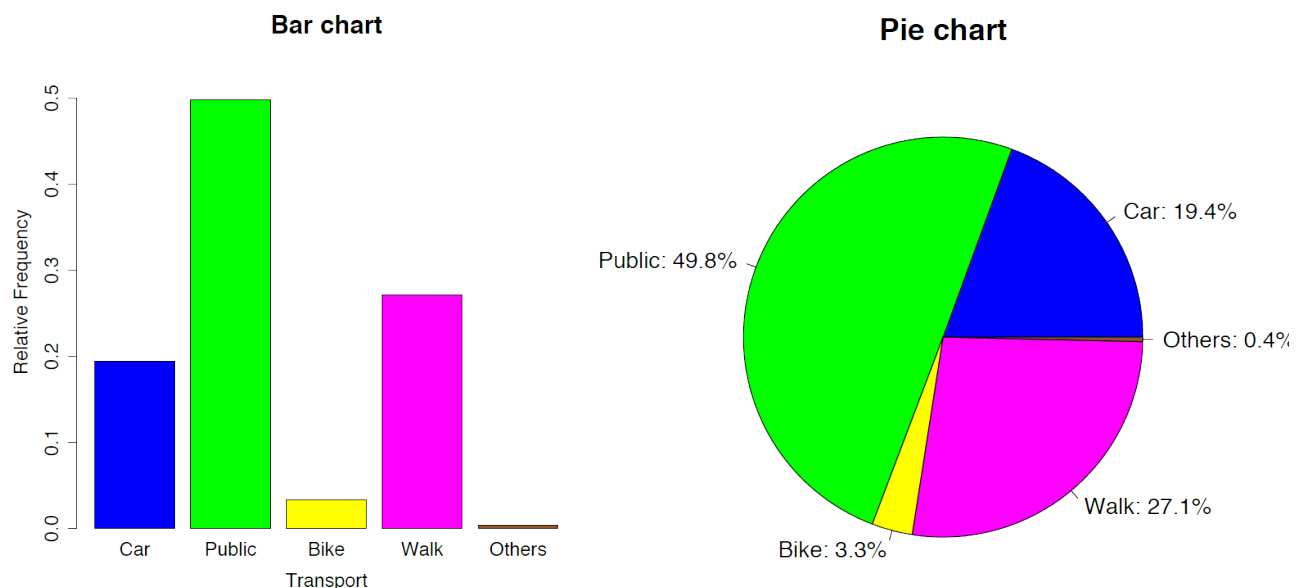


Figure 1.2: Bar Chart (left panel) and Pie Chart (right panel) of “Transportation” [[Image Description](#) (See Appendix D Figure 1.2)]

If the bar chart and pie chart are generated based on counts, the charts won't change except for the scale—the relative frequency in the bar chart and the percentage in the pie chart will be replaced by counts or frequency.

Suppose that another qualitative variable recorded in the study was “gender.”, we can also present the data characterized by two qualitative variables in what is referred to as a **contingency table**. Below is the contingency table with the two qualitative variables, gender and transport:

Table 1.4: Contingency Table of “Gender” and “Transportation”

	Car	Public	Bicycle	Walking	Others	Total
Female	25	80	5	38	0	148
Male	28	56	4	36	1	125
Total	53	136	9	74	1	273

The variable “**Gender**” is called **the row variable** (shown in red font in the table) and “**Transportation**” is the **column variable** (shown in blue font in the table). The totals “148” and “125” (in red) are called **row totals** (sum across transportation for each category of “gender”). The totals “53,” “136,” “9,” “74,” “1” are called **column totals** (sum across gender for each category of “Transportation”), and “273” is the **grand total**. Those 10 numbers in bold are called **cells**.

One interesting question is whether the pattern of transportation among females is the same as that among males. We can compare the relative frequencies of all the categories for females with their counterparts among males. There are 148 female students and 125 male students in total; therefore, the relative frequencies of the five categories for females are $25/148$, $80/148$, $5/148$, $38/148$, $0/148$ as compared to $28/125$, $56/125$, $4/125$, $36/125$, $1/125$ for males.

The distributions of “transportation” for females and males can be compared graphically using a side-by-side pie chart and a side-by-side bar chart.

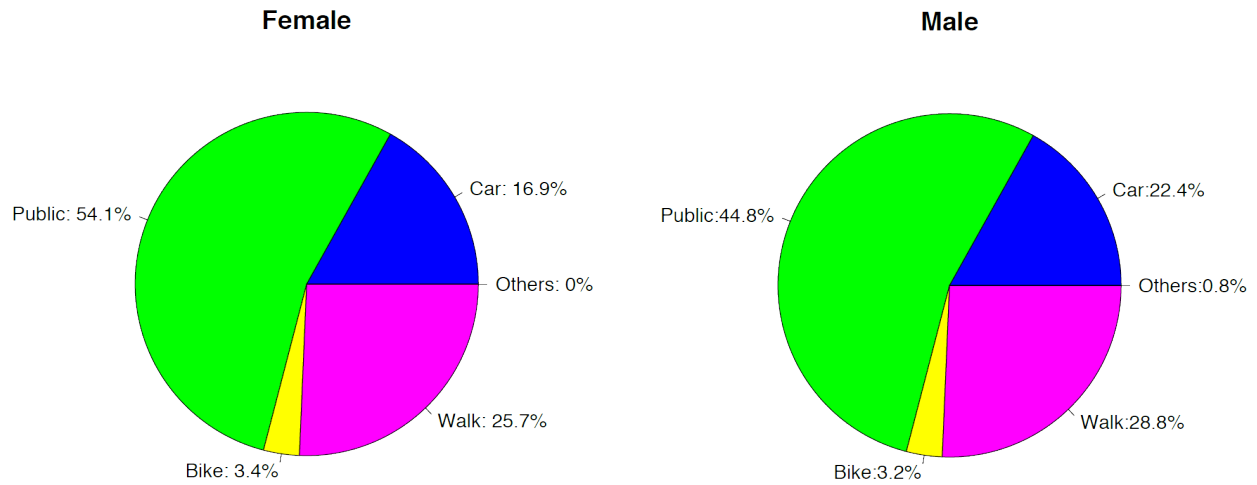


Figure 1.3: Side-by-side pie Chart of “Transportation” for Females and Males. [[Image Description \(See Appendix D Figure 1.3\)](#)]

The side-by-side pie chart shows that the patterns in transportation among females and male are very similar, since the two pie charts are almost identical. The side-by-side bar chart based on the **relative frequency** gives the same conclusion: the distributions of “transportation” for female and male are very similar, which implies there is no significant difference between female and male in the way they come to school.

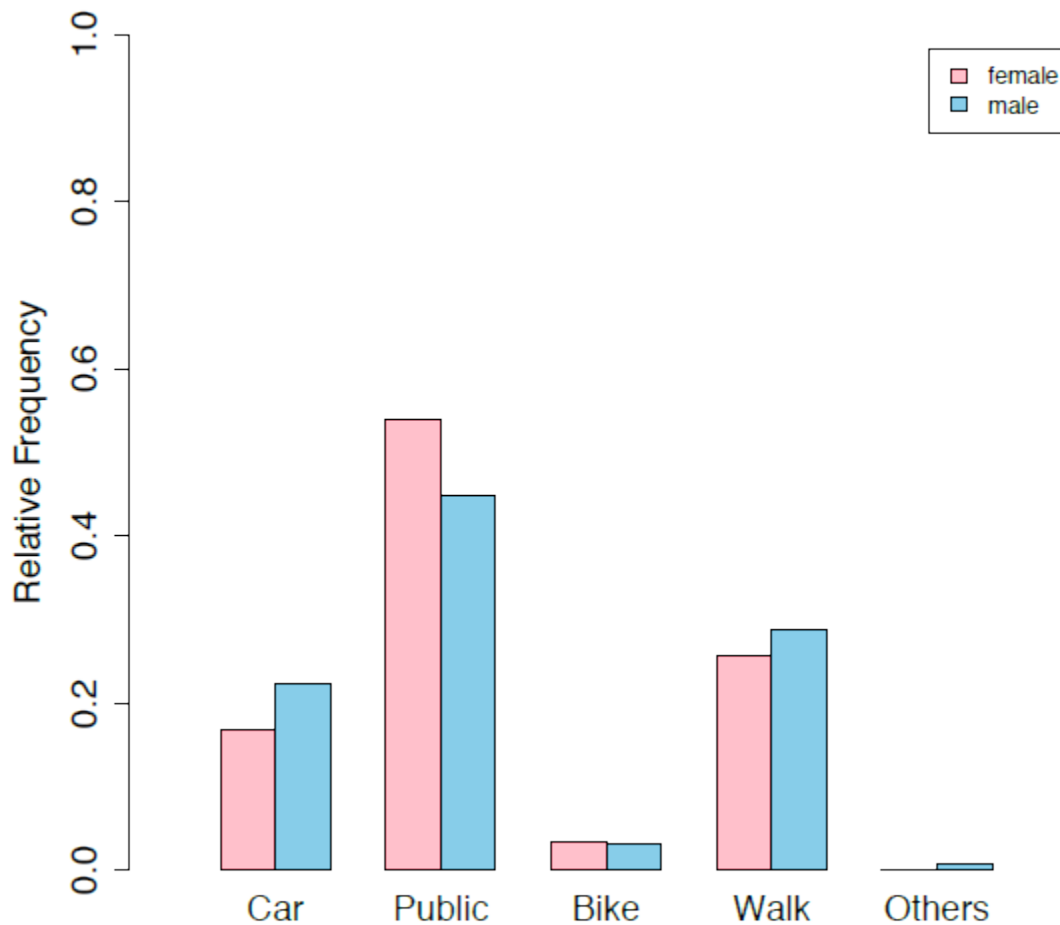


Figure 1.4: Side-by-Side Bar Chart of “Transportation” for Female and Male. [[Image Description \(See Appendix D Figure 1.4\)\]](#)



Instructor's Note

When we compare the distributions of two different groups using a side-by-side bar chart, we should use the relative frequency as the y-axis. Using frequency as the y-axis and comparing the frequencies alone, without taking into account the total of each group, can be misleading.

1.4.2 Organizing Quantitative Discrete Data

We are able to list all possible values for a quantitative discrete variable; therefore, for a quantitative discrete variable with only a few different values, we can describe it using tools similar to those for qualitative variables, i.e., a (relative) frequency table and histogram.

A histogram is somewhat similar to a bar chart. The x-axis shows the value of the variable of interest and the y-axis displays either frequencies or relative frequencies. Histograms can be used to describe both quantitative discrete and quantitative continuous variables. For a continuous variable, we cut the range of the variable into subintervals of equal width and draw one rectangle for each subinterval; the height of the rectangle is the number of observations falling into the corresponding subinterval. For a discrete variable with a small number of possible values, we can draw a rectangle with equal width for each value, the height of each rectangle is either the frequency or relative frequency.

Example: Organizing Quantitative Discrete Variables

There are 100 students in a class; ten have no siblings, thirty have one sibling, thirty-five have two siblings, fifteen have three siblings, and ten have more than three siblings.

We can use a (relative) frequency table and a histogram to summarize the data.

Table 1.5: Frequency and Relative Frequency Table of “# of Siblings”

# of Siblings	Frequency	Relative Frequency
0	10	0.10
1	30	0.30
2	35	0.35
3	15	0.15
>3	10	0.10
Total	100	1.00

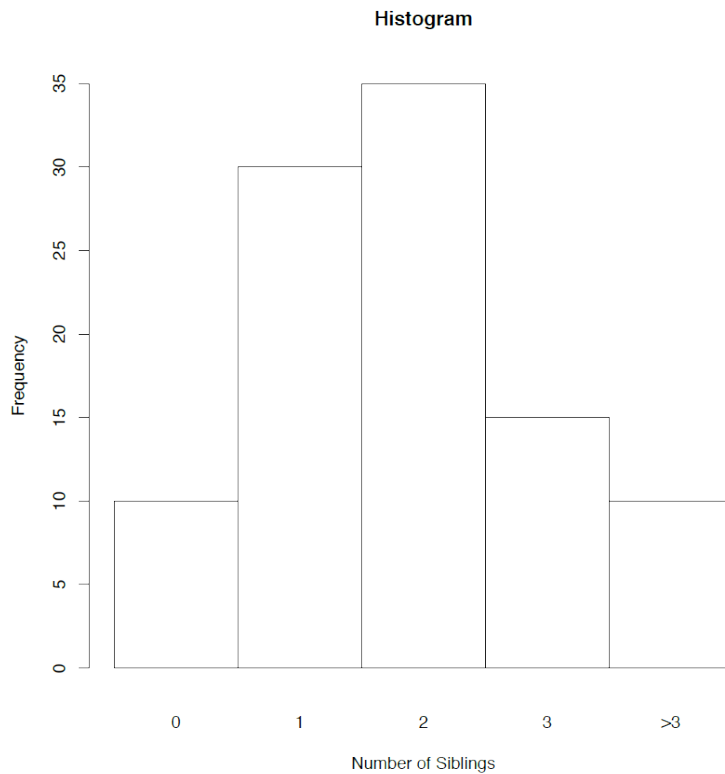


Figure 1.5: Histogram of “# of Siblings” [[Image Description \(See Appendix D Figure 1.5\)](#)] [Click on image to enlarge.](#)



Instructor's Note

Difference between a bar chart and a histogram:

- The bars of a bar chart do not touch one another. Since there is often no inherent ordering among the categories, the order among the bars is usually irrelevant (i.e., bars can be switched without affecting the usefulness of the graph).
- The adjacent bars of a histogram do touch one another. Since there is ordering among numbers, that ordering is to be preserved among the bars of a histogram. That is, the first bar corresponds to the smallest value (or the interval of the smallest values), the second bar corresponds to the second smallest value (or the interval of the second smallest values), and so on.

1.4.3 Organizing Quantitative Continuous Data

Example: Organizing Quantitative Continuous Variables

Here are the 50 grades for an exam:

68	72	59	56	60	40	55	68	76	75
46	59	37	54	83	85	29	55	56	42
50	49	65	68	61	53	55	92	68	48
79	51	24	57	48	71	90	81	34	60
47	39	65	74	49	52	59	9	62	37

How to present and summarize these data?

Grouping Table and Histogram

Recall that all values of a discrete variable can be listed. However, this is not the case for a continuous variable: we cannot list all possible values for a continuous variable. For example, even though the above 50 grades are all reported as whole numbers, there is no reason why a grade couldn't contain a decimal, such as 46.5, or 66. $\bar{6}$. For this reason, it is most appropriate to view the grade variable as a continuous variable. Even though we cannot list all possible values of a continuous variable, we can cut the range of a continuous variable into subintervals of equal width and use histograms to summarize quantitative continuous data. The range of grade is $[0, 100]$, a convenient and neat cut is by intervals with width of 10 or 20. If we cut by intervals of 10, the resulting grouping data and histogram are as follows:

Table 1.6: Grouping Table of Grade for Histogram

Interval	Frequency	Relative Frequency
[0, 10)	1	$1/50=0.02$
[10, 20)	0	$0/50=0.00$
[20, 30)	2	$2/50=0.04$
[30, 40)	4	$4/50=0.08$
[40, 50)	8	$8/50=0.16$
[50, 60)	14	$14/50=0.28$
[60, 70)	10	$10/50=0.20$
[70, 80)	6	$6/50=0.12$
[80, 90)	3	$3/50=0.06$
[90, 100]	2	$2/50=0.04$
Total	50	1.00

Please note that **we still need to keep those intervals that have no observations**. For example, the interval [10, 20) includes 10 but excludes 20, and has no observations. We need to keep this interval when we draw a histogram for the data.

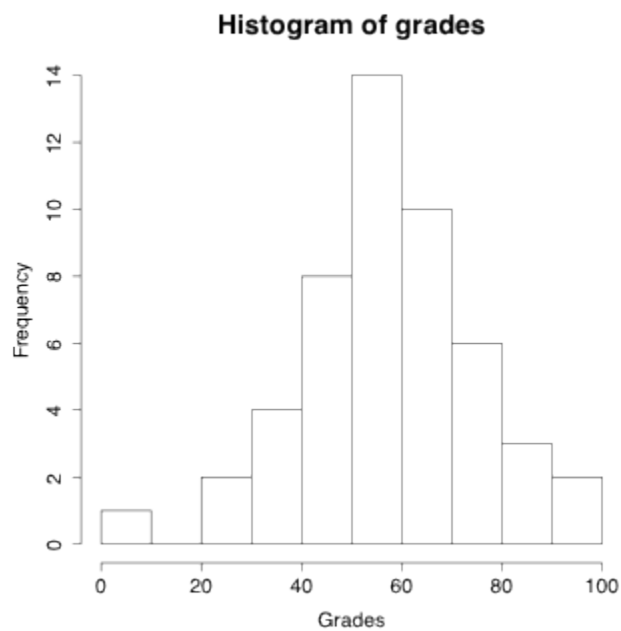


Figure 1.6: Histogram of Grade [[Image Description](#)
([See Appendix D Figure 1.6](#))]



Instructor's Note

1. A common question when drawing histograms is whether to use $[,)$ or $(,]$ intervals. Please note that different software may follow different rules. It is important to consistently follow the same rule for all intervals in your histogram.
2. Another common question is how many bins is proper. A rule of thumb is the square root of the number of observations. For the grade example, since $n = 50$ and $\sqrt{n} = \sqrt{50} = 7.07$. The range of grade is $[0, 100]$, to create convenient cuts, we can divide the range either into 10 subintervals with equal length, i.e., $[0, 10), [10, 20), \dots, [90, 100]$ or 5 subintervals with equal width, i.e., $[0, 20), [20, 40), \dots, [80, 100]$.
3. Note that histograms with different number of bins might appear very different. When investigating the shape of the distribution of a variable using a histogram, it is always better to draw a boxplot and normal Q-Q plot as well. Boxplot and normal Q-Q plot will be covered in sections 2.4 and 5.6 respectively.

Stem-and-Leaf Diagram

Another way to present quantitative data is a stem-and-leaf diagram. To construct a stem-and-leaf diagram:

- Think of each observation consisting of a stem (all but the rightmost digit) and a leaf (the rightmost digit, a single digit).
- Draw a vertical line, write the stems from the smallest to the largest in a vertical column to the left of the vertical line.
- Write each leaf to the right of the vertical line in the same row as its corresponding stem.
- Arrange the leaves in each row from the smallest to the largest.
- Indicate the decimal place of the data if applicable.

Let's return to the grades data:

68	72	59	56	60	40	55	68	76	75
46	59	37	54	83	85	29	55	56	42
50	49	65	68	61	53	55	92	68	48
79	51	24	57	48	71	90	81	34	60
47	39	65	74	49	52	59	9	62	37

We can group the grades by the first digits (in intervals of 10) as follows,

Table 1.7: Working Table for Stem-and-Leaf Diagram

Interval	Data
[0, 10)	9
[10, 20)	
[20, 30)	24, 29
[30, 40)	34, 37, 37, 39
[40, 50)	40, 42, 46, 47, 48, 48, 49, 49,
[50, 60)	50, 51, 52, 53, 54, 55, 55, 55, 56, 56, 57, 59, 59, 59
[60, 70)	60, 60, 61, 62, 65, 65, 68, 68, 68, 68, 68
[70, 80)	71, 72, 74, 75, 76, 79
[80, 90)	81, 83, 85
[90, 100]	90, 92

If we take apart the grades and mark the first digit at left side of the line and the second digit at the right side of the line, it becomes a stem-leaf diagram:

Stems	Leaves
0	9
1	
2	49
3	4779
4	02678899
5	01234555667999
6	0012558888
7	124569
8	135
9	02

Decimal place: 9|0 = 90

Figure 1.7: Stem-and-Leaf Diagram of Grade [\[Image Description \(See Appendix D Figure 1.7\)\]](#)

The part “Decimal place: 9|0 = 90” indicates that the decimal point is one digit to the right of the vertical line.

Some other useful guidelines of the stem-and-leaf diagram are as follows:

- Keep the stems within the range of the data even though they have no leaf.
- If there are too many leaves, break down each stem into two lines. Leaves from 0 to 4 are placed in the first line and 5 to 9 in the second. (See in plot on the right for the grade example. Because the interval [50, 60) has too many leaves, we break the leaves into 2 lines: the first line lists those leaves ranging from 0 to 4, the second line lists leaves ranging from 5 to 9. Do the same to all the other stems.
- Divide or multiply the numbers by 10, 100, etc., and then round if necessary to create integers that have at most three digits, and indicate the decimal point if applicable.

Stems	Leaves
0	
0	9
1	
1	
2	4
2	9
3	4
3	779
4	02
4	678899
5	01234
5	555667999
6	0012
6	558888
7	124
7	569
8	13
8	5
9	02
9	

Example

Let's consider two data sets. Data set I: 3600, 1500, 6900 and Data set II: 0.36, 0.15, 0.69. It is not a good idea to draw the stem-and-leaf diagram based on the original data sets. Take Data set I for example, all three numbers have a leaf of 0 (the right most digit) and there are many stems without leaves between 150 and 360. Therefore, we divide the numbers by 100 and transform the numbers to 36, 15, 69 to draw a stem-and-leaf diagram. Finally, we indicate the decimal point by putting 6|9=6900 at the bottom of the graph. Similarly, we multiply all three numbers 0.36, 0.15, 0.69 by 100 to create a new data set: 36, 15, 69 and then draw a stem-and-leaf diagram.

These two data sets have the same resulting stem-and-leaf diagram as the data set 36, 15, and 69. However, the decimal point is 3 digits to the right of the vertical line for Date set I, i.e., we should indicate 6|9=6900; the decimal point is one digit to the left of the vertical line for Date set II, i.e., 6|9=0.69.

Stem-and-Leaf Diagram for Data set 1: 3600, 1500, 6900 Stem-and-Leaf Diagram for Data set 2: 0.36, 0.15, 0.69

Stem	Leaf
1	5
2	
3	6
4	
6	9

Decimal 6|9=6900

Stem	Leaf
1	5
2	
3	6
4	
6	9

Decimal 6|9=0.69

1.5 Shape of a Distribution

A histogram shows the shape of the distribution of a quantitative variable. The shape of a distribution includes the following three aspects:

- Overall shape: what the distribution looks like, e.g., bell-shaped, J-shaped, triangular, and uniform. (See examples in the figures below.)
- Modality: number of peaks (highest points). A distribution is called **unimodal** if it has only one peak, **bimodal** if it has two peaks, and multimodal if it has more than two peaks.
- Symmetry and skewness. If you fold a distribution in the middle and the two parts can match, the distribution is called **symmetric**. If it has a longer left tail, it is called skewed to the left (or left skewed), and skewed to the right (or right skewed) if it has a longer right tail.

The following figure shows some special shapes of distributions.

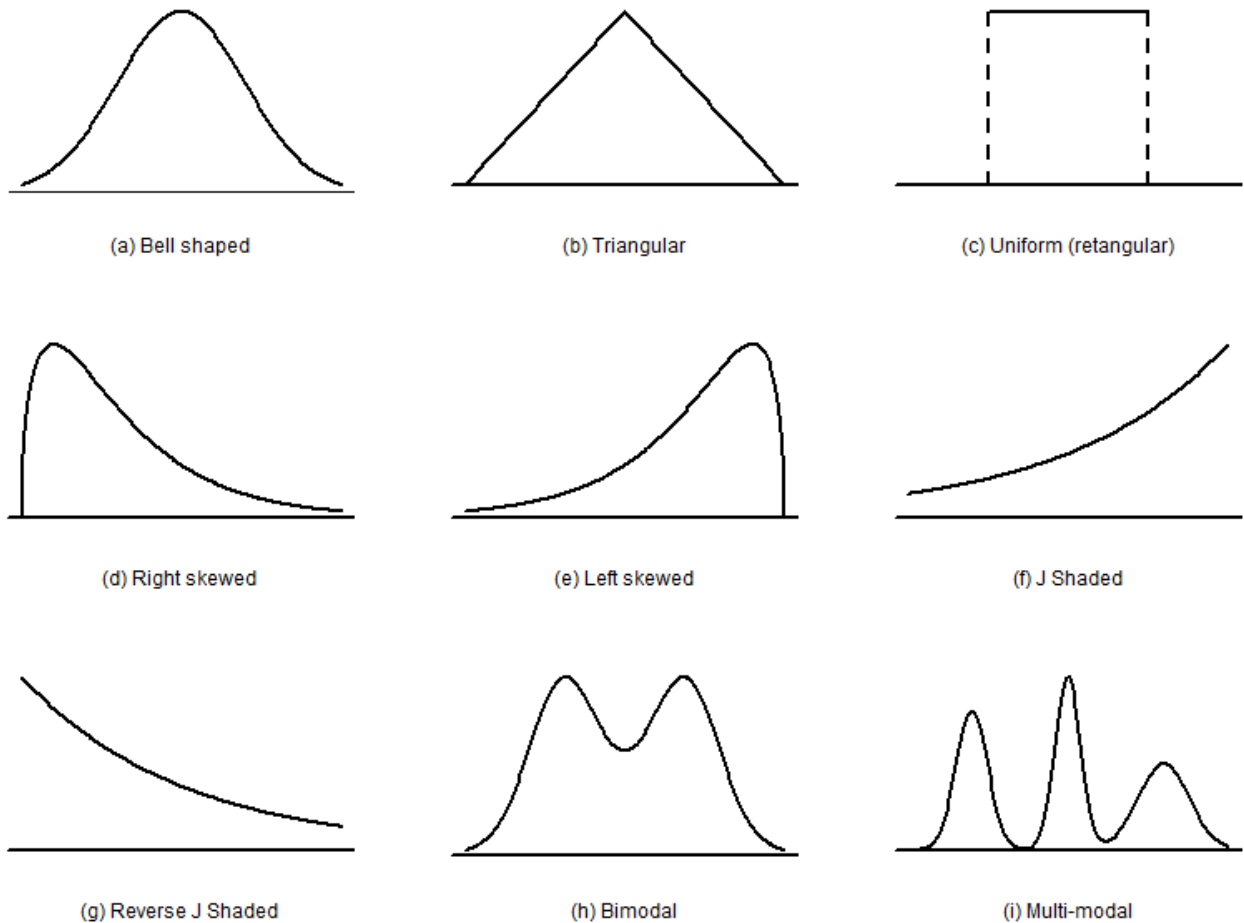


Figure 1.8: Some Special Shapes of Distributions [[Image Description \(See Appendix D Figure 1.8\)](#)]

Distributions in Figure 1.8 can be described respectively as follows:

- (a) bell-shaped, unimodal, and symmetric
- (b) triangular, unimodal, and symmetric
- (c) rectangular, no peak, and symmetric
- (d) unimodal and right-skewed
- (e) unimodal and left-skewed
- (f) J-shaped, unimodal, and left skewed
- (g) reversed J-shaped, unimodal, and right-skewed
- (h) bimodal and symmetric
- (i) multimodal and asymmetric

Comment on the overall shape and modality of the following histogram.

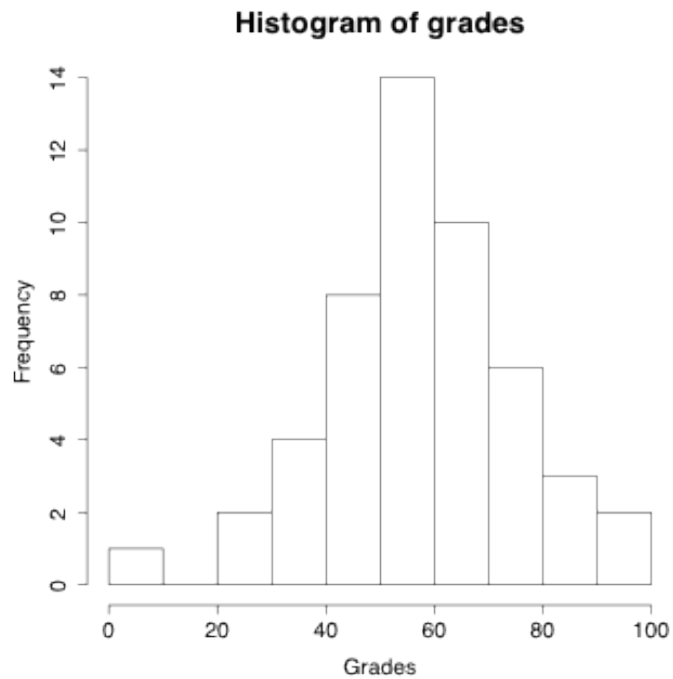


Figure 1.8.1: The histogram of the grades data above shows its distribution is bell-shaped, unimodal, and symmetric. [[Image Description \(See Appendix D Figure 1.8.1\)](#)]



Activity

Exercise: Shape of a Distribution

Figure 1.9 is the histogram of survival time in years after cancer diagnosis. Comment on the overall shape and modality of the histogram.

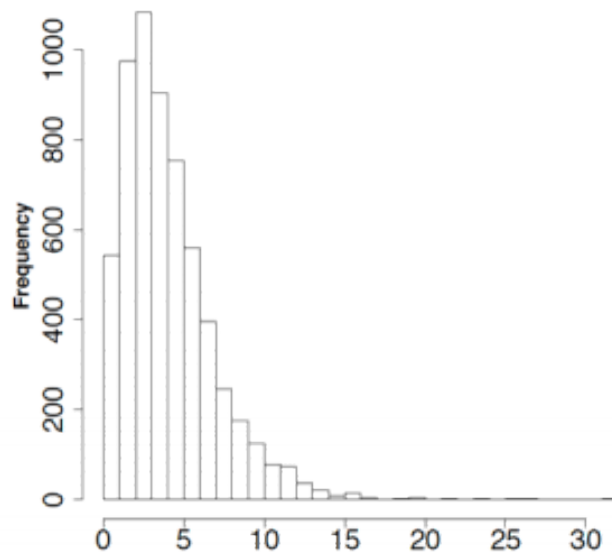


Figure 1.9: Histogram of Survival Time (in years) [[Image Description \(See Appendix D Figure 1.9\)](#)]

Show/Hide Answer

We can see the distribution of the survival time is unimodal and right-skewed.

1.6 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Tell what statistics is and how useful it is to solve real-life and work issues (Section 1.1).
- Identify the population and sample of a study (Section 1.1).
- Identify whether a study is descriptive or inferential (Section 1.1).
- Draw a simple random sample using the random-number table (Section 1.2).
- Distinguish designed experiments and observational studies (Section 1.2).
- Distinguish two types of data: quantitative and qualitative/categorical (Section 1.3).
- Describe qualitative/categorical data using pie charts and bar charts (Section 1.4).
- Describe quantitative data using stem-and-leaf diagrams and histograms (Section 1.4).
- Describe the shape of a histogram (Section 1.5).

1.7 Review Questions

1. In order to study the effect of aspirin on preventing heart disease and stroke for women over 50 years old, 30,000 initially healthy women 50 years of age or older in Canada were randomly selected. These 30,000 women were randomly assigned to either receive 100 mg of aspirin or placebo (vitamin) on alternative days and were monitored for 10 years. The results showed that the percentage of heart disease and stroke is significantly lower in the aspirin group than in the placebo group. Therefore, we claim that aspirin is effective in preventing heart disease and stroke for all women over 50 years old in Canada.
 - a. Identify the population and sample in this study.
 - b. Is this study inferential or descriptive?
 - c. Is this an observational study or a design experiment?
2. The following table gives the salaries (in thousand dollars) for physics and computer science (CS) majors obtaining a bachelor's degree, a master's degree or a PhD.

SALARY	MAJOR	DEGREE	SALARY	MAJOR	DEGREE
51.9	Physics	Bach	50.8	CS	Bach
58.2	Physics	Bach	59.4	CS	Bach
49.9	Physics	Bach	55.9	CS	Bach
50.6	Physics	Bach	45.1	CS	Bach
51.4	Physics	Bach	54.1	CS	Bach
43.7	Physics	Bach	50.7	CS	Bach
52.9	Physics	Bach	46.8	CS	Bach
59.2	Physics	Master	65.8	CS	Master
60.5	Physics	Master	57.5	CS	Master
57.1	Physics	Master	66.9	CS	Master
59.1	Physics	Master	62.8	CS	Master
54.9	Physics	Master	68.5	CS	Master
61.7	Physics	Master	69.3	CS	Master
62.4	Physics	Master	61.5	CS	Master
78.2	Physics	PhD	73.3	CS	PhD
69.6	Physics	PhD	65.7	CS	PhD
70.5	Physics	PhD	71.7	CS	PhD
73.2	Physics	PhD	72.5	CS	PhD
81.7	Physics	PhD	73.0	CS	PhD
74.8	Physics	PhD	67.2	CS	PhD
69.8	Physics	PhD	67.5	CS	PhD

- Use the most proper graphs to summarize each of the three variables "salary", "major" and "degree".
- Describe the shape of the histogram of salary.

Show/Hide Answer

- The population is all women over 50 years old in Canada. The sample is those 30,000 women.
 - This is an inferential study since it generalized the conclusion to the entire population—all women over 50 years old in Canada.

- c. This is a designed experiment, since the subjects were **randomly assigned** to the two different groups.

2.

- a. The variable “salary” is a quantitative continuous variable, we can use histogram, boxplot or stem-and-leaf diagram to summarize it. Both “major” and “degree” are qualitative variables, we can use either a pie chart or a bar chart to summarize them.
- b. It is bimodal. It is hard to tell whether it is symmetric or skewed. We might be able to tell the shape better if we draw a boxplot as well.

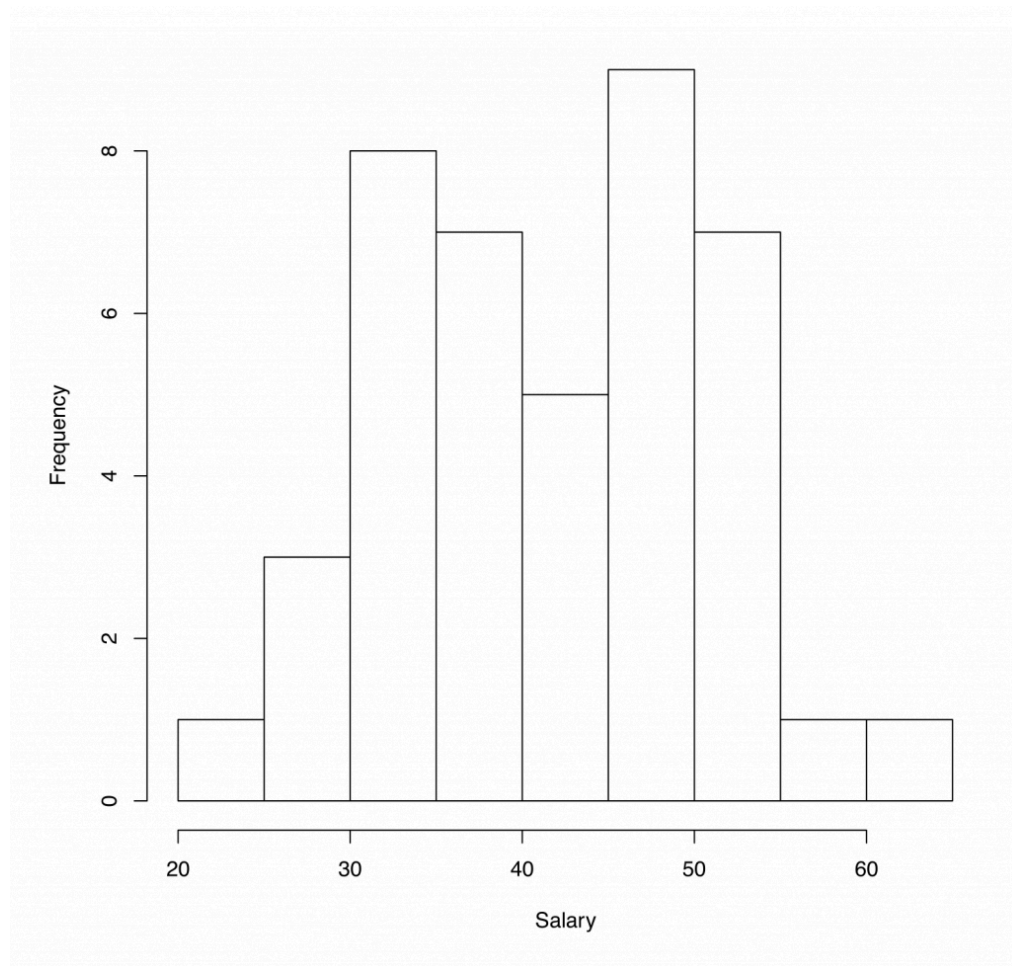


Figure 1.10 [[Image Description \(See Appendix D Figure 1.10\)\]](#)

1.8 Assignment 1

Purposes

This assignment has two parts. The first part assesses your knowledge of the concept of sample, population, descriptive vs. inferential study, using graphs to summarize data. The second part assesses your skills in using R commander to create the graphs and interpreting their meaning.

Resources

[M01_SaleHome.xlsx](#)

Instructions

Part A

Important note:

By default, in all assignments in this course, you are required to complete the questions or tasks in Part A by hand. This means that to do any calculation or drawing, you will NOT use R commander or any computer application. That is, you will do calculations manually with a non-programmable scientific calculator and use a pen or pencil to draw figures or build a distribution table on paper (or drawing pad/tablet) and take a photo of it and insert it below the answer space of the question. **Before you start your assignment, you should get a calculator that has Statistic functions.**

Before you complete Part B using R commander, you should read and practice the R commander steps by following the guidance in the Lab Manual.

Complete the following:

1. In order to investigate the effectiveness of a new method in teaching STAT 151 at MacEwan University. From all students who are taking or who will be taking STAT 151 at MacEwan University, a simple random sample of 800 students were selected for this

study. From these 800 students, a simple random sample of 200 students were chosen and assigned to sections taught with the new method, while the other 600 were taught with the standard method. At the end of the term, the class average in the final grade for students taught with the new method is 3% higher than that for students taught with the regular method.

- a. Identify the sample and population of this study. (2 marks)
 - b. Is this study a designed experiment or an observational study? Explain why. (2 marks)
 - c. Is this study descriptive or inferential? Explain why. (2 marks)
 - d. Propose two different methods to take a simple random sample of 200 students from a collection of 800 students. (2 marks)
2. The following table shows the prices of 30 sale homes (in the last column) and 9 features of the homes: size (area of the home measured in square feet), pool (indicating whether the property has a swimming pool or not), area (total area of the lot in square feet), age (age of the home), bath (# of bathrooms), stories (# of stories), garage (# of cars can be parked), traffic (whether the property faces street subject to a constant flow of daily traffic), roof (whether the home has a tile roof or non-tile roof).

Note: the following data is the first 30 ones on the spreadsheet of M01_SaleHome.xlsx you will download for the Part B tasks.

	SIZE	POOL	AREA	AGE	BATH	STORIES	GARAGE	TRAFFIC	ROOF	PRICE
1	1865	Yes	9509.4	18	2.5	1	2	No	Non-tile	145950
2	2576	Yes	11076.9	15	3.0	2	3	No	Non-tile	160000
3	2576	Yes	10168.8	15	3.0	2	2	No	Non-tile	184000
4	2056	Yes	13430.4	15	2.0	1	2	No	Non-tile	152000
5	1730	Yes	11083.5	17	2.0	1	2	No	Non-tile	149000
6	1882	Yes	10559.8	18	2.5	1	2	No	Non-tile	132000
7	2102	Yes	14533.7	16	2.0	1	2	No	Non-tile	150000
8	2461	Yes	9596.2	6	3.0	1	3	No	Tile	190000
9	2461	Yes	10231.5	6	3.0	1	3	No	Tile	226000
10	1514	Yes	10911.5	16	2.0	1	2	No	Non-tile	120000
11	1994	Yes	13605.7	17	2.0	1	2	No	Non-tile	141000
12	2455	Yes	14704.1	16	3.5	1	2	No	Non-tile	169000
13	1730	Yes	14623.4	17	2.0	1	2	No	Non-tile	138600
14	1655	No	9747.7	18	2.5	1	2	No	Non-tile	124000
15	1865	Yes	9932.9	18	2.5	1	2	Yes	Non-tile	130000
16	1882	Yes	10274.4	18	2.5	1	2	No	Tile	150000
17	2718	Yes	9675.3	6	3.5	1	3	No	Tile	243000
18	1882	Yes	11825.1	18	2.5	1	2	No	Non-tile	137900
19	1882	No	14831.5	18	2.5	1	2	Yes	Non-tile	111500
20	1994	Yes	16122.5	17	2.0	1	2	No	Non-tile	152000
21	2214	Yes	12358.3	18	2.5	1	2	No	Non-tile	147000
22	2718	Yes	16214.1	6	3.5	1	3	No	Tile	245000
23	2576	Yes	12055.5	15	3.0	2	3	No	Non-tile	175000
24	3124	No	9497.6	6	3.5	1	3	No	Tile	242500
25	2128	Yes	9823.7	15	2.5	1	2	No	Non-tile	152000
26	1655	Yes	10520.5	18	2.5	1	2	No	Non-tile	137000
27	2214	No	10739.0	18	2.5	1	2	No	Non-tile	148000
28	2576	Yes	11087.7	15	3.0	2	2	No	Non-tile	175000
29	2928	Yes	16458.6	10	3.5	1	2	No	Tile	210000
30	2576	Yes	10368.5	15	3.0	2	3	No	Non-tile	169900

[Image Description (See Appendix D Question 2)]

- Identify the type of data provided in each column as qualitative, quantitative discrete, or quantitative continuous. (10 marks)
- For the variable "size," (total 10 marks)
 - Obtain a frequency distribution using [1400, 1600) as the first sub-interval, [1600, 1800) as the second sub-interval, [1800, 2000) as the third, and etc. and insert it in the space below. (2 marks)
 - Obtain a relative frequency distribution based on part (1) and insert it in the space below. (2 marks)
 - Construct a relative-frequency histogram and insert it below. (3 marks)
 - Describe the graph you constructed in part (3) about its overall shape, modality, symmetry/skewness, if applicable. (3 marks)
- For the variable "bath," do the following: (total 9 marks)

1. Obtain a frequency distribution. (2 marks)
 2. Obtain a relative frequency distribution based on part (1). (2 marks)
 3. Construct a graph corresponding to part (1). (3 marks)
 4. Describe the graph obtained in part (3). (2 marks)
- d. For the variable “roof,” do the following: (total 12 marks)
1. Obtain a frequency distribution. (2 marks)
 2. Obtain a relative frequency distribution based on part (1). (2 marks)
 3. Construct two different types of graphs corresponding to part (2). (6 marks)
 4. Describe the graphs obtained in part (3). (2 marks)

Part B

Finish the following questions using R and R commander:

Read the data set “M01_SaleHome.xlsx” and use R commander to complete the following tasks. **For each, you need to copy or do a screenshot of the output in R commander (we later call it computer output) and paste it into the space below the questions.** To save space, you only need to copy and paste what is asked for in the questions, and sometime may need to shrink the size.

1. Use the most suitable type of graphs to summarize the prices of these 88 sale homes. Comment on the distribution of the price in terms of overall shape, modality, symmetry/skewness if applicable. (5 marks)
2. Use a suitable graph(s) we taught in Module 1 to compare the prices of homes with a tile roof and a non-tile roof. Briefly explain your findings based on the graph(s). (5 marks)
3. Use R commander to obtain a contingency table with “roof” as the row variable and “pool” as the column variable (2 marks).
Based on the computer outputs from R commander, obtain the percentages in the following four questions.
 - a. Homes with a swimming pool. (1 mark)
 - b. Homes without a swimming pool. (1 mark)
 - c. Homes with a swimming pool and with a tile roof. (1 mark)
 - d. Homes without a swimming pool and with a non-tile roof. (1 mark)
4. Use the most suitable graph to show the effect of “Size” on the “Price” of the sale home. Briefly describe the relationship you found. (5 marks)

Quiz 1



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2598#h5p-2>

CHAPTER 2: DESCRIPTIVE STATISTICS

Overview

This chapter introduces how to describe the center and spread (variation) of a distribution using descriptive measures. The measures for the center covered in this chapter are median, mean, and mode. The measures for the variation are range, interquartile range, and standard deviation.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Choose proper measures to describe the center and spread (variation) of a distribution.
- Calculate the mean, median, mode, standard deviation, range, and interquartile range of the given data, if applicable.
- Calculate the quartiles and five-number summary of the data.
- Draw a boxplot for the given data.
- Explain the meaning of a z-score.

2.1 Centre of a Distribution

The centre of a distribution is in general referred to as the most typical value of the distribution. There are three ways to describe the centre of a distribution—median, mean, and mode.

2.1.1 Median

What is the centre of a ruler? The centre could be viewed as the balance point that cuts the ruler into two halves with equal weight. A similar idea can be applied to the centre of a distribution: the centre of a distribution can be viewed as the value that divides the sorted data into two halves with an equal number of observations. This value is known as the **median**. That is, 50% of the observations are below the median and another 50% are above the median. Here are the steps to find the median of a set of data:

1. Sort the data from the smallest to the largest.
2. If the total number of observations n is odd, the median is the observation standing in the middle of the sorted list.
3. If n is an even number, the median is the average of the two values in the middle of the sorted list.

Example: Find the Median

Find the Median of the Data 3, 5, 3, 7, 7.

Steps:

- Sort into 3, 3, 5, 7, 7.
- $n = 5$ which is odd, median = 5.



Activity

Exercise: Find the Median

Find the median of the following data sets:

1. 3, 5, 3, 7, 997
2. 3, 5, 3, 7
3. Male, Female, Male, Male, Female, Female, Female

Show/Hide Answer

1. 3, 5, 3, 7, 997
Sort into 3, 3, 5, 7, 997.
 $n = 5$ which is odd, median = 5.
2. 3, 5, 3, 7
Sort into 3, 3, 5, 7
 $n = 4$ which is even, median = $\frac{3+5}{2} = 4$.
3. Male, Female, Male, Male, Female, Female, Female
Sort: cannot sort Male and Female; therefore, the median does not exist.

2.1.2 Mean

The **mean** is the average of all observations, which equals the total divided by the number of observations. Suppose the population has N observations denoted as x_1, x_2, \dots, x_N to distinguish them. Therefore, x_i refers to the i th observation, $i = 1, 2, \dots, N$. The **population mean**, denoted as μ can be calculated as

$$\mu = \frac{\text{total}}{N} = \frac{\text{sum}}{N} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N} = \frac{\sum x_i}{N}.$$

The notation " \sum " is the summation sign which means taking the sum of the observations as indicated in the index, i.e., for all values of i from 1 to N . Here N denotes the population size, the number of individuals in the population.

If we have a sample of size n, x_1, x_2, \dots, x_n , we can calculate the **sample mean**, denoted as \bar{x} (read as x-bar), as follows:

$$\bar{x} = \frac{\text{sum}}{n} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum x_i}{n}.$$

Here n is the sample size, the number of individuals in the sample.

In inferential statistics, we often use the sample mean \bar{x} to estimate the value of the population mean μ .

Example: Find the Mean

Find the Sample Mean of the Data 3, 5, 3, 7, 7.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+5+3+7+7}{5} = \frac{25}{5} = 5.$$



Activity

Exercises: Find the Mean

Find the mean of the following data sets:

1. 3, 5, 3, 7, 997
2. 3, 5, 3, 7
3. Male, Female, Male, Male, Female, Female, Female

Show/Hide Answer

1. $\bar{x} = \frac{\sum x_i}{n} = \frac{3+5+3+7+997}{5} = \frac{1015}{5} = 203.$

The sample mean is much larger than the mean in the previous example due to the extremely large observation of 997.

2. $\bar{x} = \frac{\sum x_i}{n} = \frac{3+5+3+7}{4} = \frac{18}{4} = 4.5.$

3. We cannot calculate the average of qualitative data; therefore, the sample mean does not exist.

2.1.3 Mode

The last measure of centre covered in this course is the **mode**, which is the observation that occurs most often. If at least two observations occur most often, the data set has multiple modes; if all observations occur once, there is no mode.

Example: Find the Mode

Find the Mode of the Data 3, 5, 3, 7, 7.

The observation “3” occurs twice, and so does the observation “7”. The observation “5” occurs only once. Therefore, the modes are 3 and 7.



Activity

Exercise: Find the Mode

Find the mode of the following data sets:

1. 3, 5, 3, 7, 997
2. 3, 5, 9, 7
3. Male, Female, Male, Male, Female, Female, Female

Show/Hide Answer

1. 3, 5, 3, 7, 997
The observation “3” occurs twice, the observations “5,” “7,” and “997” each occur only once. Therefore, the mode is 3.
2. 3, 5, 9, 7
Each observation occurs only once, so there is no mode.
3. Male, Female, Male, Male, Female, Female, Female
There are three males and four females. Thus, the mode is “Female.”

2.1.4 Choose the Proper Measure to Describe Centre

Mean, median, and mode are the three measures of the centre. The following section provides some practical guidelines for choosing the proper measure to describe the centre of the data.

Mean Versus Median

Both the mean and the median are measures of the center of a distribution. When a distribution is symmetric, the mean and the median are equal. However, it is better to use the mean to describe the centre of a symmetric distribution for several reasons (one of which is introduced in Chapter 6). On the other hand, when a distribution is skewed or when it contains outliers, it is better to use the median. This is because the mean includes every observation from a data set and, as such, it is easily influenced by extremely large or small values (called outliers). Conversely, the median does not include every observation from a data set but only the centralmost value(s). For this reason, the median is highly resistant to outliers. Recall the two data sets in previous examples: {3, 3, 5, 7, 7} and {3, 3, 5, 7, 799}. These two data sets have the same median of 5; however, they have very different sample means (5 versus 203) due to the extensive observation (i.e., 799) in the second data set.

The following graphs show the relationship between the mean (red solid) and the median (blue dashed) for symmetric, right-skewed, and left-skewed distributions.

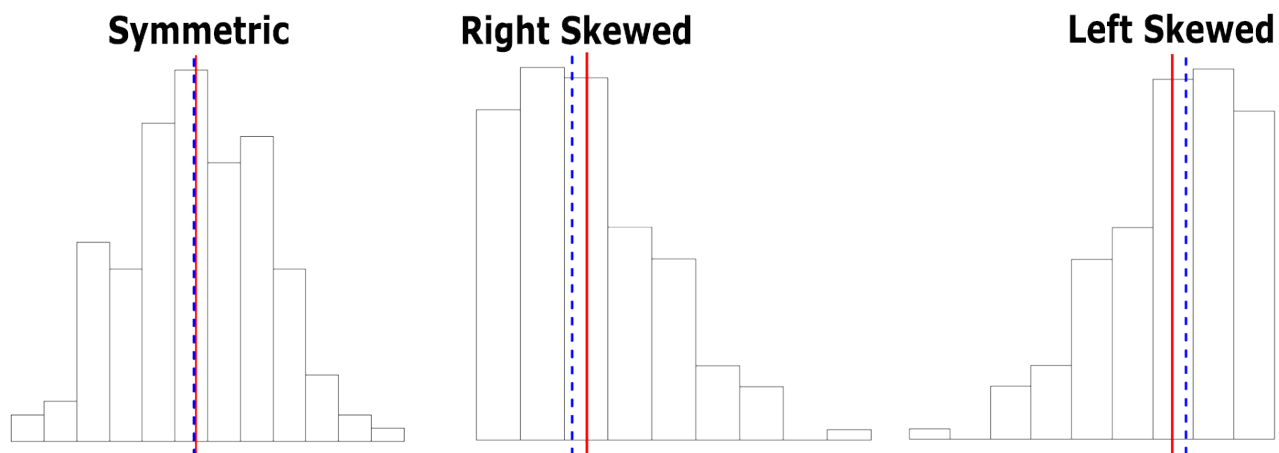


Figure 2.1: Compare Mean (red solid) and Median (blue dashed) [[Image Description \(See Appendix D Figure 2.1\)](#)]

We can tell from the figures:

- For the right-skewed distribution (longer tail on the right-hand side), mean > median because the observations on the right tail drag the mean to the right.
- For the symmetric distribution, mean = median. Both divide the distribution into two parts with roughly equal number of observations.
- For the distribution that is skewed to the left (longer tail on the left-hand side), mean < median because the observations on the left tail drag the mean to the left.

Summary of the Centre

Here are some guidelines for choosing the proper measure to describe the centre of a distribution:

- Use the median when the distribution is extremely skewed or outliers exist.
- Use the mean when the distribution is symmetric and there are no outliers.
- For qualitative/categorical data, we can only use the mode to describe the center.
- For quantitative data, the mode can also be computed. However, it is not as informative as the median or the mean.

2.2 Quartiles and Percentiles

In addition to the measures of centre, some other measures can be used to describe a distribution such as quartiles and percentiles. Recall that the median is the middle value that divides the sorted data into two halves with an equal number of observations; that is, 50% of the observations are below the median, and another 50% are above the median. Similarly, quartiles and percentiles of a distribution are defined as follows:

- **Quartiles** are the three values that divide the sorted data into four parts with an equal number of observations, denoted as Q_1, Q_2, Q_3 . Each part contains 25% of the data. Actually, the second quartile Q_2 is the median of the entire data set; the first quartile Q_1 is the median of the bottom 50% (first half) and the third quartile Q_3 is the median of the top 50% (second half). **Note that when the number of observations n is odd, we include the median in both the first half and the second half when calculating Q_1 and Q_3 .**
- **Percentiles** are those 99 values that divide the sorted data into 100 parts with an equal number of observations. Each part contains 1% of the data. The first quartile Q_1 is the 25th percentile, the second quartile Q_2 (median) is the 50th percentile, and the third quartile Q_3 is the 75th percentile. In this course, we will not calculate percentiles by hand, except for the important special cases of quartiles. The software can calculate any arbitrary percentiles for us.

Example: Find the Quartiles

Find the quartiles for 3, 1, 9, 7, 5, 11, 21

Steps:

1. Sort into 1, 3, 5, **7**, 9, 11, 21.
2. $n = 7$ is odd, $Q_2 = \text{median} = 7$.
3. The bottom half consists of the first three smallest observations and the median, i.e., 1, **3, 5**, 7. Q_1 is the median of the first half, i.e., $Q_1 = \frac{3+5}{2} = 4$.
4. The top half consists of the three largest values and the median, i.e., 7, **9, 11**, 21. Q_3 is the median of the second half, $Q_3 = \frac{9+11}{2} = 10$.

Therefore, the quartiles are $Q_1 = 4, Q_2 = 7, Q_3 = 10$.

Note: since the number of observations $n = 7$ which is odd, we include the median $Q_2 = 7$ in both the first half $\{1, 3, 5, \mathbf{7}\}$ and the second half $\{\mathbf{7}, 9, 11, 21\}$.



Activity

Exercise: Find the Quartiles

Find the quartiles for the data 3, 1, 9, 7, 5, 11, 21, 19.

Show/Hide Answer

Steps:

1. Sort into 1, 3, 5, 7, 9, 11, 19, 21.
2. $n=8$ is even, $Q_2 = \text{median} = \frac{7+9}{2} = 8$.
3. The bottom half is the first four observations in the sorted list 1, 3, 5, 7, and Q_1 is the median of the first half, i.e., $Q_1 = \frac{3+5}{2} = 4$.
4. The top half is the last four observations 9, 11, 19, 21, and $Q_3 = \frac{11+19}{2} = 15$ is the median of the second half.

Therefore, the quartiles are $Q_1 = 4$, $Q_2 = 8$, $Q_3 = 15$.

2.3 Spread (Variation) of a Distribution

Besides to the centre, we need another descriptive measure to describe how the data spread out. That is called the **spread** or **variability** of the distribution. Measures of variation covered are the range, interquartile range (IQR), and standard deviation.

2.3.1 Range and Interquartile Range (IQR)

One intuitive measure of the spread is the **range** of the data, which is defined as the difference between the largest and the smallest observations,

$$\text{range} = \text{maximum} - \text{minimum} = \text{largest} - \text{smallest}.$$

Similar to the mean, range is sensitive to outliers.

We can use the interquartile range

$$IQR = Q_3 - Q_1,$$

which is the difference between Q_3 and Q_1 to describe the spread if the distribution is extremely skewed or outliers exist. The IQR is often paired with the median to describe the spread and the centre of a distribution respectively.

2.3.2 Standard Deviation

Like the mean, the standard deviation takes into account all the observations and measures variation by indicating on average of how far the observations are away from the mean. For a data set with a large amount of variation, i.e., the observations are very different from one another, the standard deviation will be large. For a data set with a small amount of variation, on average, the observations are close to the mean, so the standard deviation will be small.

Steps to calculate the sample standard deviation are:

1. Calculate the sample mean of the data set, \bar{x} .

2. For each observation x_i , find its deviation from the mean \bar{x} , denoted as $(x_i - \bar{x})$. The sum of the deviations always equals zero, i.e., $\sum_{i=1}^n (x_i - \bar{x}) = 0$.
3. In order to obtain quantities that do not sum to zero, take the square of the deviations. The sum of squared deviations, $\sum_{i=1}^n (x_i - \bar{x})^2$ gives a measure of total variation of all the observations.
4. Finally, the **sample standard deviation**, denoted as s , is calculated as $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$.

This is referred as the defining formula of the sample standard deviation.

The term

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

is defined as the **sample variance** of the data. Roughly speaking, it gives the average squared distance from each observation x_i to the sample mean \bar{x} . The square root of the sample variance s^2 gives the sample standard deviation s . Roughly speaking, the sample standard deviation s can be interpreted as the average distance from each observation x_i to the sample mean \bar{x} . Just as the sample mean \bar{x} is used to estimate the population mean μ , the sample variance s^2 can be used to estimate the population variance σ^2 , and the sample standard deviation s can be used to estimate the population standard deviation σ .

It can be shown that

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

And the sample standard deviation becomes

$$s = \sqrt{\frac{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}{n-1}}.$$

This is referred as the **computing formula** of the sample standard deviation. The defining formula $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ is helpful in understanding the meaning of the sample standard deviation; while the computing formula $s = \sqrt{\frac{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}{n-1}}$ is useful in calculating the sample standard deviation by hand since it involves much less calculations.

The standard deviation is often paired with the mean to describe the spread and the centre of a distribution respectively.

Example: Measures of Spread (Variation)

Find the range, IQR, and sample standard deviation for 3, 5, 3, 7, 7.

1. For range

- Sort into **3, 3, 5, 7, 7**. The minimum (smallest observation) is 3, and the maximum (largest observation) is 7.
- $\text{range} = \text{maximum} - \text{minimum} = 7 - 3 = 4$

2. For IQR

- Sort into 3, 3, **5**, 7, 7.
- $n = 5$ is odd, median $Q_2 = 5$.
- The first half is 3, **3**, 5. The median of the first half is $Q_1 = 3$. The second half is 5, **7**, 7. The median of the second half is $Q_3 = 7$.
- $IQR = Q_3 - Q_1 = 7 - 3 = 4$

3. For sample standard deviation, the following table shows the calculation of the sample standard deviation.

Table 2.1: Calculate the Sample Standard deviation Using the Computing Formula

x_i	x_i^2
3	$3^2=9$
5	$5^2=25$
3	$3^2=9$
7	$7^2=49$
7	$7^2=49$
$\sum x_i = 25$	$\sum x_i^2 = 141$

$$\begin{aligned}
 s &= \sqrt{\frac{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}{n - 1}} \\
 &= \sqrt{\frac{141 - \frac{25^2}{5}}{5 - 1}} = \sqrt{\frac{141 - 125}{4}} = \sqrt{\frac{16}{4}} = \sqrt{4} = 2.
 \end{aligned}$$

Interpretation: Roughly speaking, the average distance between the observations and the sample mean is 2.

If you would like to use the defining formula, it is helpful to construct the following table:

Table 2.2: Calculate the Sample Standard deviation Using the Defining Formula

x_i	Deviation: $(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3	$3 - 5 = -2$	$(-2)^2 = 4$
5	$5 - 5 = 0$	$0^2 = 0$
3	$3 - 5 = -2$	$(-2)^2 = 4$
7	$7 - 5 = 2$	$2^2 = 4$
7	$7 - 5 = 2$	$2^2 = 4$
$\sum x = 25$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 16$

The sample standard deviation calculated by the defining formula is

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{16}{5-1}} = 2$$

which is the same as the value obtained by the computing formula.

2.3.3 Summary: Choose Proper Measures

Here are some guidelines for choosing proper measures to describe the centre and spread (variation) of a distribution:

- Use the **median** and the **IQR** for the centre and spread respectively when the distribution is skewed or outliers exist.
- Use the **mean** and the **standard deviation** for the centre and spread respectively when the distribution is roughly symmetric and there are no outliers.
- Although the mode may also be used as a measure of centre for numerical data, it is usually not as informative as the median or the mean.
- For categorical data, the mode is the only descriptive measure we can use to describe the center of qualitative/categorical data. None of the measures of spread covered in this chapter (i.e., range, IQR, standard deviation) can be applied to qualitative/categorical data.

2.4 Five-Number Summary and Boxplot

The five-number summary of a data set consists of the minimum (the smallest observation), Q_1 , Q_2 , Q_3 and the maximum (the largest observation).

These five numbers together give us a brief idea about the distribution of the data: Q_2 (the median) is the centre of the distribution, the range (the difference between the maximum and the minimum) and the IQR (the difference between Q_3 and Q_1) tell us the spread (variation) of the data. The difference between Q_1 and the minimum, between Q_2 and Q_1 , between Q_3 and Q_2 , and between the maximum and Q_3 give the range of the first, second, third and fourth 25% of the data respectively. Moreover, the five-number summary helps us identify outliers, those observations that are far away from the bulk of the data.

2.4.1 Identify Outliers

Outliers are observations far away from the majority of the data. Quantitatively, any observation that falls outside the interval of (lower limit, upper limit) is considered as an outlier. The upper and lower limits are defined as:

$$\text{lower limit} = Q_1 - 1.5 \times IQR; \quad \text{upper limit} = Q_3 + 1.5 \times IQR.$$

Example: Identify Outliers

Identify the outliers for the data 3, 1, 9, 7, 5, 11, 21 if any.

Steps:

1. Find the quartiles. Refer to Example 4, part (a), $Q_1 = 4$, $Q_2 = 7$, $Q_3 = 10$.
2. $IQR = Q_3 - Q_1 = 10 - 4 = 6$
3. lower limit $= Q_1 - 1.5 \times IQR = 4 - 1.5 \times 6 = -5$
4. upper limit $= Q_3 + 1.5 \times IQR = 10 + 1.5 \times 6 = 19$

Since $21 > 19$, it is outside the interval $(-5, 19)$, 21 is an outlier.

Exercise: Choose Proper Measures

Based on the histogram and five-number summary of the data, answer the following questions.

Table 2.3: Five-Number Summary of the Data

Summary	Min	Q ₁	Median	Q ₃	Max
	0.1	2	3.5	5	32

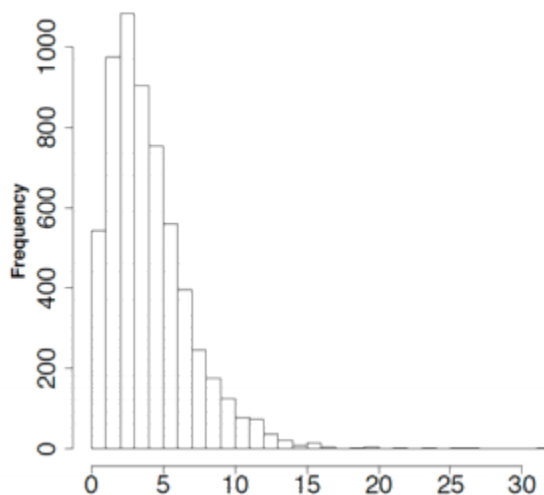


Figure 2.2: Histogram of the Data [\[Image Description\]](#)
[\(See Appendix D Figure 2.2\)\]](#)

1. Comment on the distribution (shape, centre, spread).
2. Are there any outliers in the data?
3. Provide proper measures of the centre and spread of the data. Explain why.

Show/Hide Answer

1. Comment on the distribution (shape, centre, spread).
The distribution is unimodal, skewed to the right with a median 3.5 and $IQR = 5 - 2 = 3$.
2. Are there any outliers in the data?
Yes. Upper limit = $Q_3 + 1.5 \times IQR = 5 + 1.5 \times 3 = 9.5$.
Any observation greater than 9.5 is an outlier.
3. Provide proper measures of the centre and spread of the data. Explain why.
Use median for the centre and IQR for the spread due to outliers and strong skewness.

2.4.2 Boxplot

A **boxplot**, also called a box-and-whisker plot, is a useful tool to display the centre and spread of a data set by providing a graphical representation of the five-number summary as well as potential outliers. Steps to draw a boxplot:

1. Calculate the five-number summary: minimum, Q_1 , Q_2 , Q_3 , and maximum.
2. Calculate the lower and upper limits: lower limit = $Q_1 - 1.5 \times IQR$, and upper limit = $Q_3 + 1.5 \times IQR$.
3. Find the **adjacent values**, the largest and smallest observations **within the lower and upper limits**. Identify the potential outliers (observations beyond the upper and lower limits), if any exist.
4. Draw short horizontal lines at Q_1 , Q_2 , Q_3 , and connect them with vertical lines to form a box.
5. Draw very short horizontal lines at the adjacent values and then draw the whiskers by connecting the adjacent values and the box with vertical lines.
6. Plot each potential outlier with an asterisk.
7. Put labels and the title.



Instructor's Note

- A boxplot can be drawn vertically or horizontally.
- Symbols such as circles or asterisks are often used to plot potential outliers.

Example: Draw a Boxplot

Construct a boxplot for the data 3, 1, 9, 7, 5, 11, 21.

Steps:

1. Calculate the five-number summary:
sort: 1, 3, 5, 7, 9, 11, 21
 $min = 1$, $Q_1 = 4$, $Q_2 = 7$, $Q_3 = 10$, $max = 21$
2. Calculate the lower and upper limits
 $IQR = Q_3 - Q_1 = 10 - 4 = 6$
lower limit = $Q_1 - 1.5 \times IQR = 4 - 1.5 \times 6 = -5$

upper limit = $Q_3 - 1.5 \times IQR = 10 + 1.5 \times 6 = 19$.

3. Adjacent values are 1 and 11, so the max 21 is an outlier.
4. Form a box based on $Q_1 = 4$, $Q_2 = 7$, $Q_3 = 10$.
5. Mark the adjacent values 1 and 11, “grow the whiskers,” the dashed lines connecting the box and the adjacent values.
6. Plot the potential outlier with 21.
7. Title and label the boxplot.

Example Boxplot

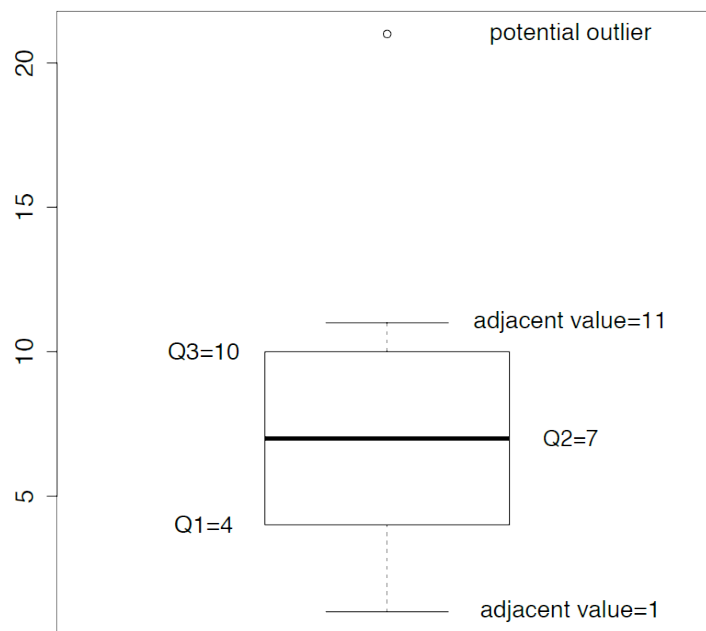


Figure 2.3: Resulting Boxplot of the Example [[Image Description \(See Appendix D Figure 2.3\)](#)]

We can describe the distribution of the data in the following aspects based on a boxplot:

- The centre: the median Q_2 .
- The spread (variation): the range and IQR. Note that, however, the range is sensitive to outliers.
- The shape of the distribution:
 - **Left skewed** if the distance between the lower adjacent value and Q_1 is larger than the distance between the upper adjacent value and Q_3 , and the distance between Q_1 and the median is larger than the distance between Q_3 and the median.
 - **Right skewed** if the distance between the lower adjacent value and Q_1 is smaller than the distance between the upper adjacent value and Q_3 , and the distance

between Q_1 and the median is smaller than the distance between Q_3 and the median.

- **Symmetry** if the distance between the lower adjacent value and Q_1 is approximately equal to the distance between the upper adjacent value and Q_3 , and the distance between Q_1 and the median is approximately equal to the distance between Q_3 and the median.
- Note that it is sometimes the case that the whiskers show skewness in one direction while the box shows skewness in the opposite direction. In such cases, it is not always possible to clearly determine skewness or symmetry.
- Identify outliers.

The following are three boxplots that show right skewed, symmetric, and left skewed distributions respectively.

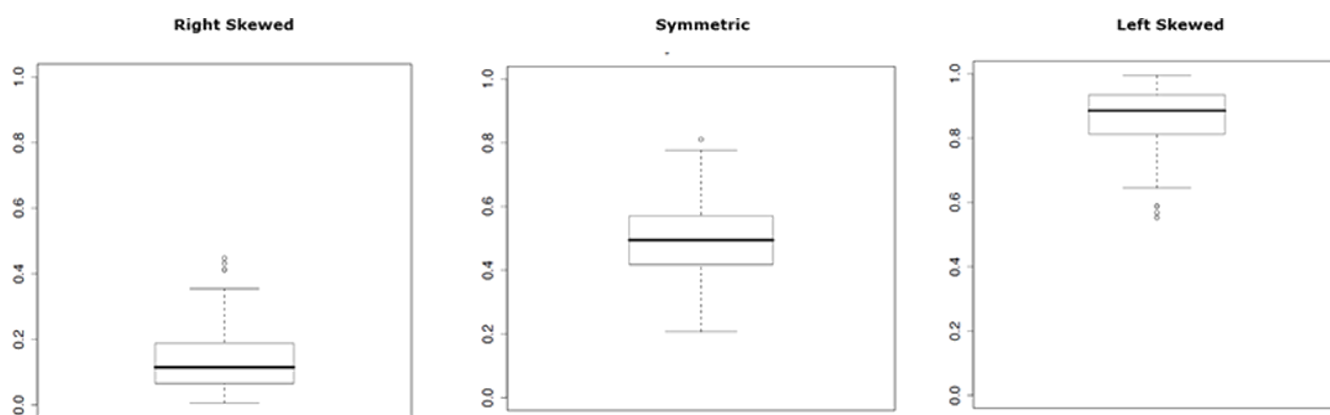


Figure 2.4: Boxplots of Skewed and Symmetric Distributions. [[Image Description \(See Appendix D Figure 2.4.\)](#)]

Similar to side-by-side histograms, we can use side-by-side boxplots to compare different groups.

Example: Side-by-Side Boxplots

I want to compare grades of students who attend lectures with those who do not. Both the table and the side-by-side boxplots tell us that:

- Attendees have a larger median score.
- Non-attendees have a slightly larger variation. Both the IQR (height of the box) and standard deviation of non-attendees are larger than that of attendees.

- Grades of both groups are slightly left skewed with a longer tail on the lower end.

Table 2.4: Numerical Summaries of Grades of Non-Attendees and Attendees

Summary	Min	Q ₁	Median	Q ₃	Max	Mean	SD
Non-attendees	35.62	52.70	64.76	77.78	87.30	63.23	15.48
Attendees	47.77	69.80	77.83	85.15	96.51	76.92	11.83

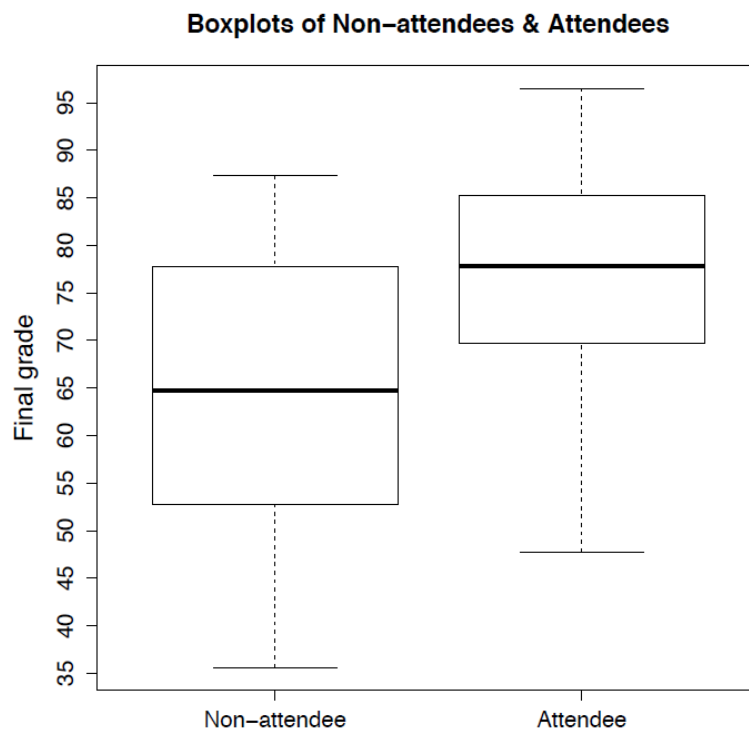


Figure 2.5: Side-by-Side Boxplots of Non-Attendees and Attendees. [[Image Description \(See Appendix D Figure 2.5\)](#)]



Exercise: Draw a Boxplot

Draw a boxplot for the data: -5, 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95.

Show/Hide Answer

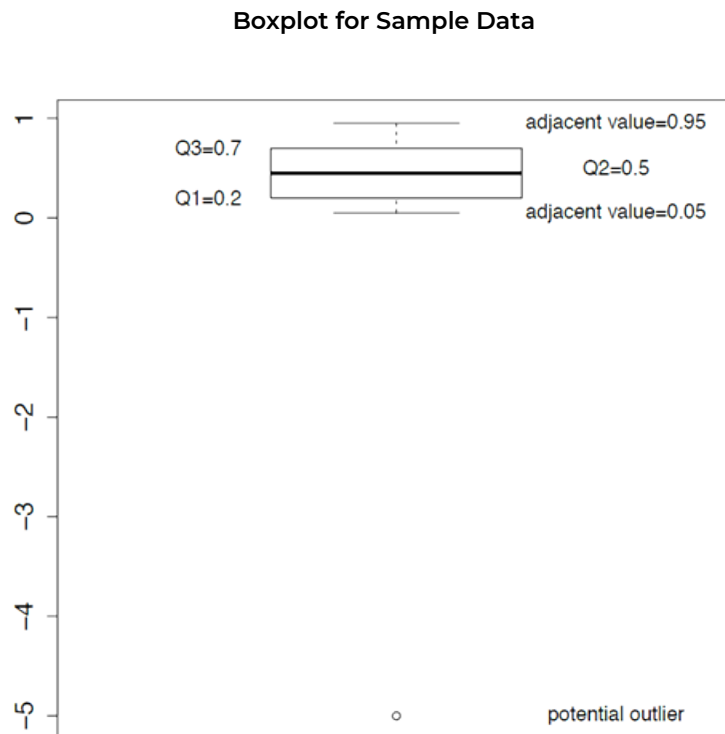


Figure 2.6: Boxplot for the Sampled Data. [\[Image Description \(See Appendix D Figure 2.6\)\]](#)

Steps:

1. Calculate five-number summary:
sort: -5, 0.05, 0.15, 0.25, 0.35, **0.45**, 0.55, 0.65, 0.75, 0.85, 0.95
 $\min = -5$, $Q_1 = 0.2$, $Q_2 = 0.45$, $Q_3 = 0.7$, $\max = 0.95$
2. Calculate the lower and upper limits
 $IQR = Q_3 - Q_1 = 0.7 - 0.2 = 0.5$
lower limit $= Q_1 - 1.5 \times IQR = 0.2 - 1.5 \times 0.5 = -0.55$
upper limit $= Q_3 + 1.5 \times IQR = 0.7 + 1.5 \times 0.5 = 1.45$
3. Adjacent values are 0.05 and 0.95, the min -5 is an outlier.
4. Form a box based on $Q_1 = 0.2$, $Q_2 = 0.45$, $Q_3 = 0.7$.
5. Mark the adjacent values 0.05 and 0.95, and then draw the whiskers, the dashed lines connecting the box and the two adjacent values.
6. Plot the outlier -5.
7. Title and label boxplot.

2.5 Descriptive Measures for Population and Sample

We summarize the descriptive measures for the population and for the sample in the following table. Note that a summation sign without indices means taking the sum of all observations in the data, e.g., the population mean $\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{\sum x_i}{N}$.

Population

Definition: The collection of all individuals under consideration in a study.

Population size N = the total number of individuals in the population.

Population mean μ : Suppose the measurement of each individual is x_1, x_2, \dots, x_N , the population mean is defined as

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N} = \frac{\sum x_i}{N}.$$

Population standard deviation, σ , is the square root of the population variance σ^2 . It is defined as

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

The following formula is helpful in calculating the population standard deviation,

$$\sigma = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \mu^2} = \sqrt{\frac{\sum x_i^2}{N} - \mu^2}$$

A descriptive measure for a population, such as μ and σ , is called a *parameter*.

Sample

Definition: Part of or a subset of the population from which information is obtained.

Sample size n = the total number of individuals in the sample.

Sample mean \bar{x} : Suppose the measurements of the sample are x_1, x_2, \dots, x_n , the sample mean is defined as

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum x_i}{n}.$$

Sample standard deviation, s , is the square root of the sample variance s^2 . It is defined as

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

The following formula is helpful in calculating the sample standard deviation,

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

A descriptive measure for a sample, such as \bar{x} and s , is called a *statistic*.

2.6 Z-Score as a Measure of Relative Standing

A college admissions officer is looking at the files of two international candidates, one with a computer-based TOEFL score of 210 and the other with an IELTS score of 7.5. Which score is better? How can we compare measurements with different scales? The solution is the z-score, which gives a relative standing among the population.

Suppose variable X follows a distribution with a mean μ and a standard deviation σ , the corresponding standardized variable is defined as

$$Z = \frac{X - \mu}{\sigma}.$$

It can be shown that the standardized variable Z has a mean 0 and a standard deviation 1.

For a given value of X , denoted as the corresponding lower-case x , the value of the standardized variable is called the z-score of x , which is given by

$$z = \frac{x - \mu}{\sigma}.$$

If the population parameters μ and σ are unknown, use the sample mean and standard deviation \bar{x} and s to estimate them, then the z-score becomes

$$z = \frac{x - \bar{x}}{s}.$$

Properties of the z-score:

- It measures how far an individual is away from the mean using standard deviation as the unit (ruler).
- It represents a relative standing of an observation.
- A z-score > 0 means the observation x is above the mean; z-score < 0 means the observation x is below the mean; z-score $= 0$ means the observation x is equal to the mean.
- A z-score has no unit.

Example: Z-score

Suppose that the TOEFL score for admission at MacEwan has a mean 200 and a standard deviation 10, and the IELTS score has a mean 6 and a standard deviation 1. Two international candidates, one with a TOEFL score of 210 and the other with an IELTS score of 7.5. Which score is relatively better?

Calculate the z-score for the student who took the TOEFL exam:

$$z_T = \frac{x - \mu}{\sigma} = \frac{210 - 200}{10} = 1.$$

This z-score means that the student who took the TOEFL exam is one standard deviation **above the average**.

The z-score for the student who took the IELTS exam:

$$z_I = \frac{x - \mu}{\sigma} = \frac{7.5 - 6}{1} = 1.5.$$

This z-score means that the student who took the IELTS exam is 1.5 standard deviation **above the average**. Therefore, the student who took the IELTS exam is better because the z-score is larger. This tells us that the IELTS student score is relatively further above the mean than the TOEFL student score is above its mean.



Activity

Exercise: Women Heptathlon Champion

Women's heptathlon in Olympics includes seven track and field events—200-m and 800-m runs, 100-m high hurdles, shot put, javelin, high jump, and long jump. How do you determine the champion?

Show/Hide Answer

The events are measured in different units, some are in metres and some are in seconds. Moreover, for those runs and high hurdles, results are the smaller the better; for jumps, shot put, and javelin, the larger the better. In order to combine the results of the seven events to a single score to determine the winner, one possible solution is the z-score. The idea is as follows:

For each athlete, calculate her z-score in each of those seven events:

Put a negative sign in front of those z-scores whose results are the smaller the better, i.e., for 200-m and 800-m runs, 100-m high hurdles. Note that a positive z-score in 100-m run indicates the athlete is below average, since she spends more time than the mean to finish the run. The athlete has the smallest z-score, which is a negative z-score, is the winner of this event. Putting a negative sign in front of the

smallest z-score will make it positive and the largest z-score for this event. This makes sense in the way that the larger the z-score the better performance.

Add the seven z-scores together and get a single value. The one has the largest value is the winner.

2.7 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Choose proper measures to describe the centre and spread (variation) of a distribution (Section 2.1).
- Calculate the mean, median, mode, standard deviation, range, and interquartile range of the given data, if applicable (Sections 2.1, 2.2 and 2.3).
- Calculate the quartiles and five-number summary of the data (Section 2.4).
- Draw a boxplot for the given data (Section 2.4).
- Explain the meaning of a z-score (Section 2.6).

2.8 Review Questions

1. In 2004, the mean net worth of families in the United States was \$448.2 thousand and the median net worth was \$93.1 thousand. Which measure of center do you think is more appropriate? Explain your answer.
2. Wayne Gretzky, a retired professional hockey player, played 20 seasons in the National Hockey League (NHL) from 1980 through 1999. The number of games in which Gretzky played during each of his 20 seasons in the NHL are as follows: 74, 80, 73, 78, 78, 45, 80, 79, 79, 80, 48, 64, 80, 70, 80, 74, 82, 81, 80, 82.
 - a. Find the mean, median, mode of these 20 numbers. Interpret the three measures for the center.
 - b. Find the quartiles of the data and interpret.
 - c. Find the range, interquartile range, sample standard deviation of the data and interpret.
 - d. Find the five-number summary of the data and draw a boxplot of these 20 numbers. Comment on the resulting boxplot.
 - e. Choose proper measures for the center and spread (variation) of the distribution. Justify your answer.
3. The following table gives the salaries (in thousand dollars) for physics and computer science (CS) majors obtaining a bachelor's degree, a master's degree or a PhD.
 - a. What can we tell from the side-by-side boxplot comparing the salaries of computer science (CS) and physics majors, the one on the left?
 - b. What can we tell from the side-by-side boxplot comparing the salaries of Bachelor, Master and PhD, the one on the right?
 - c. What can we tell from the side-by-side boxplot comparing the salaries of combinations of "Major" and "Degree"?

SALARY	MAJOR	DEGREE	SALARY	MAJOR	DEGREE
51.9	Physics	Bach	50.8	CS	Bach
58.2	Physics	Bach	59.4	CS	Bach
49.9	Physics	Bach	55.9	CS	Bach
50.6	Physics	Bach	45.1	CS	Bach
51.4	Physics	Bach	54.1	CS	Bach
43.7	Physics	Bach	50.7	CS	Bach
52.9	Physics	Bach	46.8	CS	Bach
59.2	Physics	Master	65.8	CS	Master
60.5	Physics	Master	57.5	CS	Master
57.1	Physics	Master	66.9	CS	Master
59.1	Physics	Master	62.8	CS	Master
54.9	Physics	Master	68.5	CS	Master
61.7	Physics	Master	69.3	CS	Master
62.4	Physics	Master	61.5	CS	Master
78.2	Physics	PhD	73.3	CS	PhD
69.6	Physics	PhD	65.7	CS	PhD
70.5	Physics	PhD	71.7	CS	PhD
73.2	Physics	PhD	72.5	CS	PhD
81.7	Physics	PhD	73.0	CS	PhD
74.8	Physics	PhD	67.2	CS	PhD
69.8	Physics	PhD	67.5	CS	PhD

4. The z-score corresponding to an observed value of a variable tells you _____ .
5. A positive z-score indicates that the observation is _____ the mean, whereas a negative z-score indicates that the observation is _____ the mean.
6. Suppose that you obtained 350 points in an exam. The exam has 400 possible points, the mean score is 280 and the standard deviation is 20. Did you do well on the exam? Explain your answer.
7. Each year, thousands of high school students bound for college take the Scholastic Assessment Test (SAT). This test measures the verbal and mathematical abilities of prospective college students. Student scores are reported on a scale that ranges from a low of 200 to a high of 800. In one high school graduating class, the mean SAT math score is 528 with a standard deviation of 105; the mean SAT verbal score is 475 with a standard deviation of 98. A student in the graduating class scored 740 on the SAT

math and 715 on the SAT verbal. Compared to the other students in the graduating class, on which test did the student do better?

Show/Hide Answer

1. Since the mean is much larger than the median, the distribution is extremely right skewed. It is more appropriate to use the **median** to describe the center.

2.

- a. The sum of the data is $\sum x_i = 1487$ and the sum of squares of the data is $\sum x_i^2 = 112665$. Therefore, sample mean $\bar{x} = \frac{\sum x_i}{n} = \frac{1487}{20} = 74.35$. Arrange the data from smallest to largest: 45 48 64 70 73 74 74 78 78 **79 79** 80 80 80 80 80 80 81 82 82.

Sample size $n = 20$, median is $\frac{79+79}{2} = 79$. The mode is 80.

Interpretation: 50% of observations are below 79 and another 50% are above 79 (the median); the average of the observation is 74.35 (the mean); the observation occurs most often is 80 (the mode).

- b. The first half: 45 48 64 70 **73 74** 74 78 78 79, $Q_1 = \frac{73+74}{2} = 73.5$, $Q_2 = 79$. The second half: 79 80 80 80 **80 80** 80 81 82 82, $Q_3 = \frac{80+80}{2} = 80$. Therefore, the quartiles are $Q_1 = 73.5$, $Q_2 = 79$, $Q_3 = 80$.

Interpretation: the bottom 25% of observations are below 73.5, 25% are between 73.5 and 79, 25% are between 79 and 80, the top 25% are above 80.

- c. Range=Max-min=82-45=37. IQR= $Q_3 - Q_1 = 80 - 73.5 = 6.5$.

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{112665 - \frac{1487^2}{20}}{20-1}} = 10.5295.$$

Interpretation: the data spread over an interval of length 37 (range); the middle 50% of the observations spread over an interval of length 6.5 (IQR); roughly speaking, the average distance from the observations to the sample mean 74.35 is 10.5295 (standard deviation).

- d. The 5-number summaries are $min = 45$, $Q_1 = 73.5$, $Q_2 = 79$, $Q_3 = 80$, $max = 82$. The distribution is left-skewed with two outliers at the lower end.
- e. Use median for the center and IQR for the spread (variation) since outliers exist.

3.

- a. CS majors have a higher median salary. Physics majors have a larger variation in salary. The distribution for CS majors is left skewed; while the distribution for physic majors is right skewed.
- b. The median salary for PhD is higher than master, and master is higher than bachelor. Salary for master has a larger IQR than the other two groups. The variations for PhD and bachelor are similar.
- c. For PhD and bachelor, physics majors have a slightly higher median; for master,

however, CS majors have a higher median salary than physics. The variation is similar for CS majors at the three different education level; the variations in salary increase for physics majors when the education level increases.

4. how far the observation is away from the mean in units of standard deviation.
5. A positive z-score indicates that the observation is **above** the mean, whereas a negative z-score indicates that the observation is **below** the mean.
6. The z-score is $z = \frac{350-280}{20} = 3.5$. You did extremely well since you are 3.5 standard deviations above the mean. Most x-scores are between -3 and 3.
7. The z-score for math is: $z_1 = \frac{740-528}{105} = 2.02$. The z-score for verbal is $z_2 = \frac{715-475}{98} = 2.45$. The student did better in verbal, since it has a larger z-score.

2.9 Assignment 2

Purposes

This assignment has two parts. The first part assesses your knowledge of choosing proper measures to describe the centre and spread (variation) of the distribution of the given data, calculating the mean, median, mode, standard deviation, range, and interquartile range of the given data if applicable, explaining the meaning of a z -score, calculating the quartiles and five-number summary of the data, and drawing and interpreting a box plot. The second part assesses your skills in using R commander to create a box plot and side-by-side box plot, and obtain descriptive measures of a given data set.

Resources

[M01_SaleHome.xlsx](#)

Instructions

Part A

Complete the following:

1. In one Winter Olympics, Michelle Kwan competed in the women's singles short program. From nine judges, she received the following scores in technical component ranging from 1 (poor) to 6 (perfect).

5.8 5.7 5.9 5.7 5.5 5.7 5.7 5.7 5.6 with $\sum_{i=1}^n x_i = 51.3$

- a. Find the mean, median, and mode of the data. (6 marks)
- b. Calculate the standard deviation of the data using the defining formula

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

Interpret the number obtained. (6 marks: 4+2)

- c. Calculate the standard deviation of the data using the computing formula

$$s = \sqrt{\frac{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}{n-1}}$$

Compare the result obtained in part (b). Which formula needs less calculation,

part (b) or part (c)? (6 marks)

2. A random sample of 21 patients yielded the following data on length of stay (in days) in a hospital.

4	4	12	18	9	6	12
3	6	15	7	3	55	1
10	12	5	7	1	12	9

- Obtain and interpret the five-number summary. (10 marks)
 - Obtain and interpret the interquartile range. (4 marks)
 - Is there any potential outlier? Justify by calculation. (3 marks)
 - Draw a box plot based on the data. What can you tell about the distribution of the data? (5 marks)
 - Choose the proper measures for the centre and spread (variation) of the data. Verify your choice. (3 marks)
3. Complete the following sentences.
- A standardized variable $Z = \frac{X - \mu}{\sigma}$ always has mean ____ and standard deviation _____. (2 marks)
 - A positive z -score indicates that the observation is ____ the mean; whereas a negative z -score indicates the observation is ____ the mean. (2 marks)
4. Suppose that you obtain 350 out of 400 in one exam whose mean score is 280 and the standard deviation is 20. Did you do well in the exam? Explain why. (3 marks)

Part B

Finish the following questions using R and R commander:

Read the data set “M01_SaleHome.xlsx” and use R commander to complete the following tasks. **For each, you need to copy or do a screenshot of the output in R commander (we later call it computer output) and paste it into the space below the questions.** To save space, you only need to copy and paste what is asked for in the questions, and sometime may need to shrink the size.

- Choose the most proper measure for the centre of each of the 10 variables and provide the values of the selected measures. (20 marks)
- Use proper numerical summaries to compare the prices of homes with a tile roof and a non-tile roof. Briefly explain your findings based on the numerical summaries. (5 marks)
- Draw a side-by-side box plot to compare the prices of homes with a tile roof and a

non-tile roof. Briefly explain your findings based on the graph. (5 marks)

4. Use proper descriptive statistics and graphs to compare the prices of homes with a swimming pool and without a swimming pool. Briefly summarize your findings. (10 marks)

Quiz 2



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2616#h5p-5>

CHAPTER 3: PROBABILITY CONCEPTS

Overview

In real life, people might be interested in the chance that a certain event happens. For example, when calculating your car insurance premium, the insurance company needs to estimate your chance of having a car accident. Given the fact that you just had a minor car accident, there is no doubt that your premium will be increased since the insurance company assumes your chance of having an accident has increased. This chapter introduces the basic concepts of probability that can be applied to calculate the chance of the occurrence of a certain event.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Identify the sample space of a chance experiment.
- Calculate probabilities using the equally likely outcome model (the $\frac{f}{N}$ rule) if applicable.
- Draw Venn diagrams to show relationship between events.
- Calculate probabilities of events using the addition, complementation, conditional, and multiplication rules.
- Show whether two events are independent by calculation.
- Calculate joint and marginal probabilities based on contingency tables.
- Use combinations and permutations to calculate the number of sample points of events.

3.1 Basic Concepts in Probability

Let us first introduce some basic concepts in probability theory.

- **Chance experiment** is a process producing outcomes that vary randomly when repeated.
- **Sample space**, denoted as S , is the collection of ALL possible outcomes of a chance experiment.
- Each possible outcome in the sample space is called a **sample point**.
- **An event** is a combination of sample points; it is a subset of the sample space. We use capital letters A, B, C, ..., E, ... to represent events.

Example: Basic Concepts

Table 3.1: Examples of Sample Space and Events

Chance Experiment	Sample Space S	Events
Flip a balanced coin	{H, T} where H: head, T: tail	E = observe a head = {H}
Roll a fair die	{1, 2, 3, 4, 5, 6}	E = observe a six = {6} A = observe even numbers = {2, 4, 6} B = outcome is less than 3 = {1, 2}
Flip a balanced coin twice	{HH, HT, TT, TH}	E = observe the same outcome = {HH, TT} A = observe at least one head = {HH, HT, TH}

For example, consider rolling a fair die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$ consists of six sample points, while the event observing even numbers $A = \{2, 4, 6\}$ contains three sample points, which are part of the sample space.



Exercise: Basic Concepts

Consider the chance experiment of rolling a fair die twice.

1. Identify the sample space S .
2. List all possible outcomes of the event that the two rolls give the same result.
3. List all possible outcomes of the event that at least one six is observed.

Show/Hide Answer

1. The sample space S contains 6×6 pairs in the form of $(1, 1), (1, 2), \dots, (1, 6), (2, 1), (2, 2), \dots, (2, 6), \dots, (6, 1), (6, 2), \dots, (6, 6)$.
2. $E = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$
3. $E = \{(1, 6), (2, 6), (3, 6), (4, 6), (5, 6), (6, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)\}$

3.2 Probability of An Event

Given an event E , the chance or the probability that the event E happens is denoted as $P(E)$. A probability near 0 indicates that the event is very unlikely to occur when the chance experiment is conducted, whereas a probability near 1 suggests that the event is very likely to occur.

3.2.1 Frequentist Interpretation of Probability

From the frequentist's point of view, the probability of an event can be interpreted as the proportion of times the event occurs in a large number of repetitions of the chance experiment. For instance, if we flip a balanced coin 100 times and observe 55 heads, then the probability of observing a head is

$$P(H) \approx \frac{\text{of times we observe a head}}{n} = \frac{55}{100} = 0.55.$$

The figure below shows that the proportion approaches to 0.5 as the number of experiments repetitions n increases. It is expected to observe heads half of the time, because the coin is balanced and there is a 50–50 chance heads are observed. Therefore, the proportion of observed heads will approach a constant when the coin is flipped infinite times, this constant is $P(H)$.

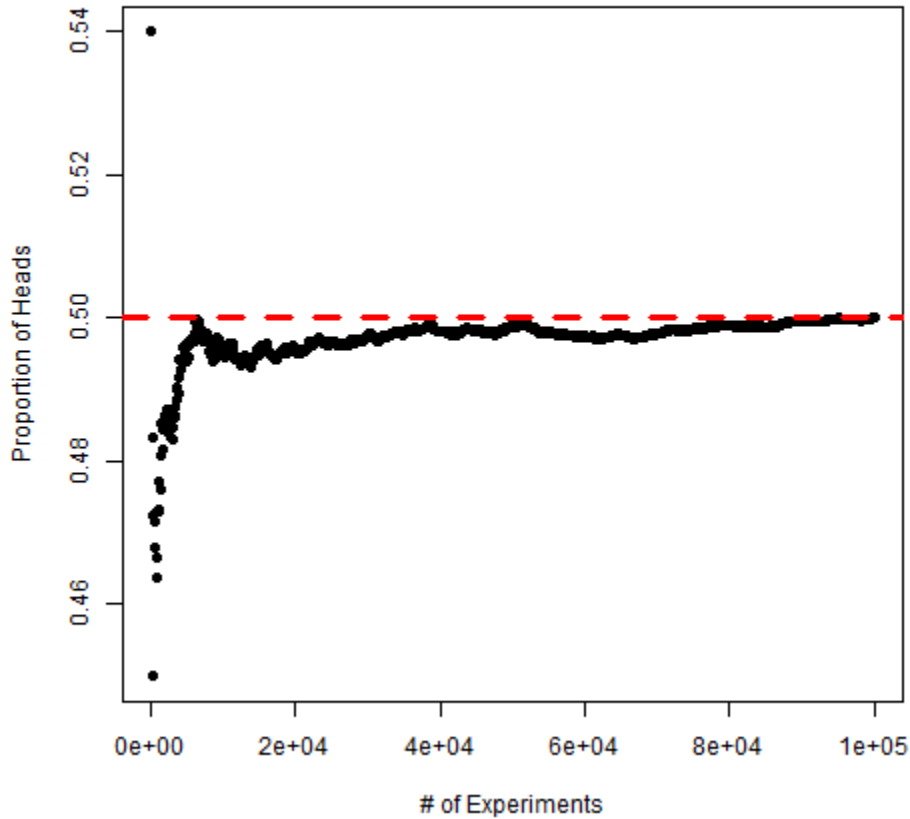


Figure 3.1: Proportion of Heads Converges to 0.5. [[Image Description \(See Appendix D Figure 3.1\)](#)]

Theoretically speaking, it is impossible to observe the probability of an event $P(E)$, because we won't repeat the chance experiment infinite times. However, we can sometimes calculate the probability based on a model or some probability rules.

3.2.2 Equally Likely Outcome Model, the f/N Rule

The simplest model, the equally likely outcome model, assumes that all possible outcomes have equal chance to be observed, such as flipping a **balanced** coin and rolling a **fair** die. For the equally likely outcome model, the probability that an event E happens is given by

$$\begin{aligned}
 P(E) &= \frac{\text{of sample points in event } E}{\text{of sample points in sample space } S} \\
 &= \frac{\text{of ways event } E \text{ can occur}}{\text{of possible outcomes}} \\
 &= \frac{f}{N}.
 \end{aligned}$$

- Recall that a standard die has sample space $S = \{1, 2, 3, 4, 5, 6\}$. Use the equally likely outcomes model to find the probability of the following events:

- Observing a six: $E = \{6\}$

$$P(E) = \frac{\text{of sample points in event } E}{\text{of sample points in sample space } S} = \frac{f}{N} = \frac{1}{6}.$$

- Observing an even number: $A = \{2, 4, 6\}$

$$P(A) = \frac{\text{of sample points in event } A}{\text{of sample points in space } S} = \frac{f}{N} = \frac{3}{6} = \frac{1}{2}.$$

- Observing an outcome that is less than 3: $B = \{1, 2\}$

$$P(B) = \frac{\text{of sample points in event } B}{\text{of sample points in space } S} = \frac{f}{N} = \frac{2}{6} = \frac{1}{3}.$$

- Suppose there are 100 students in a class, the following table summarizes the frequencies of number of siblings the students have:

Table 3.2: Frequency and Relative Frequency of # of Siblings

# of Siblings	Frequency	Relative Frequency
0	10	0.10
1	30	0.30
2	35	0.35
3	15	0.15
>3	10	0.10
Total	100	1.00

Find the probability that a randomly selected student has:

- exactly one sibling

If we randomly pick one student, each student has the same chance to be chosen; therefore, we can use the equally likely outcome model. There are $f = 30$ students with exactly one sibling. Hence

$$P(E) = \frac{\text{of sample points in event } E}{\text{of sample points in sample space } S} = \frac{f}{N} = \frac{30}{100} = 0.3.$$

- at least one sibling, which means one, two, three, or more than three siblings.

$$P(E) = \frac{\text{of sample points in event } E}{\text{of sample points in sample space } S} = \frac{f}{N} = \frac{30+35+15+10}{100} = 0.9.$$

- two to three siblings (inclusive), which means either two or three siblings.

$$P(E) = \frac{\text{of sample points in event } E}{\text{of sample points in sample space } S} = \frac{f}{N} = \frac{35+15}{100} = 0.5.$$

Key Facts: Basic Properties of the Probability of an Event

- The probability of an event $P(E)$ is always between 0 and 1, that is, $0 \leq P(E) \leq 1$.
- The probability of an event that can never occur is 0, e.g., $P(\text{observe a 7 when roll a regular die}) = 0$.
- The probability of an event that must occur is 1, e.g., $P(\text{observe a number smaller than 7 when roll a regular die}) = 1$.

All these properties can be easily shown by the f/N rule.

It is straightforward to show that using the f/N rule:

- $P(\emptyset) = 0$, where \emptyset is the empty set, a set with no element.
- $P(S) = 1$

3.3 Relationship Between Events and Venn Diagrams

Given events E , A , and B , we can construct new events using the complement, intersection and union operations:

- The **complement** of event E is the event in which E does not occur. The complement of E is denoted by E^c , or simply “not E ”.
- The **intersection** of events A and B is the event in which both A and B occur. The intersection of A and B is denoted by $A \cap B$, or $A \& B$, or simply “ A and B ”.
- The **union** of events A and B is the event in which we observe one of the following: A and B both occur, A occurs but B does not, or A does not occur but B does. The union of A and B is denoted by $A \cup B$ or simply “ A or B ”.

Two events are **mutually exclusive** if they do not overlap, i.e., they do not have outcomes in common. If two events A and B are mutually exclusive, they cannot occur at the same time; therefore $A \cap B = \emptyset$ and hence $P(A \cap B) = 0$.

Example: Relationship Between Events and Mutually Exclusive

Suppose we roll a fair die, so that the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Consider the following events:

- Observing an even number, $A = \{2, 4, 6\}$
- Observing an outcome that is at most 3, $B = \{1, 2, 3\}$
- Observing an outcome that is at least 5, $C = \{5, 6\}$

1. Define and list all possible outcomes of the following events:

- (not A): observing an odd number, not $A = \{1, 3, 5\}$
- ($A \& B$): observing an even number that is at most 3, $A \& B = \{2\}$. The only element in the overlap of events A and B , the common element in both the sets of A and B is 2.

2. Are the events A and B mutually exclusive?

No, the overlap is $\{2\}$, not an empty set. Therefore, events A and B are NOT mutually exclusive.

3. Are the events B and C mutually exclusive?

Yes, they don't overlap, i.e., there is no common element in both B and C. Therefore, events B and C are mutually exclusive, i.e., $B \cap C = \emptyset$.

We can use a **Venn diagram** to show relationships between events. In a Venn diagram, the sample space S is represented by a rectangle, events are often represented by circles, and events of interest are indicated by a shaded area. The following graphs show the Venn diagrams for the events E , (not E), ($A \& B$), and (A or B) respectively.

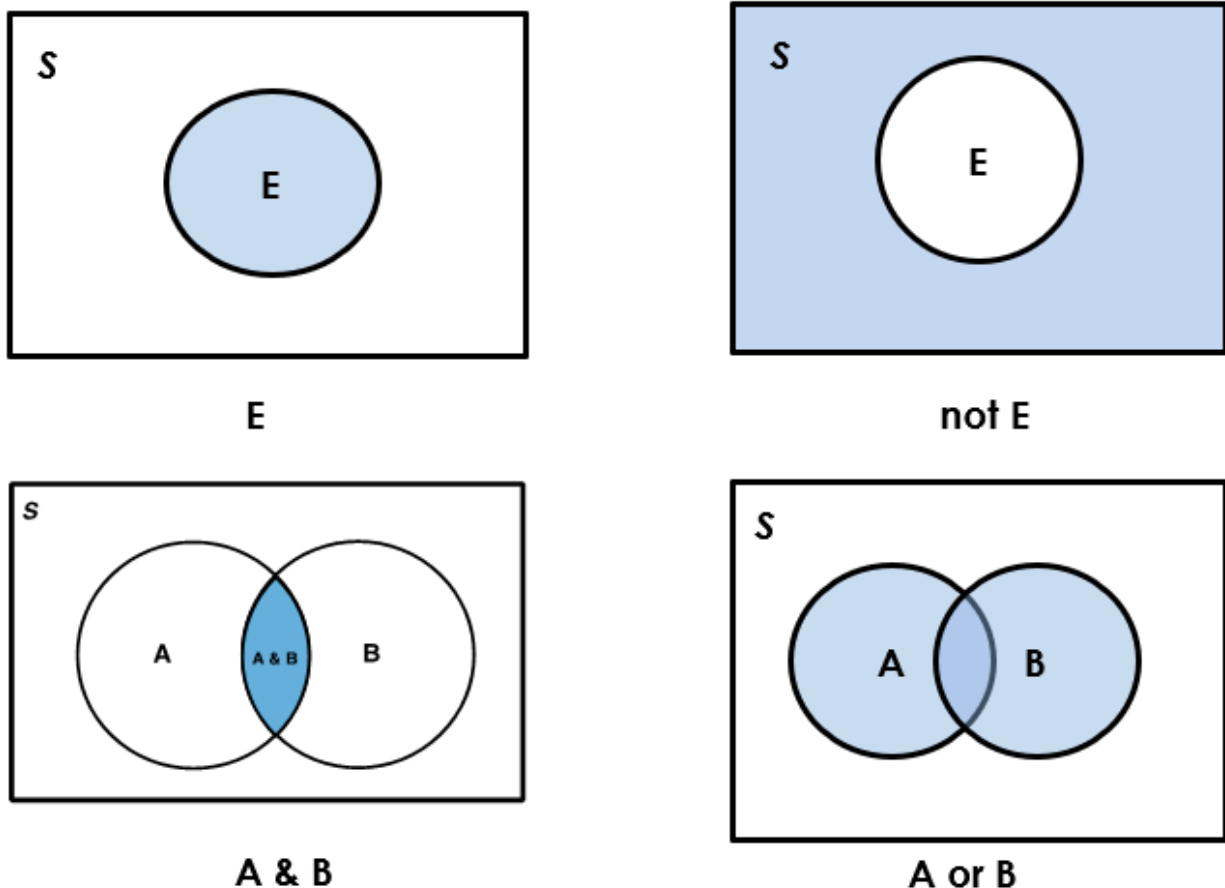


Figure 3.2 Venn Diagrams of Events [[Image Description \(See Appendix D Figure 3.2\)](#)]

It is straightforward to confirm the basic properties of the probability of an event based on the Venn diagrams:

- The total area of the rectangle is 1, which means $P(S) = 1$.
- The probability of the event E is the shaded area, between 0 and 1, which means $0 \leq P(E) \leq 1$ and $P(\emptyset) = 0$.

- If event A is a subset (part of) of event B (See Figure 3.3), denoted by $A \subseteq B$, then $P(A) \leq P(B)$. For example, observe in the above diagrams that $A \& B \subseteq A$ and $A \& B \subseteq B$. From this, it is easy to see that it is always true that $P(A \& B) \leq P(A)$ and that $P(A \& B) \leq P(B)$.

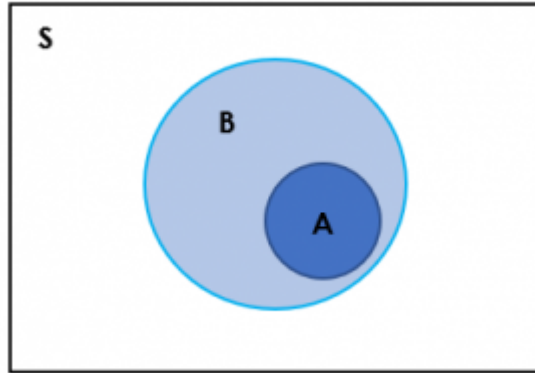


Figure 3.3 Event A is a Subset of Event B.
[[Image Description \(See Appendix D Figure 3.3\)](#)]

- For any two events A and B, since $A \subseteq (A \text{ or } B)$, we have $P(A) \leq P(A \text{ or } B)$. Similarly, $B \subseteq (A \text{ or } B)$, we have $P(B) \leq P(A \text{ or } B)$.

3.4 Probability Rules

The following are some basic rules of probability which can be easily verified with a Venn diagram:

- **Complement Rule:** $P(\text{not } E) = 1 - P(E)$ or $P(E) = 1 - P(\text{not } E)$.
- **General Addition Rule:** $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$.
- **Special Addition Rule:** when two events **A and B are mutually exclusive**,
 $P(A \text{ or } B) = P(A) + P(B)$.

This is due to the fact that $P(A \& B) = 0$, because it is impossible to observe both A and B.

More generally, if events A, B, C, \dots are mutually exclusive, then
 $P(A \text{ or } B \text{ or } C \text{ or } \dots) = P(A) + P(B) + P(C) + \dots$.

Examples: Probability Rules

Suppose we roll a fair die, so that the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Consider the following events:

- Observing a six, $E = \{6\}$.
- Observing an even number, $A = \{2, 4, 6\}$.
- Observing an outcome that is less than 3, $B = \{1, 2\}$.
- Observing an odd number, $C = \{1, 3, 5\}$.

Since the die is fair, each outcome is equally likely. Therefore we can use the f/N rule to find the probabilities.

$$\begin{aligned} P(E) &= \frac{f}{N} = \frac{1}{6}; & P(A) &= \frac{f}{N} = \frac{3}{6}, \\ P(B) &= \frac{f}{N} = \frac{2}{6}; & P(C) &= \frac{f}{N} = \frac{3}{6}. \end{aligned}$$

Find the probabilities of the following events:

1. (not E). Using the complement rule, $P(\text{not } E) = 1 - P(E) = 1 - \frac{1}{6} = \frac{5}{6}$.
2. (A & C). Events A and C are mutually exclusive, since it is impossible to observe a number that is both even and odd. Therefore, $A \cap C = \emptyset$ and $P(A \& C) = 0$.
3. (A or B). Since events A and B are NOT mutually exclusive, we use the general addition rule. The overlap of events A and B is {2}, that is, $A \cap B = \{2\}$ and therefore, $P(A \& B) = \frac{f}{N} = \frac{1}{6}$.
 By the general addition rule,
 $P(A \text{ or } B) = P(A) + P(B) - P(A \& B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}$.



Instructor's Note

Since we are rolling a fair die, we can also use the $\frac{f}{N}$ rule to find the probabilities in the previous example.

1. (not E) = not observing a six = {1, 2, 3, 4, 5}. $P(\text{not } E) = \frac{f}{N} = \frac{5}{6}$.
2. (A or B) = observing an even number or a number not more than 3 = {1, 2, 4, 6}
 $P(A \text{ or } B) = \frac{f}{N} = \frac{4}{6}$.

The results are identical to those using the probability rules.

3.5 Conditional Probability and Independence

Suppose we toss a fair die, and consider events $A = \{\text{observing an even number}\}$ and $B = \{\text{outcome} < 3\}$. If we know that event B occurs, will it affect the chance that event A occurs? If a mother has breast cancer, will it affect her daughter's probability of having breast cancer? More generally, if we acquire knowledge that some event has occurred, does it affect the probability of the occurrence of some other event? Such questions can often be addressed with conditional probabilities.

The probability that event A happens given that event B has occurred is called a **conditional probability**, denoted as $P(A|B)$, read as the conditional probability of A given B.

Two events A and B are **independent** if $P(A|B) = P(A)$, which means whether event B occurs or not won't affect the probability of A. Similarly, we have $P(B|A) = P(B)$, which means whether A occurs or not won't affect the probability of B.

In general, if A and B are any two events with $P(B) > 0$, then the conditional probability of A given B can be calculated as

$$P(A|B) = \frac{P(A \& B)}{P(B)}.$$

In the case of the breast cancer, the probability of a daughter having breast cancer given that her mother has breast cancer is the probability of both the daughter and mother having breast cancer divided by the probability that the mother has breast cancer.

Examples: Conditional Probability

Suppose we roll a fair die, so that the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Consider the following events:

- Observing an even number, $A = \{2, 4, 6\}$.
- Observing an outcome that is less than 3, $B = \{1, 2\}$.
- Observing an odd number, $C = \{1, 3, 5\}$.

1. Find the conditional probability $P(A|B)$.

- Method 1: $\frac{f}{N}$ rule

Given that B has occurred, the sample space reduces from $S = \{1, 2, 3, 4, 5, 6\}$ to $B = \{1, 2\}$. $A \cap B =$ even numbers less than 3 = $\{2\}$. Therefore, $P(A|B) = \frac{f}{N} = \frac{1}{2} = 0.5$.

- Method 2: by the conditional probability rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{2/6} = \frac{1}{2} = 0.5.$$

2. Are the events A and B independent?

By the f/N rule, $P(A) = \frac{3}{6} = 0.5$, since $P(A|B) = P(A)$, events A and B are independent.

3. Find the conditional probability $P(A|C)$.

Since events A (observing an even number) and C (observing an odd number) are mutually exclusive,

$$P(A \cap C) = 0 \implies P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{0}{P(C)} = 0.$$

4. Are the events A and C independent?

Since $P(A|C) = 0 \neq 0.5 = P(A)$, i.e., $P(A|C) \neq P(A)$, events A and C are dependent. Note that events A and C are mutually exclusive and hence must be dependent, since the occurrence of one event eliminates the chance of the other.

One common difficulty that students have with conditional probability problems is translating the text of a “word problem” into the correct probability statements. The following example illustrates different ways of writing a conditional probability.

Example: Different Ways of Writing a Conditional Probability

Suppose we roll a fair die, so that the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Consider the following events:

- Observing an even number, $A = \{2, 4, 6\}$.
- Observing an outcome that is less than 3, $B = \{1, 2\}$.

We calculated the conditional probability $P(A|B) = 0.5$ in the previous example. There are all ways we can write the conditional probability $P(A|B)$:

- Given that the outcome is less than 3, the probability of observing an even number is 0.5.
- If the outcome is less than 3, then we have a probability of 0.5 of observing an even number.
- 50% of outcomes that are less than 3 are even numbers.
- If the outcome is less than 3, there is a 0.5 probability of observing an even number.

Multiplying $P(B)$ on both sides of the conditional probability equation, we obtain the

general multiplication rule of probability:
 $P(A \& B) = P(A) \times P(B|A)$ or $P(A \& B) = P(B) \times P(A|B)$.

The proof is as follows:

$$\begin{aligned} P(A|B) &= \frac{P(A \& B)}{P(B)} \implies P(B) \times P(A|B) = P(B) \times \frac{P(A \& B)}{P(B)} \\ &\implies P(A \& B) = P(B) \times P(A|B). \end{aligned}$$

If two events A and B are **independent**, i.e., $P(A|B) = P(A)$, the general multiplication rule becomes the **special multiplication rule**:

$$P(A \& B) = P(B) \times P(A|B) = P(B) \times P(A).$$

More generally, if events A, B, C are **independent**, then

$$P(A \& B \& C \& \dots) = P(A) \times P(B) \times P(C) \times \dots$$

Key Facts: Check Whether Two Events are Independent

Two events A and B are independent, if and only if **ANY** of the following holds:

- $P(A|B) = P(A)$ or
- $P(B|A) = P(B)$ or
- $P(A \& B) = P(A) \times P(B)$

The first two equations are due to the definition of independence, and the last one is due to the special multiplication rule.



Instructor's Note

1. Any of the above equations can be used to check whether two events A and B are independent. **DO NOT** use all three, pick the one and the only one that is the easiest to calculate and check.
2. Mutually exclusive and independent are two different concepts. Mutually exclusive events are those that cannot occur simultaneously; independent events are those for which the occurrence of one does not affect the probabilities of the others. Actually, mutually exclusive events must be dependent, since the occurrence of one event eliminates the chance of all the others.

3.6 Summary of Probability Rules

We have learned the following probability rules so far:

- **Complement Rule:** $P(\text{not } E) = 1 - P(E)$ or $P(E) = 1 - P(\text{not } E)$.
- **General Addition Rule:** $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$.
- **Special Addition Rule:** $P(A \text{ or } B) = P(A) + P(B)$ if events A and B are **mutually exclusive**.
- **Conditional Probability Rule:** $P(A|B) = \frac{P(A \& B)}{P(B)}$ for $P(B) > 0$.
- **General Multiplication Rule:** $P(A \& B) = P(B) \times P(A|B)$ or $P(A \& B) = P(A) \times P(B|A)$.
- **Special Multiplication Rule:** $P(A \& B) = P(A) \times P(B)$ if events A and B are **independent**.



Activity

Exercise: Application of Probability Rules

It is believed that there is an association between breast cancer and smoking. The following table summarizes results of an observational study of 200 females who are classified by their disease status and smoking status.

	Smoker (S)	Non-smoker (not S)	Total
Breast Cancer (B)	10 (B & S)	30 (B & not S)	40 (B)
Cancer Free (not B)	20 (not B & S)	140 (not B & not S)	160 (not B)
Total	30 (S)	170 (not S)	200

1. What is the probability that a randomly selected female suffers from breast cancer?
2. What is the probability that a randomly selected female is a smoker?
3. What is the probability that a randomly selected female has breast cancer and she is a smoker?
4. What is the probability that a randomly selected female has breast cancer given that she is a smoker?
5. Are the events "Breast Cancer" and "Smoker" independent? Explain your answer by calculation.
6. Interpret the conditional probability calculated in part (4) in several ways.

Show/Hide Answer

1. What is the probability that a randomly selected female suffers from breast cancer?

It is very important to define the events in order to apply the probability rules. Let B = the event of suffering breast cancer, and S = the event of being a smoker. For this exercise, we want to compute $P(B)$. Since we randomly pick one female and each female has the same chance of being selected, we can use the f/N rule.

$$P(B) = \frac{f}{N} = \frac{40}{200} = 0.2.$$

2. What is the probability that a randomly selected female is a smoker?

$$P(S) = \frac{f}{N} = \frac{30}{200} = 0.15.$$

3. What is the probability that a randomly selected female has breast cancer and she is a smoker?

$$P(B \& S) = \frac{f}{N} = \frac{10}{200} = 0.05.$$

4. What is the probability that a randomly selected female has breast cancer given that she is a smoker?

We want to compute $P(B|S)$. By the conditional probability rule

$$P(B|S) = \frac{P(B \& S)}{P(S)} = \frac{10/200}{30/200} = \frac{1}{3} = 0.333.$$

5. Are the events “Breast Cancer” and “Smoker” independent? Explain your answer by calculation. No, since $P(B|S) \neq P(B)$. The conditional probability of breast cancer given smoker is $P(B|S) = 0.333$ and the unconditional probability of breast cancer $P(B) = 0.2$, which implies there is an association between smoking and higher rates of breast cancer.

We can also check whether $P(B \& S) = P(B) \times P(S)$. Since

$$P(B \& S) = \frac{10}{200} \neq \left(\frac{40}{200}\right)\left(\frac{30}{200}\right) = P(B) \times P(S), \text{ the two events are NOT independent.}$$

6. Interpret the conditional probability calculated in part (4) in several ways.

We got $P(B|S) = 0.333$ in part (4), and this conditional probability can be interpreted in the following ways:

- If a randomly selected woman is a smoker, her probability of having breast cancer is 0.333.
- Given that a randomly selected woman smokes, she has a probability of 0.333 of having breast cancer.
- If you are a woman who smokes, then you have a probability of 0.333 of having breast cancer.
- 33.3% of women who smoke have breast cancer.

3.7 Tree Diagrams

A **tree diagram** is a helpful tool, used to display a sequence of events and their conditional probabilities. Each branch of the tree corresponds to one possible outcome in the sequence of the events; and the probability of the outcome equals the product of all subsequent probabilities on the branch. Tree diagrams provide a useful illustration of the general multiplication rule.

Example: Tree Diagram

Suppose we have two midterms, the probability that you get at least 90 in Midterm I is 0.15. If you get at least 90 in Midterm I, then the probability that you get at least 90 in Midterm II is 0.8; if you do not receive at least 90 in Midterm I, the probability that you will receive at least 90 in Midterm II is 0.1. We could use a tree diagram to present the probabilities.

Let's define the events:

A_1 = obtaining at least 90 in Midterm I, B_1 = obtaining below 90 in Midterm I

A_2 = obtaining at least 90 in Midterm II, B_2 = obtaining below 90 in Midterm II

We can summarize the information given in the question using probability notations.

$$P(A_1) = 0.15; \quad P(B_1) = 1 - P(A_1) = 1 - 0.15 = 0.85;$$

$$P(A_2|A_1) = 0.8; \quad P(A_2|B_1) = 0.1;$$

$$P(B_2|A_1) = 1 - 0.8 = 0.2; \quad P(B_2|B_1) = 1 - 0.1 = 0.9.$$

Then the tree diagram is presented as follows:

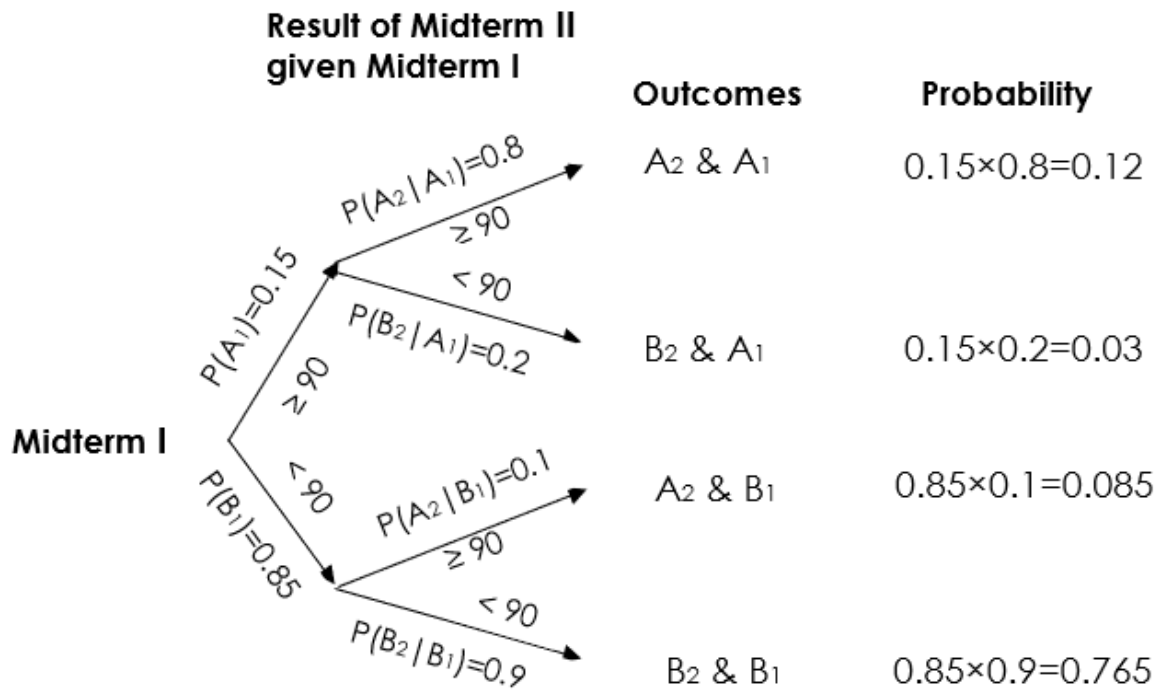


Figure 3.4: Tree Diagram of Outcomes of Two Midterm Exams. [Image Description (See Appendix D Figure 3.4)]

- Find the probability that a student obtained at least 90 in both midterms.

$$P(A_1 \& A_2) = P(A_1) \times P(A_2|A_1) = 0.15 \times 0.8 = 0.12.$$

Note that due to the multiplication rule, the probability of each outcome equals the product of all subsequent probabilities on the corresponding branch.

- Find the probability that a student obtained at least 90 in exactly one midterm.

$$\begin{aligned} P(A_1 \& B_2 \text{ or } B_1 \& A_2) &= P(A_1 \& B_2) + P(B_1 \& A_2) \\ &= P(A_1) \times P(B_2|A_1) + P(B_1) \times P(A_2|B_1) \\ &= 0.15 \times 0.2 + 0.85 \times 0.1 = 0.03 + 0.085 = 0.115. \end{aligned}$$

- Find the probability that a student obtained at least 90 in at least one of the midterms.

$$\begin{aligned} P(A_1 \& A_2 \text{ or } A_1 \& B_2 \text{ or } A_2 \& B_1) &= 1 - P(B_1 \& B_2) = 1 - P(B_1) \times P(B_2|B_1) \\ &= 1 - 0.85 \times 0.9 = 1 - 0.765 = 0.235, \end{aligned}$$

OR

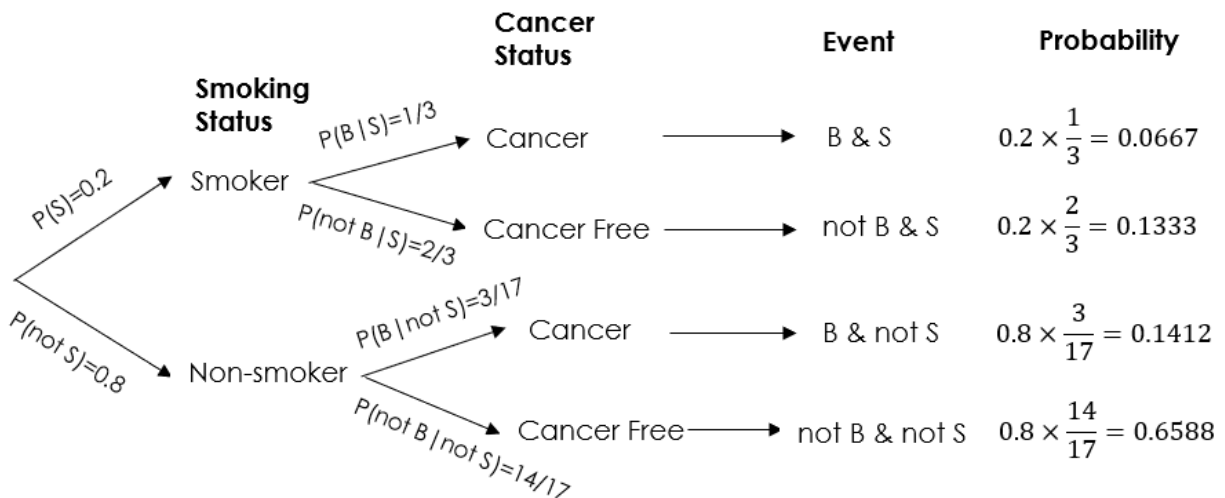
$$\begin{aligned} P(A_1 \& A_2 \text{ or } A_1 \& B_2 \text{ or } B_1 \& A_2) &= P(A_1 \& A_2) + P(A_1 \& B_2) + P(B_1 \& A_2) \\ &= 0.12 + 0.03 + 0.085 = 0.235. \end{aligned}$$

Exercise: Tree Diagram

It is believed that there is an association between breast cancer and smoking. The following table summarizes results of an observational study of 200 females who are classified by their disease status and smoking status.

	Smoker (S)	Non-smoker (not S)	Total
Breast Cancer (B)	10 (B & S)	30 (B & not S)	40 (B)
Cancer Free (not B)	20 (not B & S)	140 (not B & not S)	160 (not B)
Total	30 (S)	170 (not S)	200

Represent the information given in the contingency table above in a tree diagram, branching first on smoking status and then on breast cancer status.



[[Image Description \(See Appendix D Example 3.1\)](#)]

3.8 Counting Rules: Basic Counting Rule, Combination, and Permutation

In order to apply the equal-likely outcome model (the f/N rule) to calculate the probability of a certain event, we need to determine N (the number of all possible outcomes) and f (the number of ways we observe the event). However, if f and N are very large numbers, it can be difficult, or even impossible, to list all possible outcomes that they contain. Fortunately, we can use counting rules to determine the number of ways that something can happen without directly listing all possibilities.

3.8.1 The Basic Counting Rule

Suppose that a job consists of k separate tasks and the i th task can be done in n_i ways, $i = 1, 2, \dots, k$, the basic counting rule states that the entire job can be done in $n_1 \times n_2 \times \dots \times n_k$ different ways.

Example: Basic Counting Rule

1. An experiment involves tossing a pair of dice and observing the number on the upper faces. Find the number of possible outcomes of this experiment.

The experiment consists of two tasks: observing the outcome of the first die and observing the second die. The number of possible outcomes of the first task is 6 and the number of possible outcomes of the second task is also 6; as a result, the total number of possible outcomes is $6 \times 6 = 36$.

2. A restaurant offers a four-course meal consisting of soup, salad, a main course and a dessert. The menu provides three options for soup, two options for salad, four options for the main course, and three options for dessert. Find the total number of different orders for this four-course meal.

By the basic counting rule, number of orders is $3 \times 2 \times 4 \times 3 = 72$.

3. Determine the number of ways to arrange three objects in order.

Arranging three objects in order consists of three tasks:

- 1) choose the object to be placed in the first spot,
- 2) choose the object to be placed in the second spot, and
- 3) choose the object to be placed in the third spot.

There are three objects; therefore, there are three options for the first spot, two for the second spot and only one for the last spot. By the basic counting rule, there are $3 \times 2 \times 1 = 3! = 6$ ways to arrange three objects in order. More generally, there are $n \times (n - 1) \times \cdots \times 1 = n!$ ways to arrange n objects in order. The notation $n!$ is read as **n factorial**. Note that $0! = 1$.

3.8.2 Combination: Order Does Not Matter

A **combination** of r objects from a collection of n objects is any **unordered** arrangement of r of the n objects—in other words, any subset of r objects from the collection of n objects. The number of possible combinations of r objects that can be formed from n objects is denoted as ${}_nC_r$ (read as “ n choose r ”) and can be calculated as ${}_nC_r = \frac{n!}{r!(n-r)!}$.

For example, if a hand of 5 cards is dealt from a deck of 52 cards, the order does not matter. There are ${}_{52}C_5 = \frac{52!}{5!(52-5)!} = \frac{52!}{5!47!} = 2,598,960$ different hands.

Here is an intuitive argument for the formula. The n objects are divided into two parts: those r objects that are chosen and those $(n - r)$ objects that are unselected. The number of unordered arrangements of r objects taken from n objects equals to the number of ordered arrangements of n objects divided by the number of ways that we could arrange the r selected objects and those $(n - r)$ unselected objects in order. There are $n!$ ways to arrange n objects in order, $r!$ ways to arrange those r selected objects and $(n - r)!$ ways to arrange those $(n - r)$ unselected objects in order. By the basic counting rule, the number of ways to arrange those r selected and $(n - r)$ unselected objects is $r!(n - r)!$. Therefore, the number of unordered arrangements of r objects taken from n objects is given by ${}_nC_r = \frac{n!}{r!(n-r)!}$.

3.8.3 Permutation: Order Does Matter

A **permutation** of r objects from a collection of n objects is any **ordered** arrangement of r of the n objects. The number of possible permutations of r objects that can be formed from n objects is denoted as ${}_nP_r$ (read as “ n permute r ”) and can be calculated as ${}_nP_r = \frac{n!}{(n-r)!}$.

For example, if a passcode consists of four different digits from 0 to 9, order of the four digits matters. There are ${}_{10}P_4 = \frac{10!}{(10-4)!} = 10 \times 9 \times 8 \times 7 = 5040$ distinguish passcodes.

We can consider the permutation as a two-step procedure: we first choose r objects of n objects and then arrange them in order. There are ${}_nC_r$ ways to choose r objects of n objects and there are $r!$ different ways to arrange r objects in order. By the basic counting rule, ${}_nP_r = {}_nC_r \times r! = \frac{n!}{r!(n-r)!} \times r! = \frac{n!}{(n-r)!}$.

Alternatively, we can view the permutation as a truncated version of the factorial function. That is, we arrange r of the n objects in order: there are n ways to pick the first object, $n - 1$ ways to pick the second object, ..., and $n - r + 1$ ways to pick the r th object. Hence, by the basic counting rule, ${}_nP_r = n \times (n - 1) \times \cdots \times (n - r + 1)$.

Multiplying by $\frac{(n-r)!}{(n-r)!}$ gives ${}_nP_r = \frac{n!}{(n-r)!}$.

Exercise: Combinations and Permutations

1. Suppose a population consists of $N = 3$ students (A, B, and C). If we take a simple random sample of size $n = 2$, find the number of different samples.
2. Suppose three students run for two positions: president and vice-president. Find the number of possible outcomes.
3. Lotto 649 is a lottery that costs \$3 to play and requires you to pick 6 numbers between 1 and 49, without replacement. You win the jackpot if your six numbers match the six winning numbers, and order does not matter. According to national-lottery.com, the average Jackpot between January 1, 2020 and May 20, 2020 was about \$9,000,000. Find the probability of winning a jackpot. Do you think it is wise to play Lotto 649?

Show/Hide Answer

1. Steps:
 - Does order matter? (No, just want two students, no ordered arrangement is needed)
 - Use combination:

Sample (n=2)

A, B

A, C

B, C

$${}_3C_2 = \frac{3!}{2!(3-2)!} = \frac{3!}{2!1!} = \frac{3 \times 2 \times 1}{(2 \times 1)(1)} = 3.$$

Note that the listing of all events in the table above is not necessary to solve the problem. It is

given to demonstrate that we get the same answer by using the formula. By using the counting rules (basic counting rule, combination and permutation), we are able to obtain the number of sample points of an event without the need of listing all possible outcomes.

2. Steps:

- Does order matter? (Yes, need to choose two students and then assign them to president or vice president.)
- Use permutation: ${}_3P_2 = {}_3C_2 \times 2! = 3 \times 2 = 6$.

President	Vice President
A	B
B	A
A	C
C	A
B	C
C	B

3. Steps:

- Does order matter? (No, only need to choose six numbers between 1 and 49, order does not matter)
- Use permutation: ${}_{49}C_6 = 13,983,816$.

Choose six numbers out of 49, there are ${}_{49}C_6 = 13,983,816$ combinations, but only one combination will match those six winning numbers. Therefore, the probability of winning jackpot is

$$P(\text{Jackpot}) = \frac{f}{N} = \frac{1}{{}_{49}C_6} = \frac{1}{13,983,816} = 0.0000000715.$$

One Lotto 649 ticket costs \$3. If you try all ${}_{49}C_6 = 13,983,816$ possible combinations at a cost of

$$13,983,816 \times 3 = \$41,951,448.$$

That is, in order to guarantee winning the Jackpot, you need to spend more than \$30,000,000 more than the average winnings in the first 4 months of 2020. This is clearly not a good way to spend your money!

Exercise: Probability Rules and Counting Rules

1. Roll three balanced dice,

- find the probability of rolling all 6s.
- find the probability that all the dice come up the same number.

2. Suppose that STAT 151 has four sections, and that three students each randomly pick a section. Find the probability that

- a) all three students end up in the same section.
- b) all three students end up in different sections.
- c) nobody picks the first section.

3. Five men and three women sit together in a row. Find the probability that

- a) the same gender is at each end.
- b) the three women all sit together.

Show/Hide Answer

1.

- a. The three dice are independent and each has a $\frac{1}{6}$ probability of rolling a six. By the special multiplication rule $P(E) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}$.
- b. Using the same reasoning as in part (a), we conclude that each of the six faces has a $(\frac{1}{6})^3$ probability of occurring 3 consecutive times. Since there are 6 possible outcomes, it follows that $P(E) = 6 \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}$.

2.

- a. $\frac{f}{N} = \frac{4 \times 1 \times 1}{4 \times 4 \times 4}$. Each student has four choices, so $N = 4 \times 4 \times 4$. In order for all three students to end up in the same section, the first student has four choices, while the second and third students have to pick the same section as the first student. Hence, $f = 4 \times 1 \times 1$.
- b. $\frac{f}{N} = \frac{4 \times 3 \times 2}{4 \times 4 \times 4}$. In order for all three students to end up in different sections, the first student has four choices, the second student has three choices and third student has two choices. Hence, $f = 4 \times 3 \times 2$.
- c. $\frac{f}{N} = \frac{3 \times 3 \times 3}{4 \times 4 \times 4}$. Since the only condition is that no student can choose the first section, it follows that everyone has three choices, and so $f = 3 \times 3 \times 3$.

3.

- a. $\frac{f}{N} = \frac{\binom{5}{2}2!6! + \binom{3}{2}2!6!}{8!}$. Note that the notation $\binom{3}{2}$ is the same as ${}_3C_2$. Eight people can be ordered in $N = 8!$ ways. In order to count the number of ways for the individuals at either end to be of the same gender, we consider two cases.
 - Case 1: men on each end. First, there are $\binom{5}{2}$ (five choose 2) ways choose two of the five men to sit at either end. Next, there are $2!$ ways to arrange these two men in order. Finally, the remaining six people (three men and three women) can be arranged in $6!$ ways. By the basic counting rule, there are $\binom{5}{2}2!6!$ different arrangements with a man on each end.
 - Case 2: women on each end. First, there are $\binom{3}{2}$ (three choose 2) ways to choose two of the three women to sit at either end. Next, there are $2!$ ways to arrange these two women in order. Finally, the remaining six people (five men and one woman) can be arranged in $6!$ ways. By the basic counting rule, there are $\binom{3}{2}2!6!$ different arrangements with a woman on each end.

Combining cases 1 and 2, we see that there is a total of $\binom{5}{2}2!6! + \binom{3}{2}2!6!$ different arrangements with individuals of the same gender seated on each end.

b. $\frac{f}{N} = \frac{3!6!}{8!}$. Once again, eight people can be ordered in $N = 8!$ ways. To determine the number of reorderings with all three women together, it is easiest to view the three women as one item in an ordered arrangement with five men: $(\{W,W,W\}, M, M, M, M, M)$. Notice that six items can be reordered in $6!$ ways, and that the 3 women within the first item can be reordered in $3!$ ways. Therefore $f = 3!6!$.

3.9 Contingency Table: Joint and Marginal Probability

Recall the contingency table in the example of association between breast cancer and smoking:

Table 3.3: Contingency Table of “Cancer Status” and “Smoking Status”

	Smoker (S)	Non-smoker (not S)	Total
Breast Cancer (B)	10 (B & S)	30 (B & not S)	40 (B)
Cancer Free (not B)	20 (not B & S)	140 (not B & not S)	160 (not B)
Total	30 (S)	170 (not S)	200

The row variable is “Cancer Status” with two possible values: breast cancer or cancer free. The column variable is “Smoking Status” with two possible values: smoker and non-smoker.

The **marginal probabilities** are the row or column totals divided by the grand total. For the current example, the marginal probabilities are:

$$P(B) = \frac{40}{200} = 0.2, \quad P(\text{not } B) = \frac{160}{200} = 0.8;$$

$$P(S) = \frac{30}{200} = 0.15, \quad P(\text{not } S) = \frac{170}{200} = 0.85.$$

Note that $P(B) = 0.2$ and $P(\text{not } B) = 0.8$ give the marginal probability distribution of the row variable “Cancer Status” and they add up to 1. Similarly, $P(S) = 0.15$ and $P(\text{not } S) = 0.85$ give the marginal probability distribution of the column variable “Smoking Status” and they sum to 1.

The **joint probabilities** are the frequencies in the cells divided by the grand total. For the current example, the joint probabilities are:

$$P(B \& S) = \frac{10}{200} = 0.05, \quad P(B \& \text{not } S) = \frac{30}{200} = 0.15,$$

$$P(\text{not } B \& S) = \frac{20}{200} = 0.1, \quad P(\text{not } B \& \text{not } S) = \frac{140}{200} = 0.7.$$

3.10 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Identify the sample space of a chance experiment (Section 3.1).
- Calculate probabilities using the equally likely outcome model (the $\frac{f}{N}$ rule) if applicable (Section 3.2).
- Draw Venn diagrams to show relationship between events (Section 3.3).
- Calculate probabilities of events using the addition, complementation, conditional, and multiplication rules (Sections 3.4, 3.5 and 3.6).
- Show whether two events are independent by calculation (Section 3.5).
- Calculate joint and marginal probabilities based on contingency tables (Section 3.9).
- Use combinations and permutations to calculate the number of sample points of events (Section 3.8).

3.11 Review Questions

1. Roll five balanced dice,
 - a. find the probability of rolling all 1s. Let A be the event that the outcomes are all 1s.
 - b. find the probability that all the dice come up the same number.
 - c. Are the two events in parts (a) and (b) independent?
 - d. Are the two events in parts (a) and (b) mutually exclusive?
2. If events A and B are mutually exclusive, $P(A) = 0.25$, $P(B) = 0.4$. Find the probability of each of the following events: not A , $(A \& B)$, and $(A \text{ or } B)$.
3. A survey on 1000 employees about their gender and marital status gives the following data: 813 employees are male, 875 are married, and 572 married men. Is there anything wrong with these data? Explain why.
4. Suppose that STAT 151 has 8 sections, 3 students each randomly pick a section. Find the probability that
 - a. they end up in the same section.
 - b. they are all in different sections.
 - c. nobody picks section 1.
5. There are 10 students in our class. Assume that every student is equally likely to be born on any of the 365 days in a year. Find the probability that no two students in the class have the same birthday.

Show/Hide Answer

1.

a.

$$\begin{aligned} P(\text{all 1s}) &= P(1 \text{ for 1st die} \& 1 \text{ for 2nd die} \& 1 \text{ for 3rd die} \& 1 \text{ for 4th die} \& 1 \text{ for 5th die}) \\ &= P(1 \text{ for 1st die}) \times P(1 \text{ for 2nd die}) \times P(1 \text{ for 3rd die}) \times P(1 \text{ for 4th die}) \\ &\quad \times P(1 \text{ for 5th die}) \\ &= \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \left(\frac{1}{6}\right)^5 = 0.000128. \end{aligned}$$

b.

$$\begin{aligned} P(\text{same number}) &= P(\text{all 1s or all 2s or all 3s or all 4s or all 5s or all 6s}) \\ &= P(\text{all 1s}) + P(\text{all 2s}) + P(\text{all 3s}) + P(\text{all 4s}) + P(\text{all 5s}) + P(\text{all 6s}) \\ &= 6 \times \left(\frac{1}{6}\right)^5 = \left(\frac{1}{6}\right)^4 = 0.0007716. \end{aligned}$$

c. Events A and B are not independent, since $P(A \& B) = P(A) \neq P(A) \times P(B)$.

d. Events A and B are not mutually exclusive, since the overlap is event A which is

NOT an empty set.

2.

$$P(\text{not } A) = 1 - P(A) = 1 - 0.25 = 0.75$$

$$P(A \text{ and } B) = 0 \quad (\text{mutually exclusive})$$

$$P(A \text{ or } B) = P(A) + P(B) = 0.25 + 0.4 = 0.65.$$

3. There are $1000 - 813 = 187$ females. However, the number of married females is $875 - 572 = 303$ which is larger than 187 and this is impossible.

4.

a. Apply the basic counting rule:

$$\frac{8 \times 1 \times 1}{8 \times 8 \times 8} = \frac{1}{64} = 0.015625.$$

b. Apply the basic counting rule:

$$\frac{8 \times 7 \times 6}{8 \times 8 \times 8} = \frac{42}{64} = 0.65625.$$

c. Apply the basic counting rule:

$$\frac{7 \times 7 \times 7}{8 \times 8 \times 8} = \left(\frac{7}{8}\right)^3 = 0.6699.$$

5.

$$\frac{\binom{365}{10} \times 10!}{365^{10}} = \frac{365 \times 364 \times \cdots \times 356}{365^{10}} = 0.88305.$$

3.12 Assignment 3

Purposes

The following questions assess your knowledge of identifying the sample space of a chance experiment, calculating probabilities using equally likely outcome model (the f/N rule) if applicable, calculating probabilities of events using addition, complementation, conditional, multiplication rules, showing whether two events are independent by calculation, and using combination and permutation to calculate the number of sample points of events.

Resources

[M03_SaleHome_Recode.xlsx](#)

Instructions

Complete the following:

1. Let H be the event of observing a head and T be the event of observing a tail. A balanced coin is tossed three times.
 - a. List all possible outcomes. (2 marks)
 - b. List all possible outcomes and find the probabilities of the following events. (8 marks)
 1. A = event exactly two heads are tossed
 2. B = event the first toss is a tail
 3. C = event the first toss is a head
 4. D = event all three tosses come up the same
 - c. List all possible outcomes and find the probabilities of the following events. (10 marks: 2 for each)
 1. not A
 2. A & B
 3. B & C

4. C or D
 5. D|A
 - d. Identify all possible pairs of events defined in part (b) that are mutually exclusive. (3 marks)
 - e. Are the events A and D independent? Explain your answer mathematically. (3 marks)
2. A contingency table for injuries in the United States by circumstance (column variable) and gender (row variable) is given as follows. Note that frequencies are in millions.

	Work (C_1)	Home (C_2)	Others (C_3)	Total
Male (R_1)	8.0	9.8	17.8	?
Female (R_2)	1.3	?	12.9	25.8
Total	9.3	?	30.7	61.4

- a. Complete the contingency table. (3 marks)
 - b. Find the probability that an injured person was hurt at work, that is, $P(C_1)$. (2 marks)
 - c. Find the probability that an injured person was hurt at work and she was a female, that is, $P(C_1 \& R_2)$. (2 marks)
 - d. Find the probability that a female injured person was hurt at work, that is, $P(C_1|R_2)$. (3 marks)
 - e. Are events C_1 and R_2 independent? Explain your answer. (2 marks)
 - f. Are events C_1 and R_2 mutually exclusive? Explain your answer. (2 marks)
 - g. Is the event that an injured person is male independent of the event that an injured person was hurt at home? Explain by calculation. (4 marks)
2. Consider a population consisting of 60 students.
- a. How many samples of size 5 are possible? (2 marks)
 - b. What is the probability of taking each sample? (2 marks)
3. The U.S. Senate consists of 100 senators, two from each state. A committee consisting of five senators is to be formed.
- a. How many different committees are possible? (3 marks)
 - b. How many different committees are possible if no state may have more than one senator on the committee? (4 marks)
4. Suppose that a rare disease occurs in the general population in only one of every 10,000 people. A medical test is used to detect the disease. If a person has the disease, the probability that the test result is positive is 0.99. If a person does not have the disease, the probability that the test result is positive is 0.02. Given that a person's test result is positive, find the probability that this person truly has the rare disease. (Bonus:

5 marks)

Part B

Finish the following questions using R and R commander:

Read the data set “**M03_SaleHome_Recode.xlsx**” and use R commander to complete the following tasks. **For each, you need to copy or do a screenshot of the output in R commander (we later call it computer output) and paste it into the space below the questions.** To save space, you only need to copy and paste what is asked for in the questions and you can shrink the size of the image if necessary.

Use R commander to obtain a proper frequency distribution or contingency table to answer the following questions.

1. If we randomly select a sale home, what is the probability that it is a small house? (4 marks)
2. If we randomly select a sale home, what is the probability that it is a small house with a swimming pool? (4 marks)
3. If we randomly select a sale home, what is the probability that it is a small house given that it has a swimming pool? (4 marks)
4. If we randomly select a sale home that has a tiled roof, what is the probability that it is a small house with a swimming pool? (4 marks)
5. If we randomly select a small house, what is the probability that this house has a tiled roof given that it does not have a swimming pool? (4 marks)

Quiz 3



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2619#h5p-3>

CHAPTER 4: DISCRETE RANDOM VARIABLES

Overview

This chapter introduces the basic concepts of discrete random variables including their probability distributions, means, and standard deviations. We describe in detail one of the most important discrete probability distribution: the binomial distribution.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Determine the probability distribution of a discrete random variable.
- Calculate the mean (expected value) and standard deviation of a discrete random variable based on its probability distribution.
- Identify situations where the binomial distribution can be applied.
- Calculate the probabilities of events using the binomial probability formula, if applicable.
- Calculate the mean and standard deviation of a random variable that follows the binomial distribution.

4.1 Random Variable

In this section we introduce discrete random variables and their probability distributions.

Given a chance experiment, the collection of possible outcomes is called the sample space, denoted as \mathbf{S} . A **random variable** is a function (or a mapping) from the sample space \mathbf{S} into real numbers. Random variables are usually denoted as uppercase letters, such as X, Y, Z . We use the corresponding lowercase letters x, y, z to represent possible values that random variables may attain.

Example: Random Variable

1. Consider the chance experiment of flipping a balanced coin twice, the sample space is $\mathbf{S} = \{HH, HT, TT, TH\}$. Let the random variable $X = \#$ of tails. It is a mapping from the sample space \mathbf{S} to integers 0, 1, and 2.

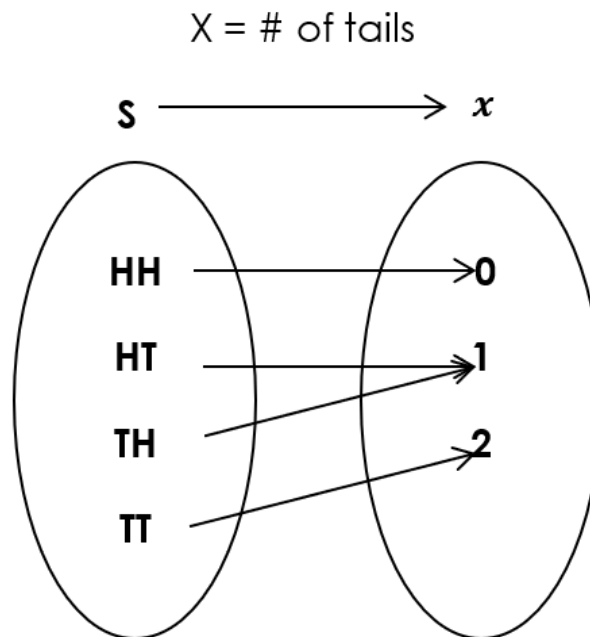


Figure 4.1: Random Variable $X = \#$ of Tails. [[Image Description](#)
([See Appendix D Figure 4.1](#))]

2. Five students are asked to report the number of siblings they have; their responses are summarized in the following table:

Name	Mark	John	Rebecca	Sarah	Mary
# of Siblings	0	1	2	2	3

Randomly pick one student and let random variable $X = \#$ of siblings the student has. Then X is a mapping from the sample space $\mathbf{S} = \{\text{Mark, John, Rebecca, Sarah, Mary}\}$ to the numbers 0, 1, 2 and 3.

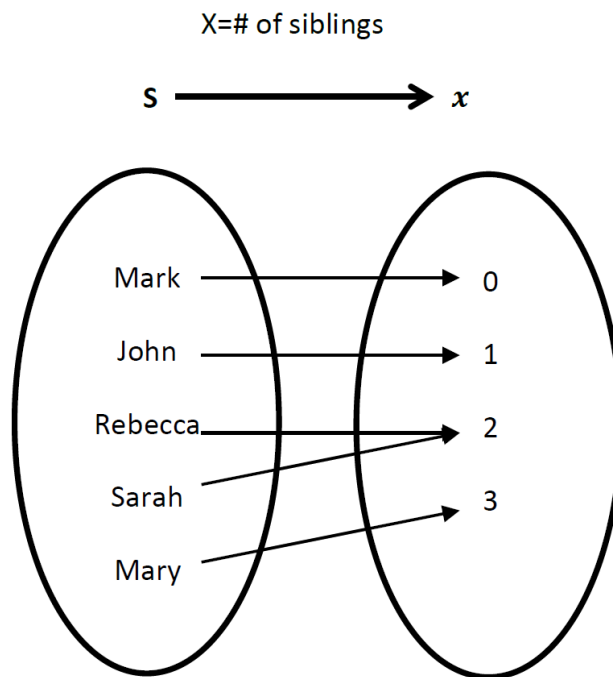


Figure 4.2: Random Variable $X = \#$ of Siblings. [\[Image Description \(See Appendix D Figure 4.2\)\]](#)

In general, a discrete random variable maps the sample space \mathbf{S} to numbers that can be listed or counted; a continuous random variable maps the sample space \mathbf{S} to an interval that is a subset of the entire real line. If you need to review discrete and continuous data, refer to [variables and data](#) in Chapter 1.3.

4.2 Probability Distribution of a Discrete Variable

The **probability distribution** of a discrete random variable X lists all possible values and their corresponding probabilities. In general, the probability distribution of a discrete variable is given in a table with two rows or two columns: one row (column) for the possible values x and the other row (column) for the corresponding probability of taking each value $P(X = x)$. A probability distribution has two important properties:

Key Fact: Two Important Properties of Probability Distribution

- $0 \leq P(X = x) \leq 1$
- $\sum_{\text{all possible } x} P(X = x) = 1$

Example: Probability Distribution of a Discrete Variable

1. Consider the chance experiment of flipping a balanced coin twice; the sample space is $S = \{HH, HT, TT, TH\}$. Let the random variable $X = \#$ of tails. Determine the probability distribution of X .
 - First, determine the possible values of X . If we flip a coin twice, we might observe zero tail (HH), one tail (HT, TH), and two tails (TT). Therefore, possible values are $x = 0, 1, 2$.
 - Next, determine the probabilities $P(X = x)$, $x = 0, 1, 2$. Since the coin is balanced, it follows that $P(H) = P(T) = 0.5$. Moreover, the two flips are independent, so the special multiplication rule applies. Thus,
 - $P(X = 0) = P(HH) = P(H) \times P(H) = 0.5 \times 0.5 = 0.25$.
 - $P(X = 1) = P(HT \text{ or } TH) = P(HT) + P(TH) = 0.25 + 0.25 = 0.5$.
 - $P(X = 2) = P(TT) = 0.25$.

Therefore, the probability distribution of X is

Table 4.1: Probability Distribution of $X = \#$ of Heads

x	0	1	2
$P(X = x)$	0.25	0.5	0.25

Note that the sum of the probabilities is one. That is

$$\begin{aligned}\sum P(X = x) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.25 + 0.5 + 0.25 = 1.\end{aligned}$$

2. A population consists of five students: Mark has no siblings, John has one sibling, both Rebecca and Sarah have two siblings, and Mary has three. Randomly pick one student and let X be the number of siblings the student has. Determine the probability distribution of X .

- First, observe that the possible values of X are $x = 0, 1, 2, 3$
- Next, determine the probabilities $P(X = x)$, $x = 0, 1, 2, 3$. One student has no siblings, two students have two siblings, and one student has three siblings. Thus,

- $P(X = 0) = \frac{f}{N} = \frac{1}{5} = 0.2.$
- $P(X = 1) = \frac{f}{N} = \frac{1}{5} = 0.2.$
- $P(X = 2) = \frac{f}{N} = \frac{2}{5} = 0.4.$
- $P(X = 3) = \frac{f}{N} = \frac{1}{5} = 0.2.$

The probability distribution of X is

Table 4.2: Probability Distribution of X =# of Siblings

x	0	1	2	3
$P(X = x)$	0.2	0.2	0.4	0.2

Note that the sum of the probabilities is one. That is

$$\begin{aligned}\sum P(X = x) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.2 + 0.2 + 0.4 + 0.2 = 1.\end{aligned}$$

A probability distribution can also be represented as a histogram. Every possible value should have a separate bar for a discrete random variable.

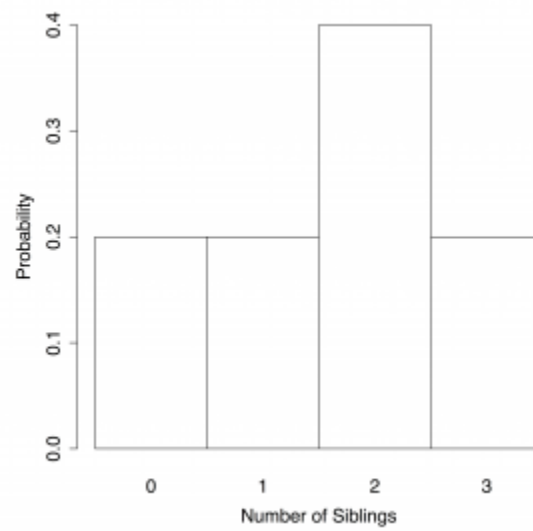


Figure 4.3: Histogram of # of Siblings [[Image Description \(See Appendix D Figure 4.3\)](#)]

4.3 Defining Events Using Random Variable Notation

We can define events using the notation of random variables and we can compute probabilities of events based on the probability distributions of the variables. For example, the event of having one sibling can be written as $\{X = 1\}$ and its probability is $P(X = 1) = 0.2$. The event of having at least one sibling is $\{X \geq 1\}$ and its probability is

$$\begin{aligned}P(X \geq 1) &= P(X = 1 \text{ or } X = 2 \text{ or } X = 3) \\&= P(X = 1) + P(X = 2) + P(X = 3) \\&= 0.2 + 0.4 + 0.2 = 0.8.\end{aligned}$$

Alternatively, we can apply the complement rule to find the probability:

$$\begin{aligned}P(X \geq 1) &= P(X = 1 \text{ or } X = 2 \text{ or } X = 3) \\&= 1 - P(X = 0) \\&= 1 - 0.2 = 0.8.\end{aligned}$$



Activity

Exercise: Define Events Using Random Variable

Consider the probability distribution of $X = \#$ of siblings below.

x	0	1	2	3
$P(X = x)$	0.2	0.2	0.4	0.2

Define the following events using the variable X and find the probabilities:

- Having one and a half siblings.
- Having zero to two siblings exclusively.
- Having zero to two siblings inclusively.

Show/Hide Answer

- Event: $\{X = 1.5\}$. Since the possible values of X are $x = 0, 1, 2, 3$, it is impossible for X to be 1.5; therefore, $P(X = 1.5) = 0$.

- b. Event: $\{0 < X < 2\}$. Since 1 is the only possible value that is greater than 0 and smaller than 2, $\{0 < X < 2\} = \{X = 1\}$. This implies $P(0 < X < 2) = P(X = 1) = 0.2$.
- c. Event: $\{0 \leq X \leq 2\} = \{X = 0 \text{ or } X = 1 \text{ or } X = 2\}$. Since the events are mutually exclusive, i.e., they don't overlap, the special addition rule gives:

$$\begin{aligned} P(0 \leq X \leq 2) &= P(X = 0 \text{ or } X = 1 \text{ or } X = 2) \\ &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.2 + 0.2 + 0.4 \\ &= 0.8. \end{aligned}$$

4.4 Mean and Standard Deviation of a Discrete Variable

Given the probability distribution of a discrete random variable X , we are able to calculate the mean, variance, and standard deviation.

The **mean** of a discrete variable X can be calculated as

$$\mu = \sum xP(X = x),$$

which is a weighted average over all possible values of X . Each possible value x is weighted by its probability $P(X = x)$. The mean is also called the **expected value** or the **expectation** of X . Note that a probability distribution can be viewed as the relative frequency distribution of some population. In this regard, the mean of a discrete random variable is equivalent to the population mean.



Instructor's Note

Some students might find it easier to find the mean by constructing a working table (see the following example).

Example: Mean of a Discrete Variable

Mark has no siblings, John has one sibling, both Rebecca and Sarah have two siblings, and Mary has three. Randomly pick one student and let X be the number of siblings the student has. Find the mean (expected value) of X .

We calculate the mean of X by constructing a working table. The first two columns of the table give the probability distribution of X and, in each row, the value in the third column is the product of the first two values.

Table 4.3: Working Table for the Mean (Expected Value) of a Discrete Variable

x	$P(X = x)$	$xP(X = x)$
0	0.2	$0 \times 0.2 = 0$
1	0.2	$1 \times 0.2 = 0.2$
2	0.4	$2 \times 0.4 = 0.8$
3	0.2	$3 \times 0.2 = 0.6$
Sum	$\sum P(X = x) = 1.0$	$\sum xP(X = x) = 1.6$

Taking the sum of the values in the last column gives the mean (expected value) of X , i.e.,

$$\begin{aligned}
 \mu &= \sum xP(X = x) \\
 &= 0 + 0.2 + 0.8 + 0.6 \\
 &= 1.6.
 \end{aligned}$$

Interpretation: On average, each of those five students has 1.6 siblings.



Instructor's Note

Even though the random variable $X = \#$ of siblings can only take integer values, we should keep the decimal place for the mean of X . That is, do not round the mean 1.6 to 2. To demonstrate why we keep the mean at 1.6, let us suppose that this probability distribution describes a much larger population of students. Although it is counterintuitive to say that we expect a student to have 1.6 siblings, it is quite natural to say that we expect 10 students to have a total of 16 siblings, 100 students to have a total of 160 siblings, and so on. If we sample the entire population of students, then the combined number of siblings is 1.6 times greater than the number of students. Hence, the average number of siblings per student is 1.6.

Here we explain why the population mean is given by $\mu = \sum xP(X = x)$. Suppose there are $N = 5$ students, one has no siblings ($x_1 = 0$), one has one sibling ($x_2 = 1$), two have two siblings ($x_3 = x_4 = 2$), and one has three siblings ($x_5 = 3$). Recall that the population mean μ is calculated as:

$$\begin{aligned}
 \mu &= \frac{\sum x_i}{N} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{0 + 1 + 2 + 2 + 3}{5} = \frac{0 \times 1 + 1 \times 1 + 1 \times 2 + 2 \times 2 + 3 \times 1}{5} \\
 &= 0 \times \frac{1}{5} + 1 \times \frac{1}{5} + 2 \times \frac{2}{5} + 3 \times \frac{1}{5} \\
 &= 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) \\
 &= \sum x e P(X = x).
 \end{aligned}$$

Similarly, the **variance** of a discrete variable X can be calculated as

$$\sigma^2 = \underbrace{\sum (x - \mu)^2 P(X = x)}_{\text{defining formula}}.$$

which is a weighted average of the squared distance from each value x to the population mean μ , weighted by its probability $P(X = x)$ (relative frequency). It can be shown that

$$\sigma^2 = \sum (x - \mu)^2 P(X = x) = \underbrace{\sum x^2 P(X = x) - \mu^2}_{\text{computing formula}}.$$

Taking the square root of the variance σ^2 gives the standard deviation σ

$$\sigma = \sqrt{\sigma^2} = \underbrace{\sqrt{\sum (x - \mu)^2 P(X = x)}}_{\text{defining formula}} = \underbrace{\sqrt{\sum x^2 P(X = x) - \mu^2}}_{\text{computing formula}}.$$

Example: Standard Deviation of a Discrete Variable

Mark has no siblings, John has one sibling, both Rebecca and Sarah have two siblings, and Mary has three. Randomly pick one student and let X be the number of siblings the student has. Find the standard deviation of X .

We can find the standard deviation using a working table:

Table 4.4: Standard Deviation Using Computing Formula

x	$P(X = x)$	x^2	$x^2 P(X = x)$
0	0.2	$0^2=0$	$0 \times 0.2 = 0$
1	0.2	$1^2=1$	$1 \times 0.2 = 0.2$
2	0.4	$2^2=4$	$4 \times 0.4 = 1.6$
3	0.2	$3^2=9$	$9 \times 0.2 = 1.8$
Sum	$\sum P(X = x) = 1.0$		$\sum x^2 P(X = x) = 3.6$

The standard deviation of X is $\sigma = \sqrt{\sum x^2 P(X = x) - \mu^2} = \sqrt{3.6 - 1.6^2} = \sqrt{1.04} = 1.02$.

Interpretation: Roughly speaking, on average, the number of siblings of those five students is 1.02 away from the mean 1.6.



Activity

Exercise: Discrete Random Variable and Its Probability Distribution

For one insurance policy, the company pays out \$10,000 if the customer dies, \$5,000 if the customer is disabled and \$0 for other situations. Suppose the probability of death is 0.001 and the probability of being disabled is 0.002. Let X be the amount of money the company pays.

1. Find the probability distribution of X . Complete the following table:

$x()$	$P(X = x)$
10000	
5000	

2. Find the mean (expected value) of X .
3. Find the standard deviation of X .
4. Suppose the company wants to make an average profit of \$50 per customer. Calculate the premium it should charge each customer.

Show/Hide Answer

1. The probability distribution consists of two components: possible values and probabilities.

$x()$	$P(X = x)$
10000	0.001
5000	0.002
0	0.997

2. Based on the working table below, the mean is calculated as

$$\mu = \sum xP(X = x) = 10 + 10 + 0 = 20.$$

$x()$	$P(X = x)$	$xP(X = x)$
10000	0.001	$10000 \times 0.001 = 10$
5000	0.002	$5000 \times 0.002 = 10$
0	0.997	$0 \times 0.997 = 0$
Sum	$\sum P(X = x) = 1.000$	$\sum xP(X = x) = 20$

Interpretation: On average, the company pays out \$20 for each customer.

3. Based on the working table below, the standard deviation is given by

$$\sigma = \sqrt{\sum x^2 P(X = x) - \mu^2} = \sqrt{150000 - 20^2} = \sqrt{149600} = 386.78.$$

$x()$	$P(X = x)$	$x^2 P(X = x)$
10000	0.001	$10000^2 \times 0.001 = 100000$
5000	0.002	$5000^2 \times 0.002 = 50000$
0	0.997	$0^2 \times 0.997 = 0$
$\sum P(X = x) = 1.000$		$\sum x^2 P(X = x) = 150000$

Interpretation: Roughly speaking, on average, the payout is \$386.78 different from the mean \$20.

Note: The standard deviation in this example does not have much of a practical meaning. Compared to the expected value, the standard deviation is large, indicating a large variation in the payout. This is due to the fact that the payout is one of 0, 5000 or 10,000 and majority of customers will receive 0 payout.

4. On average the company pays out \$20 for customer. If the company wants to make an average profit of \$50 per customer, it must ask for \$20 more on the top of \$50. Therefore, the company should charge $50 + 20 = 70$ dollars per customer.



Activity

Exercise: Mean and Standard Deviation of a Discrete Random Variable

Let X be the number of patients arriving at an emergency centre from 9 to 9:30 PM. The probability distribution of X is given in the following table.

x	0	1	2
$P(X = x)$	0.3	$4a$	$3a$

- a. Find the value of a .
- b. Find the mean of X .
- c. Find the standard deviation of X .

Show/Hide Answer

- a. Since the sum of probabilities $P(X = x)$ is one, $0.3 + 4a + 3a = 1 \implies 7a = 0.7 \implies a = 0.1$.
- b. The mean is $\mu = \sum xP(X = x) = 0 \times 0.3 + 1 \times 0.4 + 2 \times 0.3 = 1$.

c. The standard deviation is given by:

$$\begin{aligned}\sigma &= \sqrt{\sum x^2 P(X = x) - \mu^2} = \sqrt{(0^2 \times 0.3 + 1^2 \times 0.4 + 2^2 \times 0.3) - 1^2} \\ &= \sqrt{1.6 - 1^2} = \sqrt{0.6} = 0.775.\end{aligned}$$

It might be helpful to construct the working table below:

x	$P(X = x)$	$xP(X = x)$	$x^2P(X = x)$
0	0.3	$0 \times 0.3 = 0$	$0^2 \times 0.3 = 0$
1	0.4	$1 \times 0.4 = 0.4$	$1^2 \times 0.4 = 0.4$
2	0.3	$2 \times 0.3 = 0.6$	$2^2 \times 0.3 = 1.2$
Sum	1.0	$\sum xP(X = x) = 1.0$	$\sum x^2P(X = x) = 1.6$

4.5 Binomial Distribution

A special and useful **discrete** probability distribution is the **binomial distribution**. Before introducing binomial distribution, we first introduce Bernoulli trials.

Chance experiments satisfying the following conditions are called **Bernoulli trials**:

1. Only two possible outcomes: success and failure.
2. The probability of success, p , is the same on every trial.
3. Trials are independent. Observing a head on the current trial won't affect the result of the next trial. If trials are not strictly independent, it is still okay as long as the sample is less than 10% of the population.
4. The number of trials is fixed.

4.5.1 Probability Distribution of a Binomial Random Variable

Let X be the number of successes in a sequence of n Bernoulli trials with probability of success p . Then X follows a binomial distribution with the probability distribution given by

$$P(X = x) = {}_n C_x p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n.$$

Note that even though the probability distribution is not given in the table form, it lists all possible values of x and their probabilities.

Example: Binomial Probability Distribution

If we roll a balanced die three times, find the probability of rolling exactly one six.

Let X be the number of six observed, then X follows a binomial distribution with $n = 3$, $p = P(\text{observing a six}) = \frac{1}{6}$. The probability of rolling exactly one six, $P(X = 1)$, can be calculated using the binomial probability distribution formula with $n = 3$, $p = \frac{1}{6}$, $x = 1$:

$$P(X = 1) = {}_3 C_1 \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^{3-1} = 3 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^2 = 0.3472.$$

Interpretation of the binomial probability equation is as follows: There are n Bernoulli trials and hence n outcomes that are either a success or a failure. We want to find the probability of observing x successes. If we observe n successes, there must be $(n - x)$ failures. The probability of success is p and the probability of failure is $(1 - p)$. Hence, the probability of x consecutive successes followed by $(n - x)$ failures is $p^x(1 - p)^{n-x}$ because the trials are independent and thus the [special multiplication rule](#) applies. Next, observe that there are ${}_nC_x$ ways to rearrange x successes and $(n - x)$ failures and so the probability of x successes, disregarding order, is ${}_nC_x p^x(1 - p)^{n-x}$. The term ${}_nC_x$ is called the binomial coefficient. For example, if we conduct the trial three times and observe one success, it can occur in the first, second, or third trials. The number of ways to observe one success out of three trials is ${}_3C_1 = 3$.



Activity

Exercise: Binomial Distribution

Does the following random variable X follow a binomial distribution? If yes, identify the parameters n (total number of Bernoulli trials) and p (probability of success).

1. Flip a balanced coin six times, and let X = # of heads observed.
2. A multiple-choice exam has ten questions, each with four answers: A, B, C, and D. I didn't study and guess the answers. Let X = # of correct answers.
3. Roll a balanced die 10 times, let X = # of sixes observed.
4. There are 10 students: 6 female and 4 male. Randomly pick 5 students without replacement, let X = # of female students.
5. There are 10,000 students: 6,000 female and 4,000 male. Randomly pick 5 students without replacement, let X = # of female students.
6. A bad basketball player has a 10% chance of making a basket each time he tries. Assume trials are independent. He will continue trying until he has made 2 baskets. Let X = number of trials.

Show/Hide Answer

1. Yes. Flipping a coin and observing the outcome is a Bernoulli trial since the trials are independent and the probability of success (observing a head) is constant. Therefore, X has a binomial distribution with $n = 6$, $p = 0.5$.
2. Yes. I assume I have no idea how to answer any of the questions, so I randomly pick one out of four choices. Each guess is a Bernoulli trial since the trials are independent (the outcome of any guess should have no influence on the outcome of another guess), and each trial has a $1/4$ probability of success (since there are 4 answers, but only 1 is correct). Therefore, since there is a total of 10 guesses, it follows that X has a binomial distribution with $n = 10$, $p = \frac{1}{4} = 0.25$.

3. Yes. Rolling a die and observing whether the outcome is a six is a Bernoulli trial. There is a total of 10 independent trials and rolling a six is the success event. Therefore, X has a binomial distribution with $n = 10, p = \frac{1}{6}$.
4. No. Even though there are only two possible options (female and male), the trials are not independent since we are sampling without replacement (the chance of picking a female student in the current trial depends on the outcome of the previous ones). Therefore, the probabilities of success are different from one trial to another.
5. This can be regarded as a binomial distribution but is not exact. Again, because of sampling without replacement, the probability of picking a female student is not exactly the same. But they are almost the same given the fact that the sample $n = 5$ is way less than 10% of the population size $N = 10000$, i.e., $n = 5 < 0.1 \times 10000 = 1000$.
6. No. The number of trial n here is a random variable rather than a fixed value; therefore, X does not follow a binomial distribution.

4.5.2 Mean and Standard Deviation of Binomial Distribution

Recall that the general formulas for the mean and the standard deviation of a discrete variable are:

$$\mu = \sum xP(X = x), \quad \sigma = \sqrt{\sum x^2P(X = x) - \mu^2}.$$

And the probability distribution of a binomial random variable is given by

$$P(X = x) = {}_n C_x p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n.$$

Plugging the third equation in the first and second equations gives the mean and standard deviation of a binomial random variable:

$$\begin{aligned} \mu &= \sum xP(X = x) = \sum x \underbrace{{}_n C_x \times (p^x) \times (1 - p)^{n-x}}_{\text{prob dist'n of binomial}} = np; \\ \sigma &= \sqrt{\sum x^2P(X = x) - \mu^2} = \sqrt{\sum x^2 \underbrace{{}_n C_x \times (p^x) \times (1 - p)^{n-x}}_{\text{prob dist'n of binomial}} - (np)^2} = \sqrt{np(1 - p)}. \end{aligned}$$

Note that we can use $\mu = np, \sigma = \sqrt{np(1 - p)}$ to find the mean and standard deviation of a

discrete random variable X if and only if X follows a binomial distribution with parameters n and p .

4.5.3 Steps to Find Probabilities Related to Binomial Distribution

In order to apply the binomial probability formula, we need to make sure that the variable follows a binomial distribution by checking:

1. Does each trial in the experiment have only two possible outcomes?
2. Are the trials independent?
3. Does each trial have the same probability of success?
4. Is the number of trials fixed?

If the answers to all the above questions are yes and we perform n trials, let $X = \#$ of successes, then we can claim that X follows a binomial distribution. We can apply the binomial probability formula as follows:

1. Identify the success event.
2. Determine the probability of success p .
3. Determine n the total number of trials.
4. Write down the event of interest in terms of the binomial variable X .
5. Apply the binomial probability formula $P(X = x) = {}_n C_x p^x (1 - p)^{n-x}$ to calculate the probability of each outcome in the event, and then add these probabilities together.

Example: Application of Binomial Distribution

A quiz consists of 10 multiple-choice questions with four choices: A, B, C, and D. I did not study and randomly picked one answer for each question.

1. Find the probability that I get six correct answers.
2. Find the probability that I get at least one correct answer.
3. Find the probability that I get at least nine correct answers.
4. How many correct answers do you expect me to get?

Solutions:

For each question, I either get the correct answer or not. Since I randomly picked one answer, each of the four choices had the same chance of being chosen. Therefore, since there is only one correct answer, the probability of obtaining the correct answer is $\frac{1}{4}$. Whether I obtain the correct answer for the current question will not affect the chance of getting the correct answer for the next question; therefore, the trials are independent with a constant probability of success. Let X = # of correct answers, then X follows a binomial distribution.

1. Identify the success event. Since X = # of correct answer = # of successes, getting a correct answer in each guess is a success.
2. Determine the probability of success p .
 $p = \text{probability of getting a correct answer} = \frac{1}{4} = 0.25$.
3. Determine n the total number of trials. Since the quiz has 10 questions, and we randomly pick one answer per question, this is a sequence of 10 Bernoulli trials. Therefore, $n = 10$.

Let X = # of correct answers, then X follows a binomial distribution with 0.25. The probability distribution is

$$P(X = x) = {}_n C_x p^x (1-p)^{n-x} = {}_{10} C_x 0.25^x (1-0.25)^{10-x} = {}_{10} C_x 0.25^x 0.75^{10-x},$$

for $x = 0, 1, \dots, 10$.

1. Find the probability that I get six correct answers.

Event: $\{X = 6\}$ with probability

$$P(X = 6) = {}_{10} C_6 (0.25^6)(0.75^{10-6}) = 210(0.25^6)(0.75^4) = 0.01622.$$

2. Find the probability that I get at least one correct answer.

Event: $\{X \geq 1\}$. That is one or more correct answers. By the complement rule,

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + \dots + P(X = 9) + P(X = 10) = 1 - P(X = 0) \\ &= 1 - {}_{10} C_0 (0.25^0)(0.75^{10-0}) = 1 - 0.75^{10} = 0.9437. \end{aligned}$$

By using the complement rule, we only need to apply the binomial probability formula once; however, we would need to apply it 10 times if we add $P(X = 1)$ to $P(X = 10)$.

3. Find the probability that I get at least nine correct answers.

Event: $\{X \geq 9\}$. That is 9 or 10 correct answers.

$$\begin{aligned} P(X \geq 9) &= P(X = 9) + P(X = 10) \\ &= {}_{10} C_9 (0.25^9)(0.75^{10-9}) + {}_{10} C_{10} (0.25)^{10} (0.75^{10-10}) \\ &= 10(0.25^9)(0.75^1) + 1(0.25)^{10}(0.75^0) \\ &= 2.861023 \times 10^{-5} + 9.536743 \times 10^{-7} \\ &= 2.9564 \times 10^{-5}. \end{aligned}$$

4. How many correct answers do you expect me to get?

Since X follows a binomial distribution, the expected value (the mean) of X is

$$\mu = np = 10 \times 0.25 = 2.5.$$

Interpretation: For every 10 questions, I expect to obtain 2.5 correct guesses.

Note that we do not round the expected value. Even though it is not possible to observe 2.5 correct

answers, this would be the long running average if I was to repeatedly conduct this experiment. An alternative viewpoint is this: expecting 2.5 correct answers for every 10 guesses is equivalent to expecting 25 correct answers for every 100 guesses, 250 correct answers for every 1000 guesses, and so on.



Activity

Exercise: Lotto 649

Lotto 649 launched in 1982 is one of three national lottery games in Canada. Each play costs \$3 and includes one set of 6 numbers ranging from 1 to 49 for the Main Jackpot Draw. If the 6 numbers a player chosen match all the 6 winning numbers (order does not matter), he wins the Jackpot. The six winning numbers were 2, 8, 9, 16, 39, and 49 for the Wednesday, April 6 Lotto 649, and Jackpot's winning prize was \$18.7 million.

If I buy one Lotto 649 ticket each month for the next 10 years, what is the probability that I will win at least one jackpot?

1. Do we have independent Bernoulli trials?
2. Let X = # of jackpots I win over the next 10 years. Does X follow a binomial distribution?

Show/Hide Answer

1. For each Lotto 649 ticket, I have only two possible outcomes: either win the jackpot or do not win the jackpot. Purchasing one ticket for each per month for the next 10 years yields a total of $12 \times 10 = 120$ tickets, i.e., 120 independent Bernoulli trials.
2. The question asks for # of jackpots to be won in the coming 10 years; therefore, winning a jackpot is a success. The probability of success is $p = \frac{1}{49C_6}$.

Let X = # of jackpots won in the coming 10 years. Then X follows a binomial distribution with $n = 120$, $p = \frac{1}{49C_6} = 0.0000000715$. The probability distribution is

$$\begin{aligned} P(X = x) &= {}_n C_x p^x (1 - p)^{n-x} \\ &= {}_{120} C_x 0.0000000715^x (1 - 0.0000000715)^{120-x}. \end{aligned}$$

Event: at least one Jackpot = $\{X \geq 1\}$ with probability

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - {}_{120} C_0 (0.0000000715^0) [(1 - 0.0000000715)^{120-0}] \\ &= 1 - (1 - 0.0000000715)^{120} \\ &= 1 - 0.9999914 = 0.0000086. \end{aligned}$$



Activity

Exercise: Application of Binomial Distribution

Roll a balanced die four times,

- Find the probability of observing a six at least once.
- Find the probability of observing a six exactly once.
- Find the probability of observing a six between two and four times inclusively.
- How many times do we expect to observe a six?

Show/Hide Answer

Let X = number of times observing a six is observed among the four rolls, then X follows a binomial distribution with $n=4$, $p=1/6$.

a.

$$\begin{aligned}P(X \geq 1) &= 1 - P(X = 0) \\&= 1 - {}_4C_0 \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^{4-0} \\&= 1 - 0.482 = 0.518.\end{aligned}$$

b.

$$\begin{aligned}P(X = 1) &= {}_4C_1 \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^{4-1} \\&= 0.386.\end{aligned}$$

c.

$$\begin{aligned}P(2 \leq X \leq 4) &= P(X = 2) + P(X = 3) + P(X = 4) \\&= 1 - P(X = 0) - P(X = 1) \\&= 1 - {}_4C_0 \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^{4-0} - {}_4C_1 \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^{4-1} \\&= 1 - 0.482 - 0.386 = 0.132.\end{aligned}$$

- d. Since X follows a binomial distribution, its mean (expected value) is $\mu = np = 4 \times \frac{1}{6} = \frac{2}{3} = 0.667$.

4.6 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Determine the probability distribution of a discrete random variable (Section 4.2).
- Calculate the mean (expected value) and standard deviation of a discrete random variable based on its probability distribution (Section 4.4).
- Identify situations where the binomial distribution can be applied (Section 4.5).
- Calculate the probabilities of events using the binomial probability formula, if applicable (Section 4.5).
- Calculate the mean and standard deviation of a random variable that follows the binomial distribution (Section 4.5).

4.7 Review Questions

1. Let X be the number of repair calls an appliance repair shop may receive during an hour. The probability distribution of X is given in the following table:

x	0	1	2
$P(X=x)$	$2a$	$2a$	a

- Find the value of a .
 - Find the mean of X .
 - Find the standard deviation of X .
 - Are the events “receiving no more than one call” and “receiving two calls” mutually exclusive?
 - Are the events “receiving no more than one call” and “receiving two calls” independent? Explain using calculations.
2. Roll a balanced die four times,
- Find the probability of observing six at least once.
 - Find the probability of observing six exactly once.
 - Find the probability of observing six two to four times inclusively.
 - How many times do we expect to observe a six?
3. An insurance company wants to design a homeowner's policy for mid-priced homes. From data compiled by the company, it is known that the annual claim amount, X , in thousands of dollars, per homeowner is a random variable with the following probability distribution.

x	0	10	50	100	200
$P(X=x)$	0.95	0.045	0.004	0.0009	a

- Determine the value of a .
 - Find the expected annual claim amount per homeowner.
 - Determine the expected annual claim amount for every 1000 homeowners.
 - How much should the insurance company charge for the annual premium to average a net profit of \$50 per policy?
4. A sales representative for a tire manufacturer claims that the company's steel-belted radials last at least 35,000 miles. A tire dealer decides to check that claim by testing eight of the tires. If 75% or more of the eight tires he tests last at least 35,000 miles, he will purchase tires from the sales representative. If, in fact, 90% of the steel-belted

radials produced by the manufacturer last at least 35,000 miles, what is the probability that the tire dealer will purchase tires from the sales representative?

5. From past experience, the owner of a restaurant knows that, on average, 4% of the parties that make reservations never show. How many reservations can the owner accept and still be at least 80% sure that all parties that make a reservation will show?

Show/Hide Answer

1.

a. $2a + 2a + a = 1 \implies 5a = 1 \implies a = 0.2.$

b. $P(X = 0) = 2a = 0.4, P(X = 1) = 2a = 0.4, P(x = 2) = 0.2.$

$$\mu = \sum xP(X = x) = 0 \times 0.4 + 1 \times 0.4 + 2 \times 0.2 = 0.8.$$

$$\sigma = \sqrt{\sum x^2P(X = x) - \mu^2} = \sqrt{0^2 \times 0.4 + 1^2 \times 0.4 + 2^2 \times 0.2 - 0.8^2} =$$

c. 0.74833.

d. Let A be the event of receiving no more than one call and B be the event of receiving two calls. The two events are mutually exclusive since they don't overlap. That is $P(A \& B) = 0.$

e. Let A be the event of receiving no more than one call and B be the event of receiving two calls, then

$$P(A) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.4 + 0.4 = 0.8, \quad P(B) =$$

$$P(X = 2) = 0.4.$$

The two events

are NOT independent, since $P(A \& B) = 0 \neq P(A) \times P(B).$

2.

a. Let X be the number of six, then X follows a binomial distribution with $n = 4$ and $p = P(\text{rolling a six}) = \frac{1}{6}.$ We want

$$P(X \geq 1) = 1 - P(X = 0) = 1 - {}_4C_0 \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^{4-0} = 1 - 0.4823 = 0.5177.$$

b. We want

$$P(X = 1) = {}_4C_1 \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^{4-1} = 0.3858.$$

c. We want

$$P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4)$$

$$= {}_4C_2 \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{4-2} + {}_4C_3 \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{4-3} + {}_4C_4 \left(\frac{1}{6}\right)^4 \left(1 - \frac{1}{6}\right)^{4-4}$$

$$= 0.1157 + 0.0154 + 0.0008 = 0.1319.$$

$$\text{OR} = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - {}_4C_0 \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^{4-0} - {}_4C_1 \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^{4-1}$$

$$= 1 - 0.4823 - 0.3858 = 0.1319.$$

d. $\mu = np = 4 \times \frac{1}{6} = 0.6667.$

3.

$$\sum P(X = x) = 1 \implies 0.95 + 0.045 + 0.004 + 0.0009 + a = 1 \implies a =$$

a. $1 - 0.9999 = 0.0001$.

$$\mu = \sum xP(X = x) = 0 \times 0.95 + 10 \times 0.045 + 50 \times 0.004 + 100 \times 0.0009 +$$

b. $200 \times 0.0001 = 0.76(\$1000)$.

c. $1000 \times \mu = 1000 \times 0.76 = 760(\$1000)$.

d. $\mu + 50 = 760 + 50 = \$810$.

4. Let X be the number of steel-belted radials lasting at least 35000 miles, then X follows a binomial distribution with $n = 8$ and $p = 0.9$. Since 75% of eight is $0.75 \times 8 = 6$, we want

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8)$$

$$= {}_8C_6(0.9)^6(1 - 0.9)^2 + {}_8C_7(0.9)^7(1 - 0.9)^1 + {}_8C_8(0.9)^8(1 - 0.9)^0$$

$$= 0.1488 + 0.3826 + 0.4305 = 0.9619.$$

5. Since 4% of the parties never show, 96% will show. We want

$$0.96^n \geq 0.8 \implies n \leq \frac{\ln 0.8}{\ln 0.96} = 5.47 \implies n = 5.$$

4.8 Assignment 4

Purposes

The following questions assess your knowledge of the concept of the probability distribution of a discrete random variable, finding the mean (expected value) and the standard deviation of a discrete random variable as well as applications of binomial distribution.

Instructions

Part A

Complete the following:

1. A variable X of a finite population takes values 5, 7, or 8 and has the following frequency distribution:

x	5	7	8
frequency	25	40	60

- a. Determine the probability distribution of the random variable X . (3 marks)
 - b. Use random variable notation to describe the events that X takes the value 6, a value of at most 6, and a value greater than 6. (3 marks) $\{X = 6\}$, $\{X \leq 6\}$
 - c. Find $P(X = 6)$, $P(X \leq 6)$, and $P(X > 6)$. (5 marks: 1+2+2)
 - d. Construct a probability histogram for the random variable X . (3 marks)
 - e. Find the mean and the standard deviation of the random variable X . (8 marks: 3+5)
2. An American roulette wheel contains 38 numbers: 18 are red, 18 are black, and 2 are green. When the roulette wheel is spun, the ball is equally likely to land on any of the 38 numbers. Suppose that you bet \$1 on red. If the ball lands on a red number, you win \$1; otherwise you lose your \$1. Let X be the amount you win on your \$1 bet.
 - a. Determine the probability distribution of the random variable X . (3 marks)
 - b. Find the expected value of the random variable X . Interpret the expected value. (5 marks)

- c. Approximately how much would you expect to lose if you bet \$1 on red 100 times? 1,000 times? (4 marks)
- d. Is this game fair for the players? Explain. (2 marks)
3. There are four investments. The return on each investment depends on whether next year's economy is strong or weak (column variable). The following table summarizes the possible payoffs, in dollars, for the four investments (row variable). Let V , W , X , and Y denote the payoffs for the certificate of deposit, office complex, land speculation, and technical school, respectively. Then V , W , X , and Y are random variables. Assume that next year's economy has a 40% chance of being strong and a 60% chance of being weak.

- a. Determine the expected value of each random variable. (8 marks)
- b. Which investment has the best expected payoff? Which is the worst? (4 marks)
- c. Which investment would you select? Explain. (3 marks)

	Strong	Weak
Certificate of deposit	6,000	6,000
Office complex	15,000	5,000
Land speculation	33,000	-17,000
Technical school	5,500	10,000

4. A quiz consists of 10 multiple choices questions with five choices A, B, C, D, and E. I did not study and randomly picked one answer for each question. Find the probability that
- a. I get at least one question correct. (4 marks)
- b. I pass the quiz. A passing grade is 60% or better. (5 marks)
- c. I receive an "A" on the quiz (i.e., 90% or better). (3 marks)
- d. How many questions would you expect me to get correct? (2 marks)
- e. Obtain the standard deviation of the number of correct answers. (2 marks)

Part B

Finish the following questions using R and R commander

Use R commander to doublecheck your answers for Question 4 parts (a)–(c). Please copy and paste the computer outputs in the space below. (8 marks: 3+3+2)

Quiz 4



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2623#h5p-6>

CHAPTER 5: THE NORMAL DISTRIBUTION

Overview

A random variable can be either discrete or continuous. As mentioned in Chapter 4, a discrete random variable can be described by a probability distribution that lists all the possible values of the random variable and their corresponding probabilities. For a continuous random variable, however, we cannot list all the possible values, so a different approach is required to describe the probability distribution. The so-called density curve describes the probability distribution of a continuous random variable, and the area under the curve describes probabilities related to continuous random variables. This chapter introduces the normal distribution, one of the most important continuous distributions.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Describe the properties of a normal density curve.
- Describe the standard normal distribution.
- Use the standard normal table (Table II) in order to:
 - Determine the area bounded by some given values under any normal density curve.
 - Find values that correspond to given areas under any normal density curve.
- Use a normal probability plot to assess whether a given data set seems to come from a normal population.

5.1 Density Curve

We use a density curve to describe the distribution of a continuous variable. A density curve is the continuous analogue of a relative-frequency histogram with the density (scaled relative frequency) as the y-axis. For example, 1,000 grades are shown in the following grouping table and the relative-frequency histogram. The total area of the bins in the relative-frequency histogram is

$$\text{area} = 10 \times 0.003 + 10 \times 0.019 + 10 \times 0.150 + \dots + 10 \times 0.144 + 10 \times 0.022 = 10.$$

The total area of the three leftmost bins is

$$\text{area} = 10 \times 0.003 + 10 \times 0.019 + 10 \times 0.150 = 1.72,$$

which accounts for $1.72/10 = 0.172$ or 1.72% of the data. If we re-scale the y-axis of the relative-frequency histogram by dividing relative frequency (probability) by 10 (the bin width) and draw a smooth curve on the top of the histogram, we obtain the density curve of the grades with the y-axis density = relative frequency/width of the bins = relative frequency/10.

Interval	Relative Frequency
[30, 40)	0.003
[40, 50)	0.019
[50, 60)	0.150
[60, 70)	0.330
[70, 80)	0.332
[80, 90)	0.144
[90, 100]	0.022
Sum=1.000	

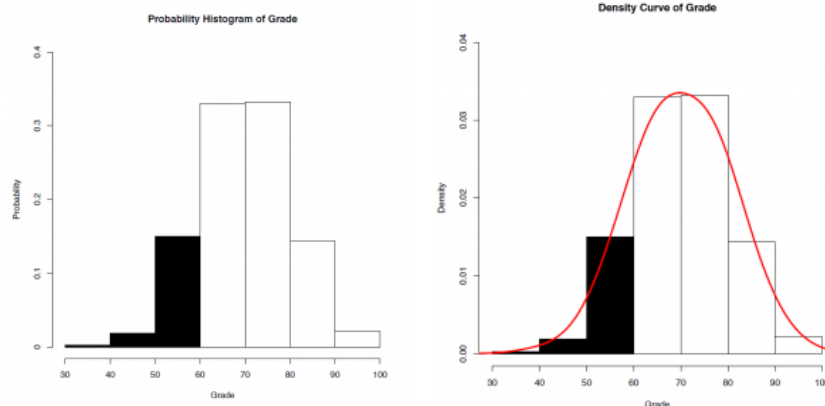


Table 5.1: Grouping Table of Grade

Figure 5.1: Probability Histogram (Left) and Density Curve (Right) of Grade. [\[Image Description \(See Appendix D Figure 5.1\)\]](#) Click on image to enlarge.

Let X = grade, the proportion of grades below 60 is given by

$$P(X < 60) = 0.003 + 0.019 + 0.150 = 0.172$$

= total area of the leftmost three bins
= area to the left of 60 under the red curve.

Similarly, the percentage of grades above 80, $P(X > 80) = 0.144 + 0.022 = 0.166$, which is the total area of the rightmost two bins or the area to the right of 80 under the red curve. The proportion (percentage) of grades between 60 and 80 is given by

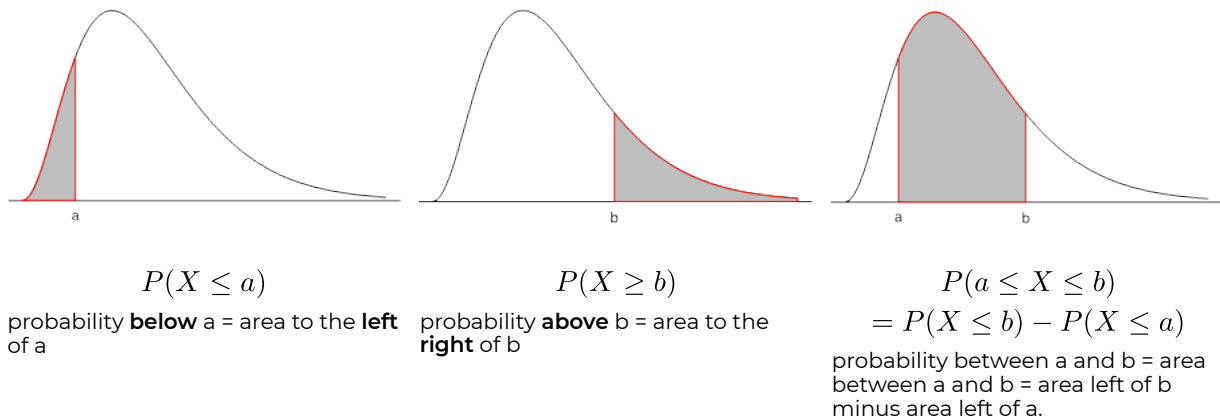
$$P(60 < X < 80) = 0.330 + 0.332 = 0.662$$

= total area of the two bins in the middle
= the area between 60 and 80 under the red curve.

A density curve has the following properties:

Key Facts: Properties of a density curve

- Total area under the curve is one.
- Area of a region under the curve gives the probability of an event.



[[Image Description \(See Appendix D Figure 5.1.1\)](#)]

5.2 Normal Density Curve

The normal density curve characterizes the normal distribution, which is the most widely used probability distribution for continuous variables. The normal distribution is symmetric and bell-shaped (for this reason it is often referred to as the “bell curve”). The normal density function has two parameters: the mean μ and the standard deviation σ . The parameter μ controls the centre (location) of the distribution and σ controls the shape of the distribution. When σ is larger, the curve appears shorter and fatter; when σ is smaller, the curve appears taller and slimmer.

Figure 5.2 shows three normal density curves— $N(0, 2)$, $N(0, 1)$ and $N(4, 1)$. $N(0, 1)$ and $N(4, 1)$ have the same standard deviation; therefore, they have the same shape; if you shift the location of $N(0, 1)$ to the right by 4, the two distributions are exactly the same. $N(0, 1)$ and $N(0, 2)$ have the same mean; therefore, they center at the same location. $N(0, 2)$ has a larger standard deviation; therefore, the density curve is shorter and fatter.

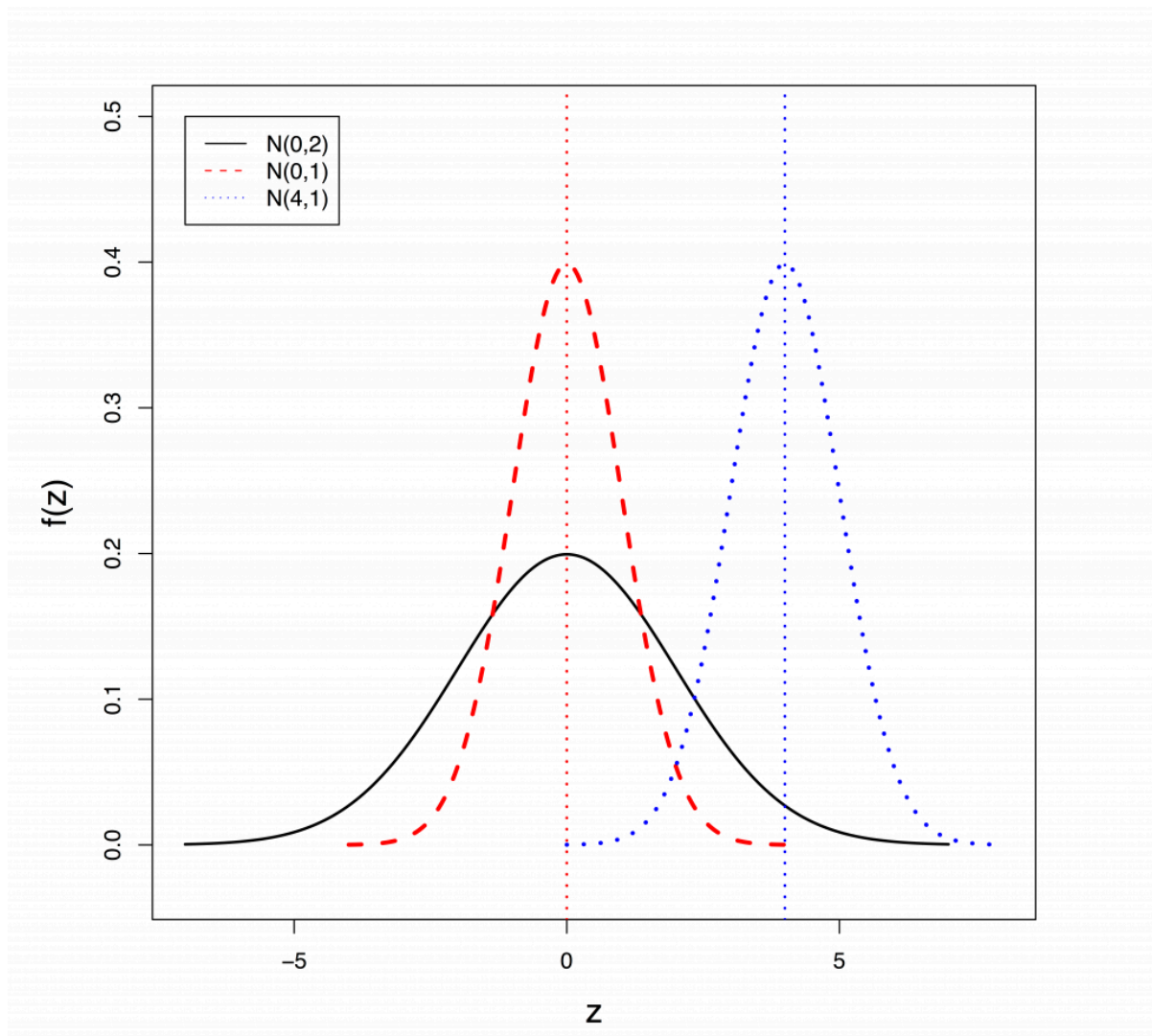


Figure 5.2: Three normal density curves. [[Image Description \(See Appendix D Figure 5.2\)](#)]

If a random variable X follows a normal distribution with mean μ and standard deviation σ , we write $X \sim N(\mu, \sigma)$. The probability density function of a normal random variable X is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \text{ with } \pi \approx 3.142 \text{ and } e \approx 2.718.$$

The normal density curve has the following properties:

Key Facts: Properties of a Normal Density Curve

- The curve extends from negative infinity ($-\infty$) to positive infinity (∞), i.e., the entire real line.

- The total area under the curve is 1. This is a common property for all density curves.
- The curve is bell-shaped, unimodal, and symmetric at the mean μ .
- Empirical rule (68.3-95.4-99.7 rule) for a normal curve:
 1. 68.26% of the observations are within the interval $[\mu - \sigma, \mu + \sigma]$ (one standard deviation to either side of the mean), i.e., $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6826$.
 2. 95.44% of the observations are within the interval $[\mu - 2\sigma, \mu + 2\sigma]$ (two standard deviations to either side of the mean), i.e., $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$.
 3. 99.74% of the observations are within the interval $[\mu - 3\sigma, \mu + 3\sigma]$ (three standard deviations to either side of the mean), i.e., $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9974$.

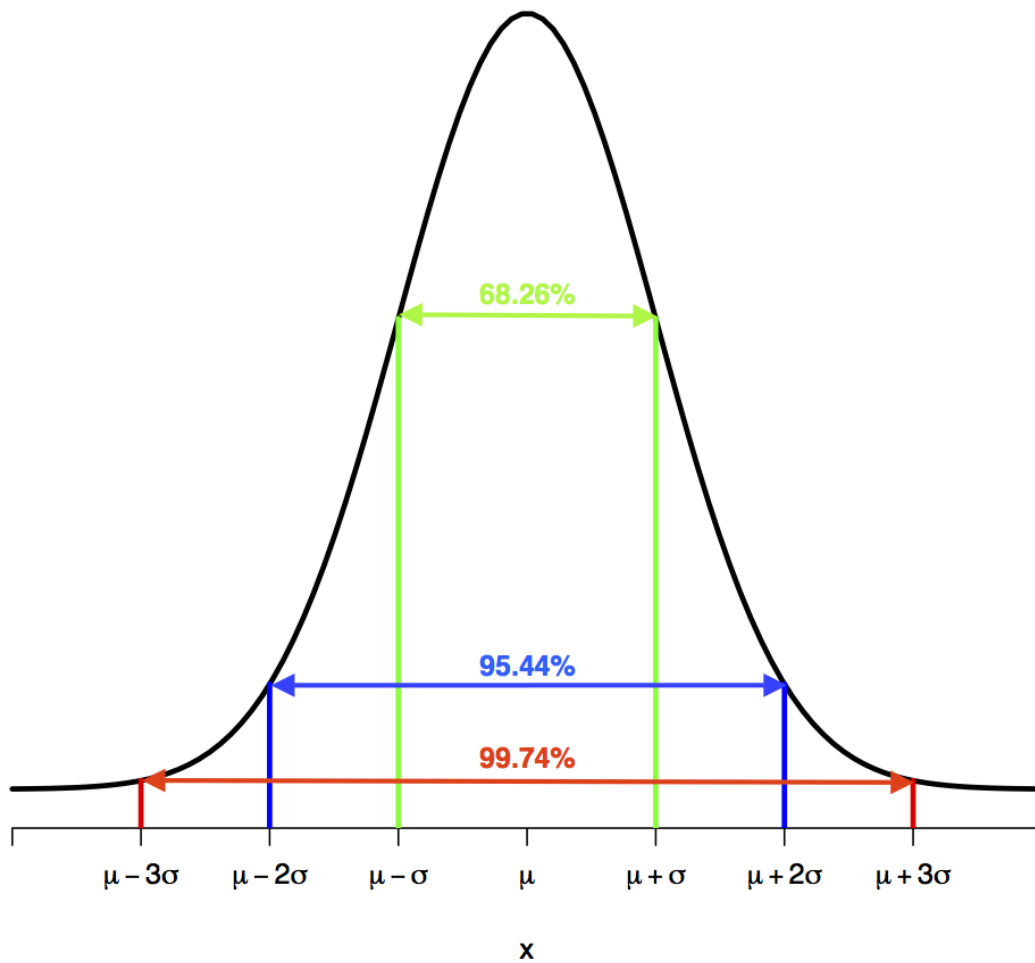


Figure 5.3: Empirical Rule of Normal Distribution. [[Image Description \(See Appendix D Figure 5.3\)](#)]

5.3 Standard Normal Density Curve

The standardized variable (z-score) has a mean of 0 and a standard deviation of 1. We can also standardize the normal variable $X \sim N(\mu, \sigma)$ by $Z = \frac{X - \mu}{\sigma}$ and Z follows a standard normal distribution with mean 0 and standard deviation 1, i.e., $Z \sim N(0, 1)$.

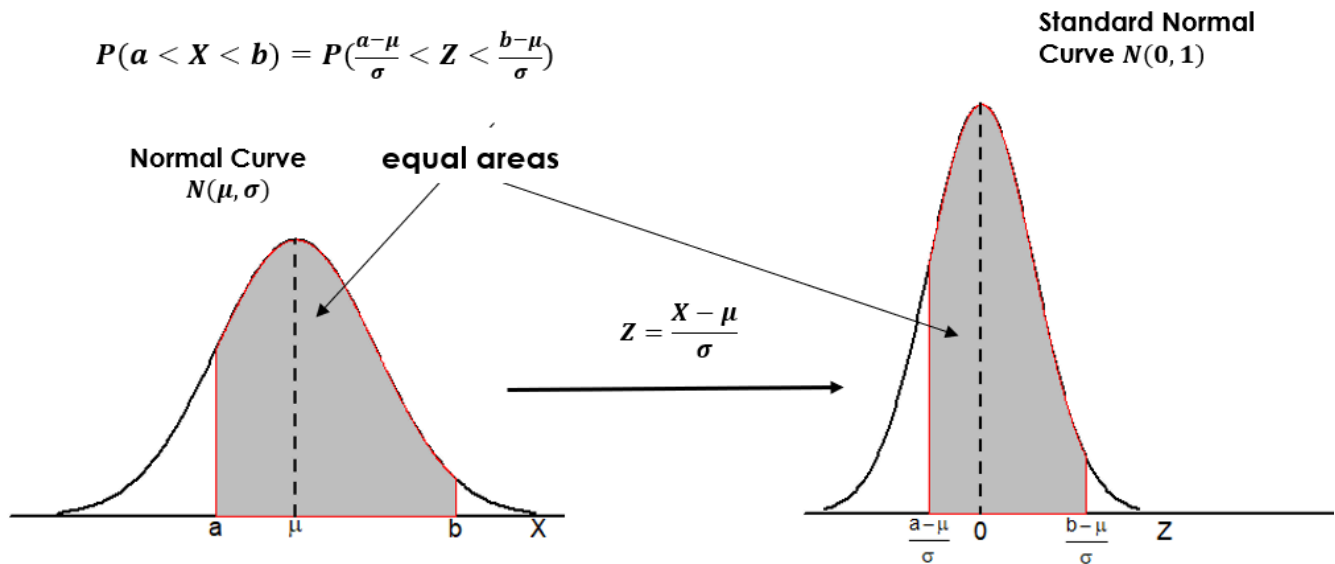


Figure 5.4: Mapping from Normal $N(\mu, \sigma)$ to Standard Normal $N(0, 1)$. [[Image Description \(See Appendix D Figure 5.4\)\]](#)

The standard normal density curve has the following properties:

Key Facts: Properties of a Standard Normal Density Curve

- Total area under the curve is 1.
- The curve extends from $-\infty$ to $+\infty$.
- Symmetric at 0, which means the area to the **right** of a positive number a is equal to the area to the **left** of $-a$. For example, $P(Z \geq 2) = P(Z \leq -2)$.
- Empirical rule:
 1. 68.26% of the observations are within the interval $[-1, 1]$. The area under the curve between -1 and 1 is 0.6826.
 2. 95.44% of the observations are within the interval $[-2, 2]$. The area under the curve between

-2 and 2 is 0.9544.

3. 99.74% of the observations are within the interval $[-3, 3]$. The area under the curve between -3 and 3 is 0.9974.

Standardization converts all normal distributions to a single one—the standard normal distribution. Therefore, we can calculate the probabilities of events relative to any normal distribution using only the standard normal density curve.

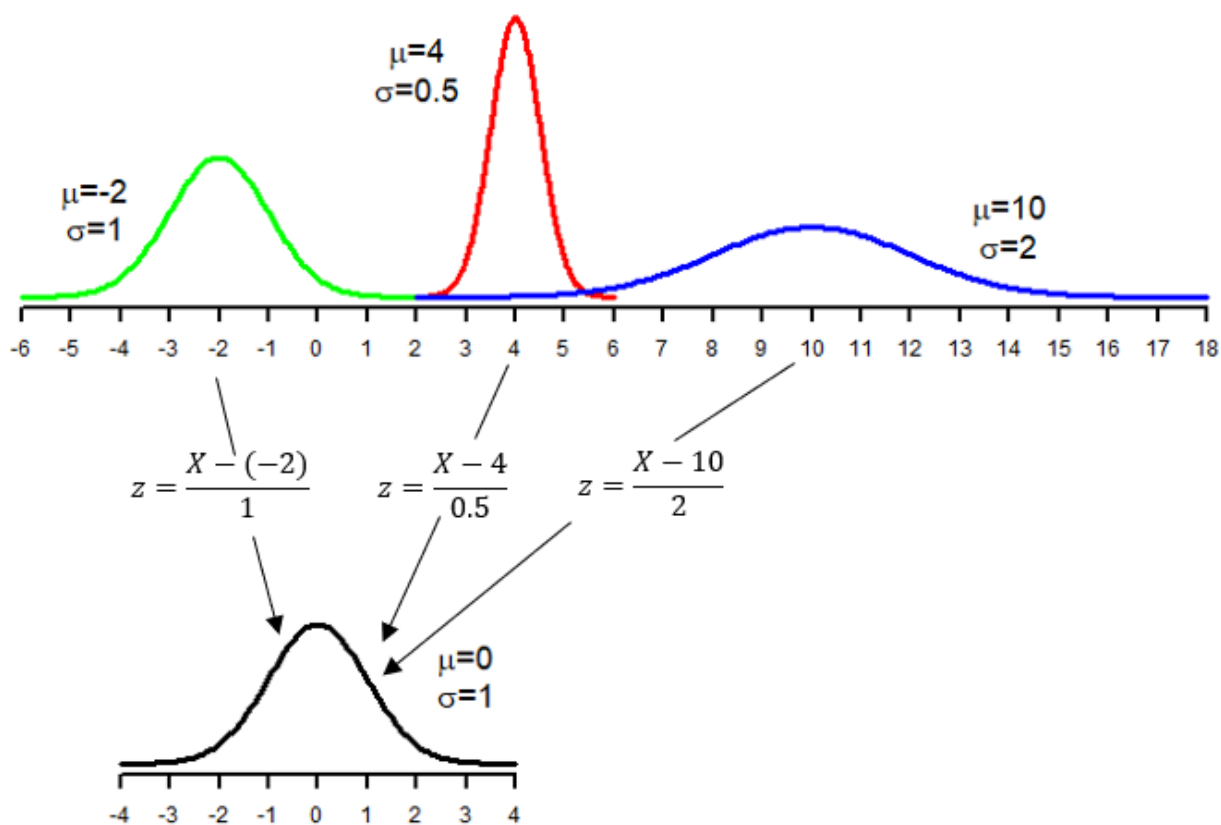


Figure 5.5: Converting Normal Distributions to Standard Normal. [[Image Description \(See Appendix D Figure 5.5\)](#)]

Example: Density Curve and Probability

Suppose the grades in a Statistics class are approximately normally distributed with mean $\mu = 70$ and standard deviation $\sigma = 10$.

- a. The percentage (proportion, to be more precise) of students with a grade **below** 60 equals the area under the standard normal curve to the **left** of **-1**. The z -score is calculated by
$$z = \frac{x-\mu}{\sigma} = \frac{60-70}{10} = -1.$$
- b. The percentage (proportion) of students with a grade **above** 90 equals the area under the standard normal curve to the **right** of **2**. The z -score is calculated by
$$z = \frac{x-\mu}{\sigma} = \frac{90-70}{10} = 2.$$
- c. The percentage (proportion) of students with a grade between 60 and 90 equals the area under the standard normal curve between **-1** and **2**. Figure 5.4 gives a graphical presentation of this question with $\mu = 70$, $\sigma = 10$, $a = 60$ and $b = 90$.

5.4 Using the Standard Normal Table

The standard normal table (usually found in the appendix of a Statistics textbook) can be used to solve problems related to normal distributions.

5.4.1 Find the Area (Probability) for a Given Z-Score

In general, the standard normal table gives the area under the standard normal curve to the left of a specified z-score. Using the table, we can calculate the area under the curve to the left of a z-score, to the right of a z-score, between two z-scores, or beyond two z-scores. Figure 5.6 shows that the area to the left of 1.96 under the standard normal curve is 0.975.

Table II: Area under the standard normal curve for positive z

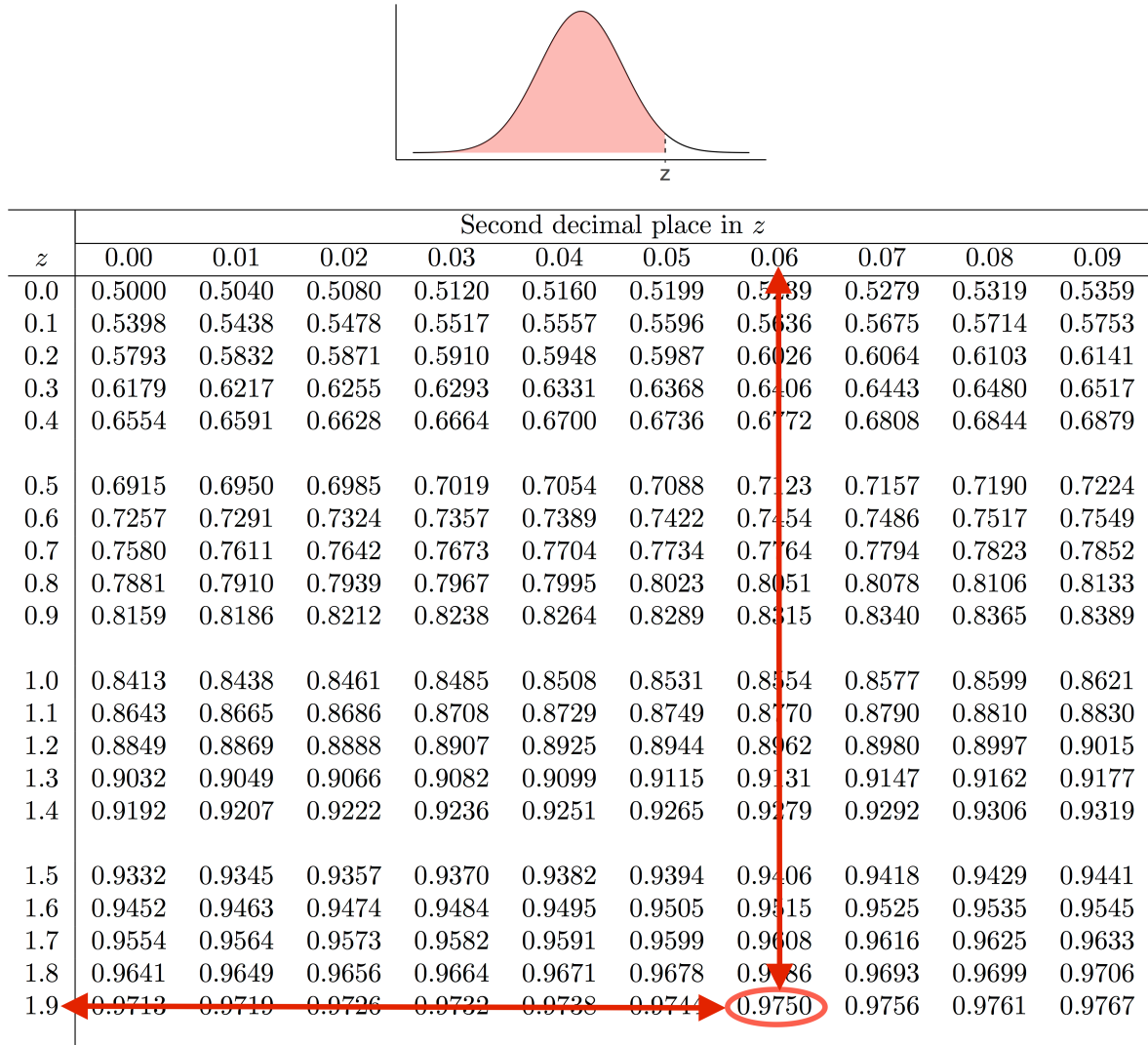


Figure 5.6: Area Under the Standard Normal Curve (Table II). [Image Description (See Appendix D Figure 5.6)]

If random variable Z follows a standard normal distribution, more detailed examples of using the standard normal table can be found in Figure 5.7:

- Left panel: the area to the left of 1.96 is 0.975, i.e., $P(Z < 1.96) = 0.975$.
- Middle panel: the area to the right of 1.96 is 0.025. There are two ways to solve this problem:
 1. Recall that the total area under any density curve is one, the area to the right of 1.96 equals one minus the area to the left of 1.96, i.e,

$$\begin{aligned}
 P(Z > 1.96) &= \text{area under the standard normal curve to the right of 1.96} \\
 &= 1 - \text{area under the curve to the left of 1.96} = 1 - 0.975 = 0.025.
 \end{aligned}$$

2. $P(Z > 1.96) = P(Z < -1.96) = 0.025$. This is because the standard normal curve is symmetric at 0. The area to the right of 1.96 equals the area to the left of -1.96.

- Right panel: the area between -1.96 and 1.96 is 0.95, i.e.,

$$\begin{aligned} P(-1.96 < Z < 1.96) &= \text{area between } -1.96 \text{ and } 1.96 \\ &= (\text{area to the left of } 1.96) - (\text{area to the left of } -1.96) \\ &= 0.975 - 0.025 = 0.95. \end{aligned}$$

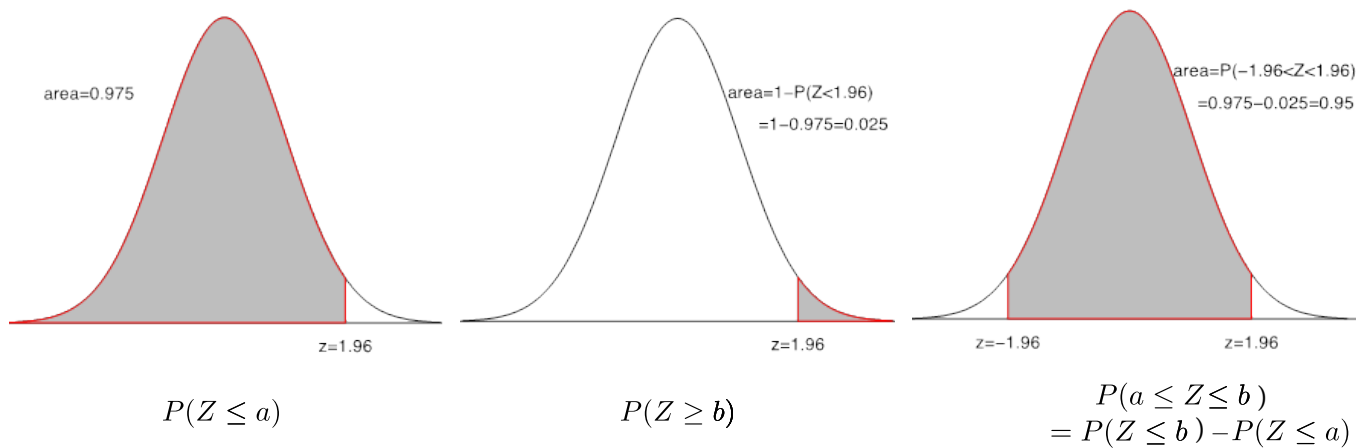


Figure 5.7: Area to the Left (left panel), Right (middle panel) of a Z-score, and Between Two Z-scores (right panel). [[Image Description \(See Appendix D Figure 5.7\)](#)]

Example: Finding Areas Under Standard Normal Curve

Suppose that $Z \sim N(0, 1)$, follows a standard normal distribution.

1. Draw a graph to show and find $P(Z < -2)$.

We can find the area to the left of -2 using the standard normal table directly.

$P(Z < -2) = P(Z < -2.00) = 0.0228$. Graph showing the area can be found in Panel (I) of Figure 5.6.

2. Draw a graph to show and find $P(Z > 2)$.

This is the area to the right of 2. Recall that the table gives the area to the left of a z -score. There are two ways to answer this question:

- Apply the symmetry property of the standard normal curve. The standard normal curve is symmetric at 0, and the area to the right of 2 equals the area to the left of -2.

$$P(Z > 2) = P(Z < -2) = 0.0228.$$
- Use the property of a density curve: all density curves have an area of one under the curve. The area to the right of 2 equals one minus the area to the left of 2.

$$P(Z > 2) = 1 - P(Z < 2) = 1 - P(Z < 2.00) = 1 - 0.9772 = 0.0228.$$
 Graph showing the area can be found in Panel (2) of Figure 5.8.

3. Draw a graph to show and find $P(Z < -2 \text{ or } Z > 2)$.

Area beyond -2 and 2, i.e., to the left of -2 or to the right of 2. The two events $\{Z < -2\}$ and $\{Z > 2\}$ don't overlap and, hence, are mutually exclusive; the special addition rule applies.

$$P(Z < -2 \text{ or } Z > 2) = P(Z < -2) + P(Z > 2) = 0.0228 + 0.0228 = 0.0456.$$

Graph showing the area can be found in Panel (3) of Figure 5.8.

4. Draw a graph to show and find $P(-4 < Z < 5)$.

The area between -4 and 5 equals the area to the left of 5 minus the area to the left of -4.

$$P(-4 < Z < 5) = P(Z < 5) - P(Z < -4) = 1 - 0 = 1.$$
 Graph showing the area can be found in Panel (4) of Figure 5.8.

Note: the standard normal table gives the area (in four decimal places) to the left of the z-score between -3.90 and 3.90. Therefore, the area to the left of any z-score below -3.90 is 0, and the area to the left of any z-score above 3.90 is 1.

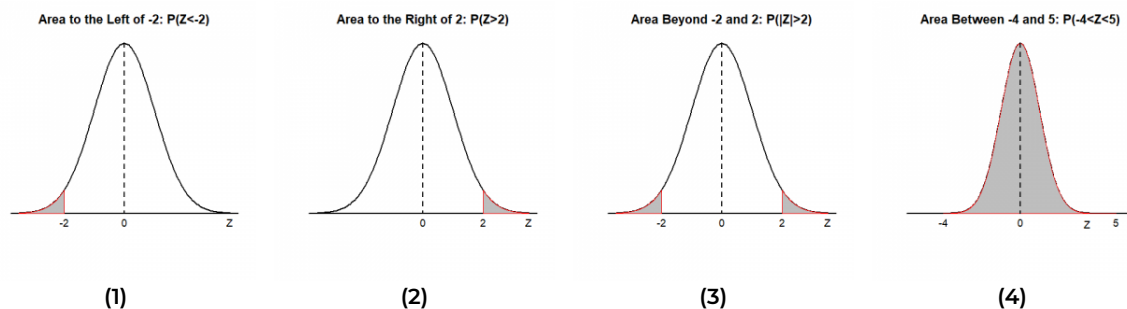


Figure 5.8: Graphs showing the areas under the standard normal curve corresponding to the probabilities in the example. [\[Image Description \(See Appendix D Figure 5.8\)\]](#) Click on the image to enlarge it.

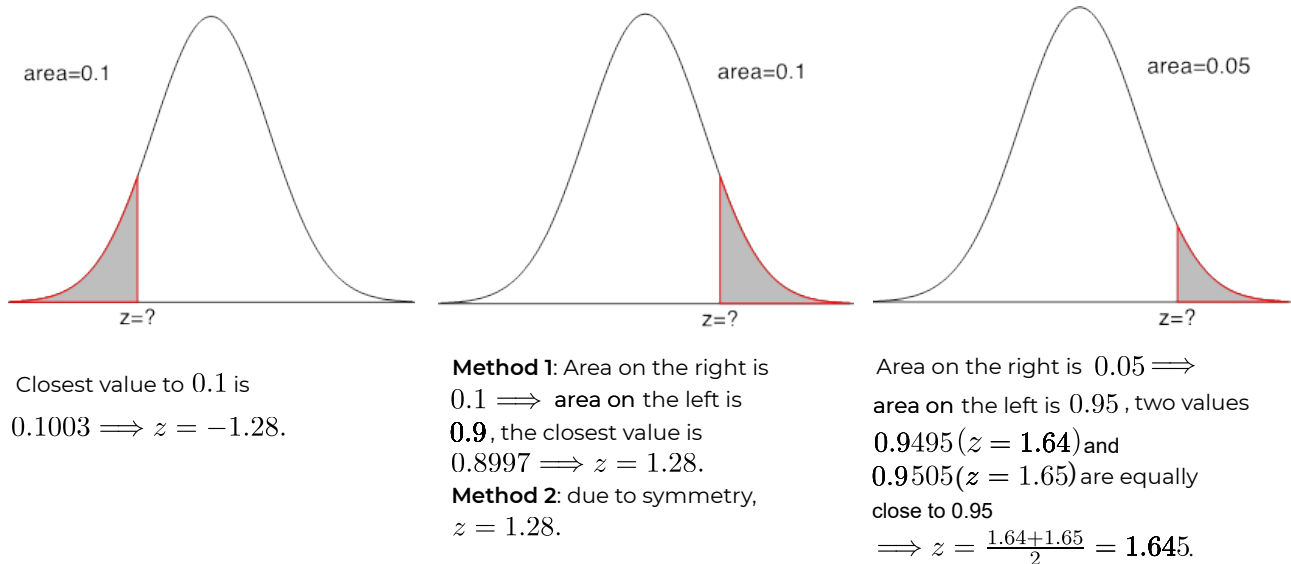
5.4.2 Find the Z-Score for a Given Area (Probability)

We can use the standard normal table in another way: find the z -score for a specified area or probability (percentage). The steps are as follows:

1. Express the given area in terms of a left-tailed probability (or probabilities if there are 2 z-scores).
2. Search the main body of the standard normal table for the closest value to the left-tailed probability.
3. Obtain the z -score that corresponds to the given area. If multiple values are equally close to the given left-tailed probability, take the average of their corresponding z -scores.

Example: Given the Area, find the corresponding z-score

Find the z -score corresponding to the shaded area in each graph:



[\[Image Description \(See Appendix D Example 5.1\)\]](#)

The notation z_α (read as z alpha) has a special meaning: the z -score that has an area of α to its **right** under the standard normal curve. The middle and right panels of the example above show that $z_{0.1} = 1.28$ and $z_{0.05} = 1.645$.

Exercise: Finding Z-score Z_α

Use the standard normal table to find

- a. $Z_{0.25}$: z-score with an area of 0.25 to its right
- b. $Z_{0.6}$: z-score with an area of 0.6 to its right
- c. $Z_{0.005}$: z-score with an area of 0.005 to its right

Show/Hide Answer

- a. $Z_{0.25}$: z-score with an area of 0.25 to its right, so the area to the left of $Z_{0.25}$ is $1-0.25=0.75$. Search the main body of the table; the closest value to 0.75 is 0.7486, which corresponds to the z-score 0.67; therefore, $Z_{0.25} = 0.67$.
- b. $Z_{0.6}$: z-score with an area of 0.6 to its right, so the area to the left of $Z_{0.6}$ is $1-0.6=0.4$. Search the main body of the table; the closest value to 0.4 is 0.4013, which corresponds to the z-score -0.25; therefore, $Z_{0.6} = -0.25$.
- c. $Z_{0.005}$: z-score with an area of 0.005 to its right, so the area to the left of $Z_{0.005}$ is $1-0.005=0.995$. Search the main body of the table for 0.995. Two values that are equally close to 0.995 are 0.9949 and 0.9951; the corresponding z-scores are 2.57 and 2.58, respectively. Therefore, $Z_{0.005} = \frac{2.57+2.58}{2} = 2.575$.

5.5 Working With Any Normal Distribution

Through standardization, we can solve problems related to any normal distribution $N(\mu, \sigma)$ using the standard normal table. The following diagram shows the procedure.

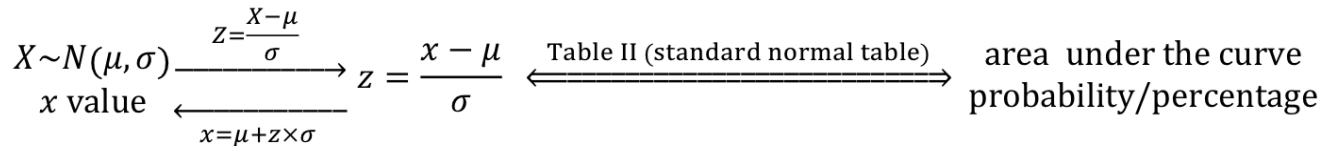


Figure 5.9: Diagram for Working with Any Normal Distribution [[Image description \(See Appendix D Figure 5.9\)](#)]

Example: Working With Any Normal Distribution Using the Standard Normal Table (Table II)

Suppose the grades in a Statistics class are approximately normally distributed with a mean 70 and a standard deviation 10, i.e., $X \sim N(\mu = 70, \sigma = 10)$.

- a. Find the percentage of students whose grades are below 60.

$$P(X < 60) = P\left(\frac{X - \mu}{\sigma} < \frac{60 - \mu}{\sigma}\right) = P\left(Z < \frac{60 - 70}{10}\right) = P(Z < -1) = 0.1587.$$

15.87% of the students have a grade below 60.

- b. Find the percentage of students whose grades are above 95.

$$\begin{aligned} P(X > 95) &= P\left(\frac{X - \mu}{\sigma} > \frac{95 - \mu}{\sigma}\right) \\ &= P\left(Z > \frac{95 - 70}{10}\right) \\ &= P(Z > 2.5) \\ &= 1 - P(Z < 2.5) \\ &= 1 - 0.9938 = 0.0062. \end{aligned}$$

$$\text{or } P(X > 95) = P(Z > 2.5) = P(Z < -2.5) = 0.0062.$$

Only 0.62% of students have a grade above 95.

- c. Find the percentage of students whose grades are between 60 and 90. The z -score for 60 is

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 70}{10} = -1; \text{ the } z\text{-score for 90 is } z = \frac{x - \mu}{\sigma} = \frac{90 - 70}{10} = 2.$$

$$\begin{aligned}
P(60 < X < 90) &= P(-1 < Z < 2) \\
&= P(Z < 2) - P(Z < -1) \\
&= 0.9772 - 0.1597 \\
&= 0.8185.
\end{aligned}$$

81.85% of the students have a grade between 60 and 90.

- d. Suppose the bottom 5% of students fail. Find the minimum grade needed in order to pass the course.

Find the grade x that bounds the bottom 5% of grades, i.e., $P(X \leq x) = 0.05$. The z -score that captures the bottom 5% of the standard normal distribution is $z = -1.645$. Therefore, the corresponding x value is

$$x = \mu + z \times \sigma = 70 + (-1.645) \times 10 = 70 - 16.45 = 53.55.$$

The passing grade is 53.55.

- e. Suppose the top 2% of students will get an A. Find the minimum grade needed in order to obtain an A.

Find the grade x that bounds the top 2% of grades, i.e., $P(X \geq x) = 0.02$. The z -score that captures the top 2% of the standard normal distribution is $z_{0.02} = 2.05$. Therefore, the corresponding x value is

$$x = \mu + z \times \sigma = 70 + (2.05) \times 10 = 70 + 20.5 = 90.5.$$

The cutoff is 90.5.

- f. Find the quartiles of student grades.

First, observe that the first, second and third quartiles of the standard normal distribution are $z_1 = -0.67$, $z_2 = 0$ and $z_3 = 0.67$, since $P(Z < -0.67) \approx 0.25$, $P(Z < 0) = 0.5$, and $P(Z < 0.67) \approx 0.75$. Thus, the first, second and third quartiles of student grades are:

$$Q_1 = \mu + z_1 \times \sigma = 70 + (-0.67) \times 10 = 63.3,$$

$$Q_2 = \mu + z_2 \times \sigma = 70 + (0) \times 10 = 70,$$

$$Q_3 = \mu + z_3 \times \sigma = 70 + (0.67) \times 10 = 76.7.$$

Thus, the quartiles of student grades are:

$$Q_1 = 63.3, \quad Q_2 = 70, \quad Q_3 = 76.7.$$

Note: Recall that one of the properties of a symmetric distribution is that the mean and median are equal. So, we could have alternatively used the symmetry of a normal distribution to establish that $Q_2 = \mu = 70$.

Exercise: Work With Any Normal Distribution Using Standard Normal Table (Table II)

Suppose the weight of boxes of a certain brand of cereal follows a normal distribution with a mean of 1,000 grams and a standard deviation of 40 grams.

- A box is rejected by the quality control department if its weight is below 950 grams. What percentage of boxes will be rejected?
- Find the percentage of boxes with weight in between 980 grams and 1,010 grams.
- Find the percentage of boxes with weight above 1,010 grams.
- Determine the 40th percentile for the weight of the boxes of cereal.
- A particular box of cereal is weighed, and it is determined that 5% of boxes are heavier than this particular box. What is the approximate weight of this box of cereal?

Show/Hide Answer

- $P(X < 950) = P(Z < (950 - 1000)/40) = P(Z < -1.25) = 0.1056$. That is 10.56%.
- $P(980 < X < 1010) = P(-0.5 < Z < 0.25) = P(Z < 0.25) - P(Z < -0.5)$
 $= 0.5987 - 0.3085 = 0.2902$, that is, 29.02%.
- $P(X > 1010) = P(Z > (1010 - 1000)/40) = P(Z > 0.25)$
 $= 1 - P(Z < 0.25) = 1 - 0.5987 = 0.4013$, that is, 40.13%.
- The 40th percentile of the standard normal distribution is $z = -0.25$. Therefore,
 $x = 1000 + (-0.25) \times 40 = 990$, so that the 40th percentile among boxes of cereal is 990 grams.
- The 95th percentile of the standard normal distribution is $z = 1.645$. Therefore,
 $x = 1000 + (1.645) \times 40 = 1065.8$, so that the 95th percentile among boxes of cereal is 1065.8 grams.

5.6 Assessing Normality: Normal Probability Plot

In later chapters, it will be necessary for us to assume our sample is selected from a normally distributed population; an easy way to check this assumption is to do so via graphical methods. For example, we can construct a histogram of our sampled data and if the histogram looks to be somewhat bell-shaped, then it is reasonable for us to assume the population is normally distributed (or at least approximately normally distributed). However, histograms only tend to inherit features of the population when the sample size is reasonably large.

A more effective alternative to a histogram is a normal probability plot, which plots observed data points against normal quantiles (for this reason, normal probability plots are often referred to as normal Q-Q plots, where “Q” stands for “Quantile.”). If the distribution of the data is roughly normal, the points on a normal probability plot will roughly fall on a straight line. Deviations from a straight line indicate that the underlying distribution is not normal.

Typically, software such as R commander is used to make a normal probability plot. Some software plots the observed quantiles in the y-axis by default (e.g., R), and some plots the normal quantiles in the y-axis (e.g., Minitab). The steps to draw a normal probability plot are illustrated in the following example.

Example: Assessing Normality Using Normal Probability Plot

Suppose the data are: 75, 80, 90, 85, 75, and 40. Check whether the data are from a normal distribution by drawing a normal probability plot.

Steps:

1. Sort the data from smallest to largest. We refer to the sorted data as the observed quantiles and put them in the first column of a table.
2. Refer to a table of normal scores (such as Table III in the appendix of the course textbook) in order to find the normal quantiles (sometimes called theoretical quantiles). In this example, there are $n = 6$ data points and so we copy the column with $n = 6$ into the second column of our table.
3. Plot the observed quantiles (y-axis) versus the theoretical quantiles (x-axis) or the other way.

4. If the data points roughly fall on a straight line, then we assume the data are from a normal distribution; otherwise, the data are not from a normal distribution.

Table 5.2: Observed and Theoretical Quantiles for a Normal Q-Q plot

Sorted value (observed quantile)	Normal score (theoretical quantile)
40	-1.28
75	-0.64
75	-0.20
80	0.20
85	0.64
90	1.28

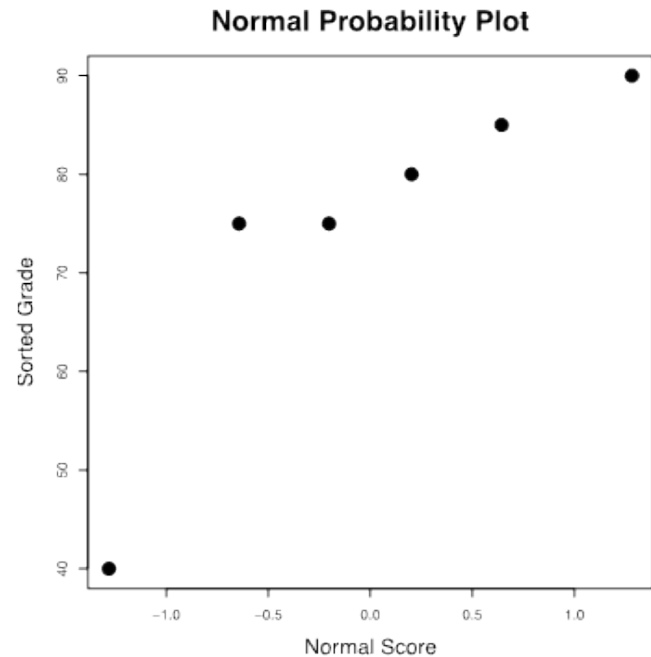


Figure 5.10: Normal Probability Plot for Six Grades. [\[Image Description \(See Appendix D Figure 5.10\)\]](#)

The six points do not fall in a straight line; the data do not seem to come from a normal distribution. However, the point on the lower left corner might be an outlier. If we remove this potential outlier, the other five points roughly fall on a straight line.

Exercise: Normal Probability Plot

Comment on the following normal probability plots and answer whether the data seem to come from a normal distribution.

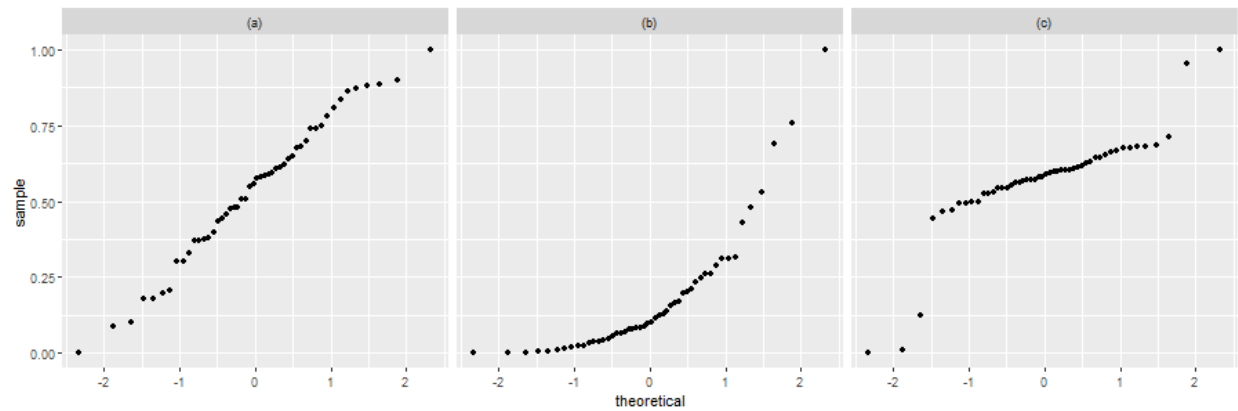
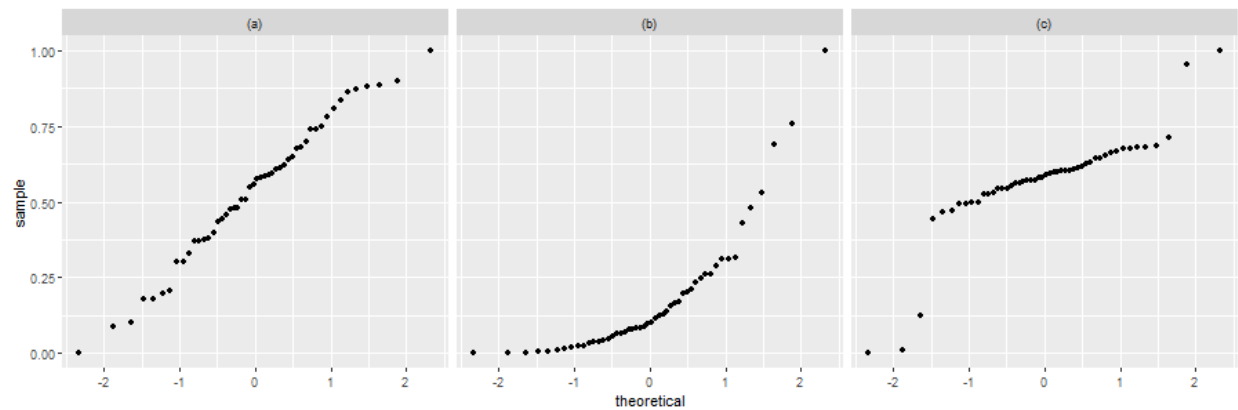


Figure 5.11: Example of Normal Probability Plots. [[Image Description \(See Appendix D Figure 5.11\)](#)]

Show/Hide Answer



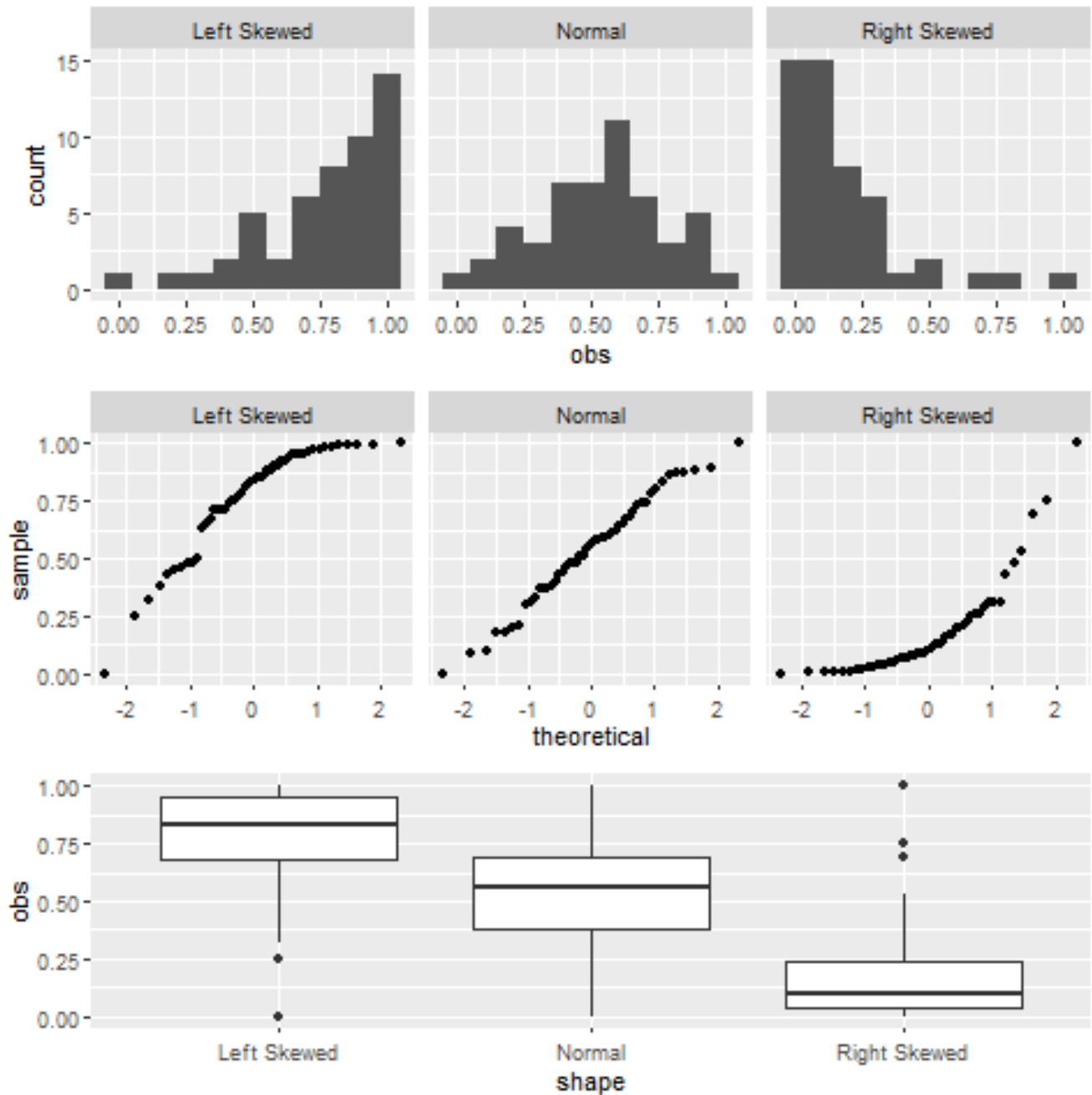
The points form an approximate straight line. Thus, it is reasonable to assume the data are from a normal distribution.

The points do not form a straight line (there is obvious curvature). This suggests that the data are not from a normal distribution.

Excluding the outlier, the points form an approximate straight line. Thus, it is reasonable to assume the data are from a normal distribution (if we disregard the outlier).

Histogram, boxplot and normal probability (Q-Q) plot are popular graphs used to explore the distribution of data. If the data are taken from a normal population, the histogram should appear to be bell-shaped, the boxplot should be symmetric, the normal probability plot should show a linear pattern. When the number of observations is not large, however, the histogram might not show bell-shape with different bin widths. Note that a boxplot cannot be used to confirm that data follow a normal distribution since some distributions, such as uniform and multimodal, are also symmetric. Therefore, the normal probability plot is the best graphical method to assess normality. Figure 5.12 shows histograms, normal probability plots, and boxplots for six typical distributions: left skewed, normal, right

skewed, multimodal and symmetric, normal with outliers, and uniform. Based on the graphs, we can see how histogram features correspond to boxplot and normal probability features.



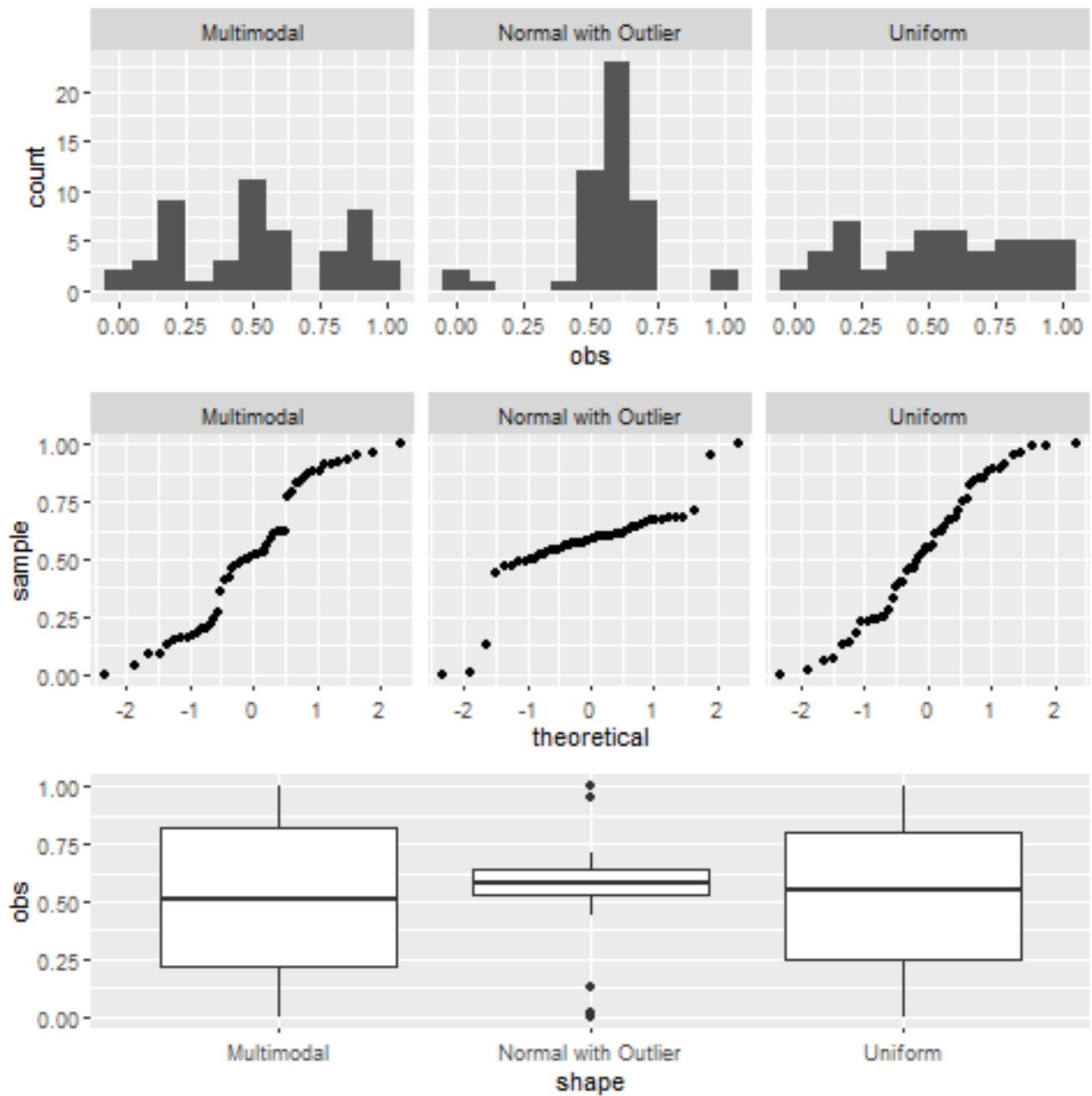


Figure 5.12: Histograms, Normal Probability Plots and Boxplot for Six Typical Distributions. [[Image Description \(See Appendix D Figure 5.12\)](#)]

5.7 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Describe the properties of a normal density curve (Section 5.2).
- Describe the standard normal distribution (Section 5.3).
- Use the standard normal table (Table II) in order to:
 - Determine the area bounded by some given values under any normal density curve (Section 5.5).
 - Find values that correspond to given areas under any normal density curve (Section 5.5).
- Use a normal probability plot to assess whether a given data set seems to come from a normal population (Section 5.6).

5.8 Review Questions

1. The standard normal distribution has mean _____ and standard deviation _____.
2. The area under the density curve that lies between 30 and 40 is 0.832. What percentage of all possible observations of the variable are either less than 30 or greater than 40?
3. The finishing times in the New York City 10-km run are normally distributed with mean 61 minutes and standard deviation 9 minutes.
 - a. Determine the percentage of finishers with times between 50 and 70 minutes.
 - b. Determine the percentage of finishers with times less than 75 minutes.
 - c. Obtain and interpret the 40th percentile for the finishing times.
 - d. Find and interpret the 8th decile for the finishing times.

Show/Hide Answer

1. The standard normal distribution has a mean **0** and standard deviation **1**.
2. Since the total area under any density curve is 1, the total area to the left of 30 and to the right of 40 is $1 - 0.832 = 0.168$. The percentage of observations that are either less than 30 or greater than 40 is 16.8%.
3.
 - a. Want to find the area between 50 and 70, we need to find the z-scores first and then use Table II.

$$\begin{aligned} P(50 < X < 70) &= P\left(\frac{50 - 61}{9} < Z < \frac{70 - 61}{9}\right) = P(-1.22 < Z < 1) \\ &= P(Z < 1) - P(Z < -1.22) = 0.8413 - 0.1112 = 0.7301. \end{aligned}$$

There are 73.01% of finishers with times between 50 and 70 minutes. We could also double check with R commander, the results are close. The results differ since we round the z score to two decimals and use the table.

```
## [1] 0.8413447 0.1108118
```

```
## [1] 0.7305329
```

- b. If you want to find the area to the left of 75, find the z-score first and then use Table II to find the area.

$$P(X < 75) = P(Z < \frac{75 - 61}{9}) = P(Z < 1.56) = 0.9406.$$

There are 94.06% of finishers with times less than 75 minutes. We could also double check with R commander, the results are close.

```
## [1] 0.9400931
```

- c. Want to know 40% of finishers with times below what value. First, use Table II to find the z-score with the area to its left is 0.4. The closest value in the main body of Table II is 0.4013 with the corresponding z-score of -0.25. Therefore, the normal quantile is

$$x = \mu + z \times \sigma = 61 + (-0.25) \times 9 = 58.75.$$

Interpretation: There are 40% of finishers with times less than 58.75 minutes.

We could also double check with R commander, the results are close.

```
## [1] 58.71988
```

- d. 8th decile is the same as the 80th percentile. Want to know 80% of finishers with times below what value. First, use Table II to find the z-score with the area to its left is 0.8. The closest value in the main body of Table II is 0.7995, with the corresponding z-score of 0.84.

$$\text{Therefore, the normal quantile is } x = \mu + z \times \sigma = 61 + 0.84 \times 9 = 68.56.$$

Interpretation: 80% of finishers have times less than 68.56 minutes.

We could also double check with R commander, the results are close.

```
## [1] 68.57459
```

5.9 Assignment 5

Purposes

This assignment has two parts. The first part assesses your knowledge of the properties of normal density curves, calculating probabilities related to normal distributions, and finding the quantiles of normal distributions. The second part assesses your skills in using R commander to find probabilities and quantiles of normal distributions.

Resources

[M05__MonthlyFees_labQ4.xlsx](#)

Instructions

Part A

Complete the following:

1. The standard normal distribution has mean ____ and standard deviation _____. (2 marks)
2. The area under the density curve between 30 and 40 is 0.832. What percentage of all possible observations of the variable are either less than 30 or greater than 40? (2 marks)
3. A curve has an area of 0.425 to the left of 4 and an area of 0.685 to the right of 4. Could this curve be a density curve for some variable? Explain your answer. (3 marks)
4. Determine the z -scores $z_{0.03}$ and $z_{0.005}$. (4 marks)
5. The finishing times in the New York City 10-km run are normally distributed with a mean of 61 minutes and a standard deviation of 9 minutes.
 - a. Determine the percentage of finishers with times between 50 and 70 minutes. (4 marks)
 - b. Determine the percentage of finishers with times less than 75 minutes. (3 marks)
 - c. Obtain and interpret the 40th percentile for the finishing times. (4 marks: 3+1)
 - d. Find and interpret the 8th decile for the finishing times. (4 marks: 3+1)

6. The weight of boxes of cereal follows a normal distribution with a mean of 1,000 grams and a standard deviation of 40 grams.
- The quality control department rejects a box if its weight is below 950 grams. What percentage of boxes will be rejected? (4 marks)
 - Find the percentage of boxes weighing 980 grams and 1,010 grams. (4 marks)
 - Find the percentage of boxes whose weight is above 1,010 grams. (4 marks)
 - Determine the 40th percentile for the weight of the boxes of cereal. (4 marks)
 - Determine the weight so that only 5% of boxes are heavier than the weight you choose. (4 marks)
 - Randomly pick five boxes, find the probability that at least one box is rejected. (6 marks)

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs into the space below each question, and write down your answers in statements.

- Use R commander to find the z -scores $z_{0.03}$ and $z_{0.005}$. (4 marks)
- The finishing times in the New York City 10-km run are normally distributed with a mean of 61 minutes and a standard deviation of 9 minutes.
 - Determine the percentage of finishers with times between 50 and 70 minutes. (3 marks)
 - Determine the percentage of finishers with times less than 75 minutes. (2 marks)
 - Obtain the 40th percentile for the finishing times. (2 marks)
 - Find the 8th decile for the finishing times. (2 marks)
- The weight of boxes of cereal follows a normal distribution with a mean of 1,000 grams and a standard deviation of 40 grams.
 - The quality control department rejects a box if its weight is below 950 grams. What percentage of boxes will be rejected? (2 marks)
 - Find the percentage of boxes weighing 980 and 1,010 grams. (3 marks)
 - Find the percentage of boxes whose weight is above 1,010 grams. (2 marks)
 - Determine the 40th percentile for the weight of the boxes of cereal. (2 marks)
 - Determine the weight such that only 5% of boxes are heavier than that. (2 marks)
 - Randomly pick five boxes, find the probability that at least one box is rejected. (4 marks)
- The monthly fees, in dollars, for a sample of the providers and plans are as follows: 40, 110, 90, 30, 70, 70, 30, 60, 60, 50, 60, 70, 35, 80, 75. Use R commander to assess whether the data follow a normal distribution. Is there any outlier? The data can be found in

the file M05_MonthlyFees_labQ4.xlsx. (5 marks)

Quiz 5



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2625#h5p-7>

CHAPTER 6: DISTRIBUTION OF THE SAMPLE MEAN AND THE CENTRAL LIMIT THEOREM

Overview

There are two types of statistics: descriptive and inferential. Chapters 1 and 2 focus on descriptive statistics; chapters 3, 4, and 5 are about probability (topics from these chapters will reappear in later chapters); chapters seven and all chapters after that will focus on inferential statistics. This chapter introduces the distribution of the sample mean and the central limit theorem, which provides a foundation for inferential statistics.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the difference between a statistic and a parameter.
- Explain the relationships among a parameter, an estimator, and a point estimate.
- Describe how to obtain the sampling distribution of the sample mean.
- Explain and describe the distribution of sample mean in three aspects: mean, standard deviation, and shape.
- Explain the central limit theorem in plain English.
- Apply the central limit theorem to answer questions about the sample mean.

6.1 Parameter and Statistic

We first review the relationship between parameters and statistics.

A **parameter** is a constant (usually unknown) used to describe some aspect of a **population**. For example, the population mean μ is the average value of a characteristic of interest for all individuals in a population (such as the average height of all individuals in Canada).

A **statistic** describes some aspect of a **sample**, much like a parameter describes some aspect of a population. However, unlike a parameter (the value of which is assumed to be constant), the value of a statistic varies from one sample to the next. Thus, we generally view a statistic as a random variable before obtaining a random sample, while we view a statistic as a fixed number after the data are collected.

For example, consider the sample mean \bar{X} , which is the average of all individuals in a sample. If we have not yet obtained a random sample, then the value of X has not yet been realized, and therefore, we can view \bar{X} as a random variable whose value depends on which individuals from the population end up in our sample. However, once we have obtained our sample and computed the value of the sample average, we now have a fixed number, denoted as \bar{x} . That is, \bar{X} denotes the sample average when viewed as a random variable. In contrast, \bar{x} denotes a particular realization of \bar{X} , which depends upon the actual data we have obtained.

Suppose there are N individuals in a population from which we plan to obtain a simple random sample of size n . If we wish to compute the value of a statistic, then it is of interest to know what values the statistic may assume and their corresponding frequencies. If we consider all possible samples of size n , there are a total of $\binom{N}{n} = {}_N C_n$ (N choose n) distinct samples, many of which give different values of the statistic. Drawing a histogram of these $\binom{N}{n}$ values gives the **sampling distribution** of the statistic. We can describe the distribution of the statistic with its mean, standard deviation, and shape.

Sometimes, we use a statistic to estimate the value of a population parameter; we call the statistic an **estimator** of the parameter. If sample data are obtained, and the value of the estimator (statistic) is computed, then we refer to this observed value as a **point estimate** of the population parameter. For example, the sample mean \bar{X} is an estimator of the population mean μ and a value of the random variable X , denoted as \bar{x} , is a point estimate of μ . Because \bar{x} is based on a sample of size n rather than on the entire population, it is generally the case that \bar{x} is not equal to μ . The difference between the point estimate and the parameter is called the **sampling error**. We call an estimator **unbiased** if the average

of the $\binom{N}{n}$ values of the estimate is equal to the population parameter. For example, the sample mean \bar{X} is an unbiased estimator of the population mean μ , i.e., the mean of the sample mean equals the population mean.

6.2 Distribution of the Sample Mean

Suppose the variable of interest is X and the population consists of N individuals. The possible values of X are the different measurements for each individual in the population. For example, suppose the variable of interest is X =height and the population is the $N = 60$ students in our class. The number $N = 60$ is called the **population size**. Suppose we measure each student's height and draw a histogram of those $N = 60$ measurements. In that case, the resulting distribution is the **population distribution**, that is, the distribution of the random variable X . The average height of all 60 students is the population mean μ .

We often use the sample mean \bar{X} to estimate the population mean μ . However, since the observed value of \bar{X} varies from sample to sample, it is helpful to know the typical accuracy of this estimator. For example, how confident are we that the error in estimating μ by \bar{x} is at most 2 cm? To answer this kind of question, we need to know the distribution of the sample mean \bar{X} .

For a population of size N , if we take a sample of size n , there are $\binom{N}{n}$ distinct samples, each of which gives one possible value of the sample mean \bar{x} . The $\binom{N}{n}$ values of \bar{x} give the distribution of the sample mean \bar{X} , which is also called the sampling distribution of the sample mean. A histogram of the $\binom{N}{n}$ values of \bar{x} shows the distribution of \bar{X} . However, $\binom{N}{n}$ is often so large that we are unable to consider all possible samples of size n directly. Fortunately, we can still obtain a reasonable approximation of the distribution of \bar{X} by obtaining a large number of random samples, say 10,000, computing each sample mean, and drawing a histogram based on our sample of the sample means. For example, if the population size is $N = 60$ and the sample size is $n = 5$, there are $\binom{60}{5} = 5,461,512$ different samples, many of which have different values of \bar{x} . Drawing a histogram of these 5,461,512 \bar{x} values gives the distribution of the sample mean \bar{X} , with sample size $n = 5$. Moreover, the sampling distribution of the sample mean \bar{X} can be described in three aspects: centre, spread (variation), and shape.

6.2.1 Mean and Standard Deviation of the Sample Mean

Let's consider a population consisting of 5 students. Suppose their heights (in cm) are

$x_1 = 155, x_2 = 165, x_3 = 175, x_4 = 185, x_5 = 195$. The population size is $N=5$ and the population mean μ and population standard deviation σ are:

$$\begin{aligned}\mu &= \frac{\sum x_i}{N} \\ &= \frac{155 + 165 + 175 + 185 + 195}{5} \\ &= 175, \\ \sigma &= \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \\ &= \sqrt{\frac{(155 - 175)^2 + (165 - 175)^2 + (175 - 175)^2 + (185 - 175)^2 + (195 - 175)^2}{5}} \\ &= 14.14.\end{aligned}$$

Consider a simple random sample of size $n = 2$, which means randomly picking two students from this population of five students. $n = 2$ is called the **sample size**. The number of ways we can pick two students out of five is ${}_5C_2 = \binom{5}{2} = 10$. For example, one possible sample is $\{x_1, x_2\}$ which gives a value of the sample mean,

$$\bar{x} = \frac{x_1 + x_2}{2} = \frac{155 + 165}{2} = 160.$$

Another possible sample is $\{x_1, x_3\}$ and the corresponding value of the sample mean is:

$$\bar{x} = \frac{x_1 + x_3}{2} = \frac{155 + 175}{2} = 165.$$

Table 6.1 lists all possible samples of sample size $n = 2, 3, 4$ and their corresponding sample mean values. The mean and standard deviation of the sample mean of all possible sample sizes are also given in the table.

$n = 2$		$n = 3$		$n = 4$	
Possible Samples	\bar{x}	Possible Samples	\bar{x}	Possible Samples	\bar{x}
x_1, x_2	160	x_1, x_2, x_3	165	x_1, x_2, x_3, x_4	170
x_1, x_3	165	x_1, x_2, x_4	168.33	x_1, x_2, x_3, x_5	172.5
x_1, x_4	170	x_1, x_2, x_5	171.67	x_1, x_2, x_4, x_5	175
x_1, x_5	175	x_1, x_3, x_4	171.67	x_1, x_3, x_4, x_5	177.5
x_2, x_3	170	x_1, x_3, x_5	175	x_2, x_3, x_4, x_5	180
x_2, x_4	175	x_1, x_4, x_5	178.33		
x_2, x_5	180	x_2, x_3, x_4	175		
x_3, x_4	180	x_2, x_3, x_5	178.33		
x_3, x_5	185	x_2, x_4, x_5	181.67		
x_4, x_5	190	x_3, x_4, x_5	185		
mean of \bar{x} , $\mu_{\bar{x}} = 175$		mean of \bar{x} , $\mu_{\bar{x}} = 175$		mean of \bar{x} , $\mu_{\bar{x}} = 175$	
SD of \bar{x} , $\sigma_{\bar{x}} = 8.66$		SD of \bar{x} , $\sigma_{\bar{x}} = 5.77$		SD of \bar{x} , $\sigma_{\bar{x}} = 3.54$	

Table 6.1: Sample Means of All Possible Samples of Sample Size $n=2, 3, 4$. [Image Description (See Appendix D Table 6.1)]

The mean and standard deviation of the sample mean \bar{X} are denoted as $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ respectively. When the sample size $n = 2$, Table 6.1 shows 10 possible values of the sample mean: 160, 165, \dots , 185, 190; there is one value of 160 and two values of 180, giving the probabilities of $\frac{1}{10}$ and $\frac{2}{10}$ observing these two values respectively. The probability distribution and distribution histogram of the sample mean \bar{X} with $n = 2$ are:

\bar{x}	$P(\bar{X} = \bar{x})$
160	$\frac{1}{10} = 0.1$
165	$\frac{1}{10} = 0.1$
170	$\frac{2}{10} = 0.2$
175	$\frac{2}{10} = 0.2$
180	$\frac{2}{10} = 0.2$
185	$\frac{1}{10} = 0.1$
190	$\frac{1}{10} = 0.1$

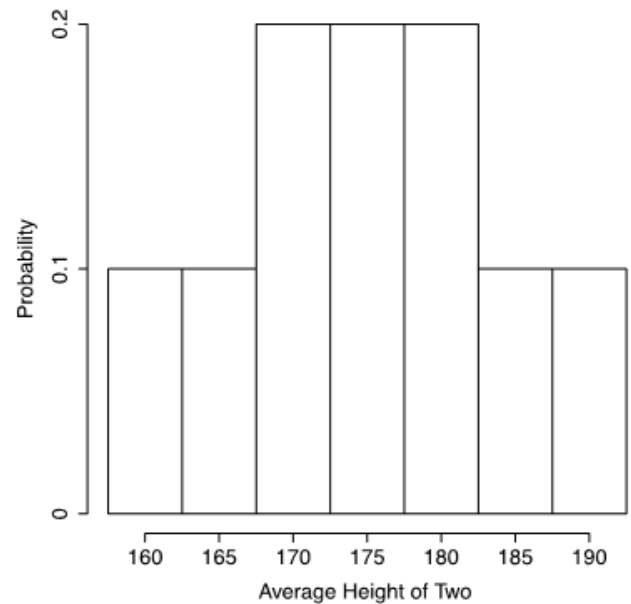


Figure 6.1: Probability Distribution and Probability Histogram of Sample Mean for $n=2$. [Image Description (See Appendix D Figure 6.1)]

The mean and the standard deviation of the sample mean with $n = 2$ are:

$$\begin{aligned}
\mu_{\bar{X}} &= \frac{160 + 165 + 170 + 175 + 170 + 175 + 180 + 180 + 185 + 190}{10} \\
&= 175, \\
\sigma_{\bar{X}} &= \sqrt{\frac{\sum (\bar{x} - \mu_{\bar{X}})^2}{N}} \\
&= \sqrt{\frac{(160 - 175)^2 + (165 - 175)^2 + \dots + (185 - 175)^2 + (190 - 175)^2}{10}} \\
&= 8.66.
\end{aligned}$$

When the sample size is $n = 3$, the mean and the standard deviation of the sample mean are:

$$\begin{aligned}
\mu_{\bar{X}} &= \frac{165 + 168.33 + 171.67 + 171.67 + 175 + 178.33 + 175 + 178.33 + 181.67 + 185}{10} \\
&= 175, \\
\sigma_{\bar{X}} &= \sqrt{\frac{\sum (\bar{x} - \mu_{\bar{X}})^2}{N}} \\
&= \sqrt{\frac{(160 - 175)^2 + (168.33 - 175)^2 + \dots + (185 - 175)^2 + (190 - 175)^2}{10}} \\
&= 5.77.
\end{aligned}$$

When the sample size is $n = 4$, the mean and the standard deviation of the sample mean are:

$$\begin{aligned}
\mu_{\bar{X}} &= \frac{170 + 172.5 + 175 + 177.5 + 180}{5} \\
&= 175, \\
\sigma_{\bar{X}} &= \sqrt{\frac{\sum (\bar{x} - \mu_{\bar{X}})^2}{N}} \\
&= \sqrt{\frac{(170 - 175)^2 + (172.5 - 175)^2 + (175 - 175)^2 + (177.5 - 175)^2 + (180 - 175)^2}{5}} \\
&= 3.54.
\end{aligned}$$

The above results show that the mean of the sample mean equals the population mean regardless of the sample size, i.e., $\mu_{\bar{X}} = \mu$, while the standard deviation of the sample mean decreases when the sample size n increases. It can be shown that when sampling without replacement from a finite population, like those listed in Table 6.1,

$$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}.$$

If we instead sample with replacement from a finite population, the standard deviation of the sample mean is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Note: If we sample without replacement, $\sigma_{\bar{X}}$ is approximately equal to $\frac{\sigma}{\sqrt{n}}$, as long as the

sample size n is much smaller than the population size N . For simplicity of notation, we only focus on the sample without replacement case for the distribution of the sample mean in the remaining chapters.

Key Facts: Mean and Standard Deviation of the Sample Mean \bar{X}

For samples of size n ,

- The mean of the sample mean \bar{X} equals the population mean μ ; that is

$$\mu_{\bar{X}} = \mu.$$

- The standard deviation of the sample mean \bar{X} equals the population standard deviation σ divided by the square root of the sample size; that is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

These two arguments are always true for any population distribution and any sample size n .

Note: The standard deviation of the sample mean $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ implies that as sample size n increases, the standard deviation of the sample mean gets smaller. This is because the sample mean gets closer to the population mean and hence has a smaller variation when the sample size increases.

6.2.2 Shape of the Distribution of the Sample Mean (Central Limit Theorem)

We discuss the shape of the distribution of the sample mean for two cases: when the population distribution is normal, i.e., the variable of interest $X \sim N(\mu, \sigma)$ and when the population distribution is not normal.

When the Population is Normally Distributed

Suppose the random variables X_1, X_2, \dots, X_n represent a simple random sample from a normal population distribution $N(\mu, \sigma)$, then the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

also follows a normal distribution, regardless of the value of the sample size n . This is a

consequence of the fact that a **linear combination** of normal random variables is itself a normal random variable.

Example: Grade of 100 Students

Suppose a population consists of 100 students and the variable of interest is X = student grades. Due to bonus questions, the maximum grade might be above 100. The histogram of the grades of these 100 students gives the population (or parent) distribution, or simply the distribution of X . The mean and standard deviation of these 100 grades give the population mean and population standard deviation $\mu = 70, \sigma = 10$. It is reasonable for us to assume grades follow a normal distribution since the histogram is bell-shaped and the points in the QQ plot form an approximate straight-line pattern.

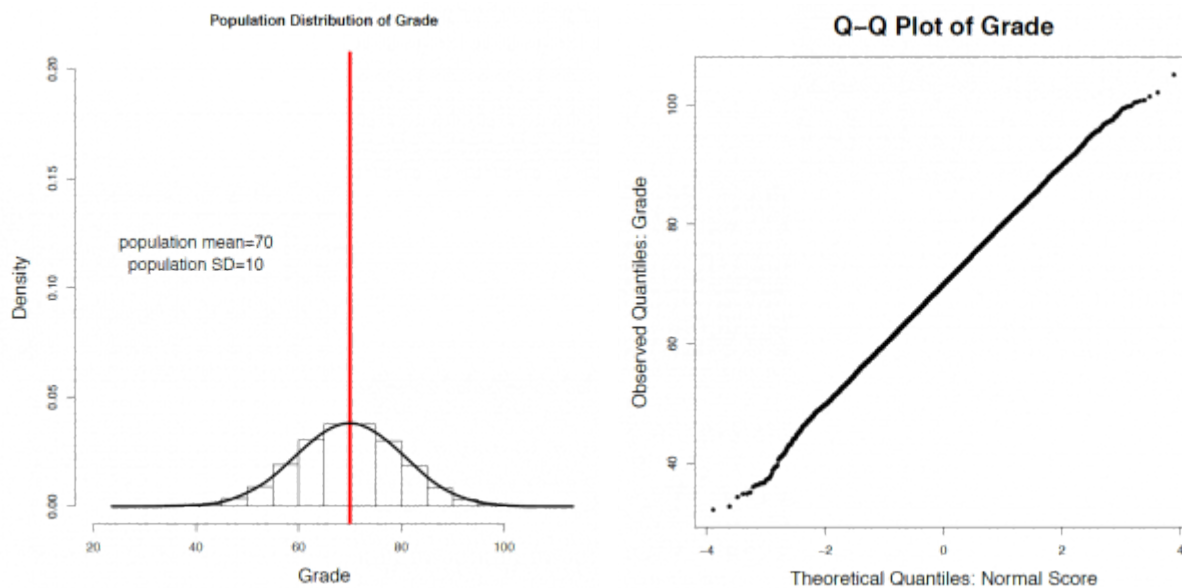


Figure 6.2: Density and Normal Probability Plot of Grade (Population). [[Image Description \(See Appendix D Figure 6.2\)](#)]

Let's examine the distributions of the sample mean \bar{X} for sample size $n = 2, 5, 30$. In each histogram, the red solid line indicates the population mean and the blue dashed line indicates the mean of the sample mean. Recall the steps to obtain the distribution of the sample mean:

1. Obtain a sample of size n from the population of 100 students and calculate the sample mean \bar{x} = average grade for this particular sample.
2. Repeat step 1 for each of the $\binom{100}{n} = {}_{100}C_n$ different samples to obtain $\binom{100}{n}$ sample means \bar{x} values.
3. Draw a histogram of those $\binom{100}{n}$ sample means.
4. If $\binom{100}{n}$ is too large, then we can approximate the distribution of the sample mean by performing

the above steps using a large number of random samples (say 10,000), instead of all $\binom{100}{n}$ samples. Note that the mean and standard deviation are $\mu_{\bar{X}} = \mu = 70, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{n}}$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{2}} = 7.07$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{5}} = 4.47$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{30}} = 1.83$$

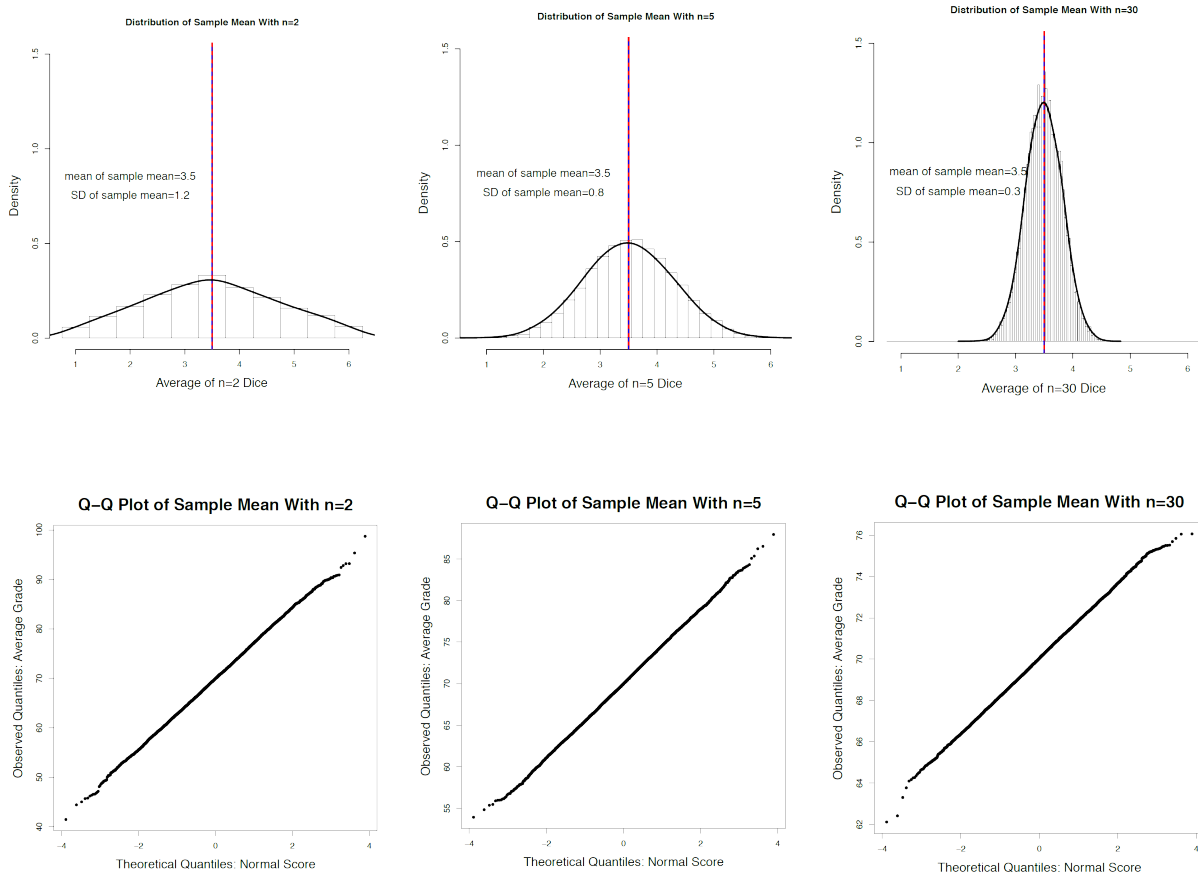


Figure 6.3: Density and Normal Probability Plot of the Average Grade (Sample Mean) for $n=2, 5, 30$. [\[Image Description \(See Appendix D Figure 6.3\)\]](#) Click on the image to enlarge it.

For each sample size, we can verify the following:

- The distribution of the sample mean \bar{X} is approximately normally distributed (symmetric, bell shape, unimodal);
- The mean of the sample mean equals the population mean of 70, and the standard deviation of the sample mean gets smaller and smaller when sample size n increases and roughly equals the population standard deviation divided by the square root of the sample size. Note that they are approximately equal because we have obtained 10,000 random samples for each sample size n ,

instead of all $\binom{100}{n} = {}_{100}C_n$ possible samples.

When the Population is not Normally Distributed

To illustrate two non-normal populations, we will discuss the uniform distribution (which is symmetric) and the exponential distribution (which is extremely right-skewed).

Example: Population Distribution is Uniform (Symmetric but not Normal)

Consider rolling a fair die. Since the die is fair, each face has the same chance to be observed; therefore, the population distribution is a uniform distribution with the following probability distribution.

Table 6.2: Working Table for the Population Mean and Standard Deviation

x	$P(X = x)$	$xP(X = x)$	$x^2P(X = x)$
1	1/6	1/6	$1^2 \times \frac{1}{6} = 1/6$
2	1/6	2/6	$2^2 \times \frac{1}{6} = 4/6$
3	1/6	3/6	$3^2 \times \frac{1}{6} = 9/6$
4	1/6	4/6	$4^2 \times \frac{1}{6} = 16/6$
5	1/6	5/6	$5^2 \times \frac{1}{6} = 25/6$
6	1/6	6/6	$6^2 \times \frac{1}{6} = 36/6$
sum=1		sum=21/ 6=3.5	sum=91/6

The population mean and standard deviation are calculated as follows:

$$\begin{aligned}
 \mu &= \sum xP(X = x) \\
 &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\
 &= 3.5, \\
 \sigma &= \sqrt{\sum x^2P(X = x) - \mu^2} \\
 &= \sqrt{\frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - 3.5^2} \\
 &= 1.71.
 \end{aligned}$$

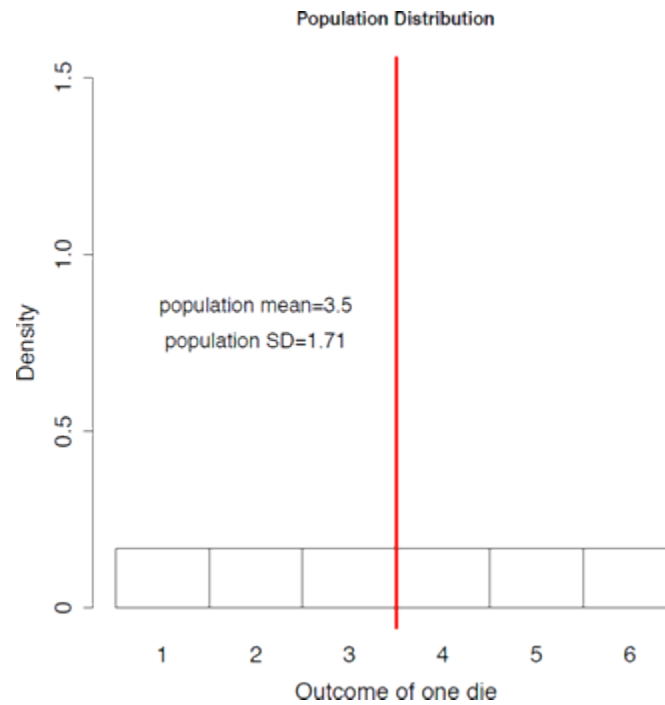


Figure 6.4: Density Curve of the Population X [\[Image Description \(See Appendix D Figure 6.4\)\]](#)

The uniform distribution is not bell-shaped and, hence, is not a normal distribution. Let's examine the distribution of the sample mean with sample sizes $n = 2, 5, 30$, that is, the distribution of the average of n rolls of a fair die. Note that the mean and standard deviation are: $\mu_{\bar{X}} = \mu = 3.5$; $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.71}{\sqrt{n}}$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.71}{\sqrt{2}} = 1.21$$

shape: triangular

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.71}{\sqrt{5}} = 0.76$$

shape: normal

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.71}{\sqrt{30}} = 0.31$$

shape: normal

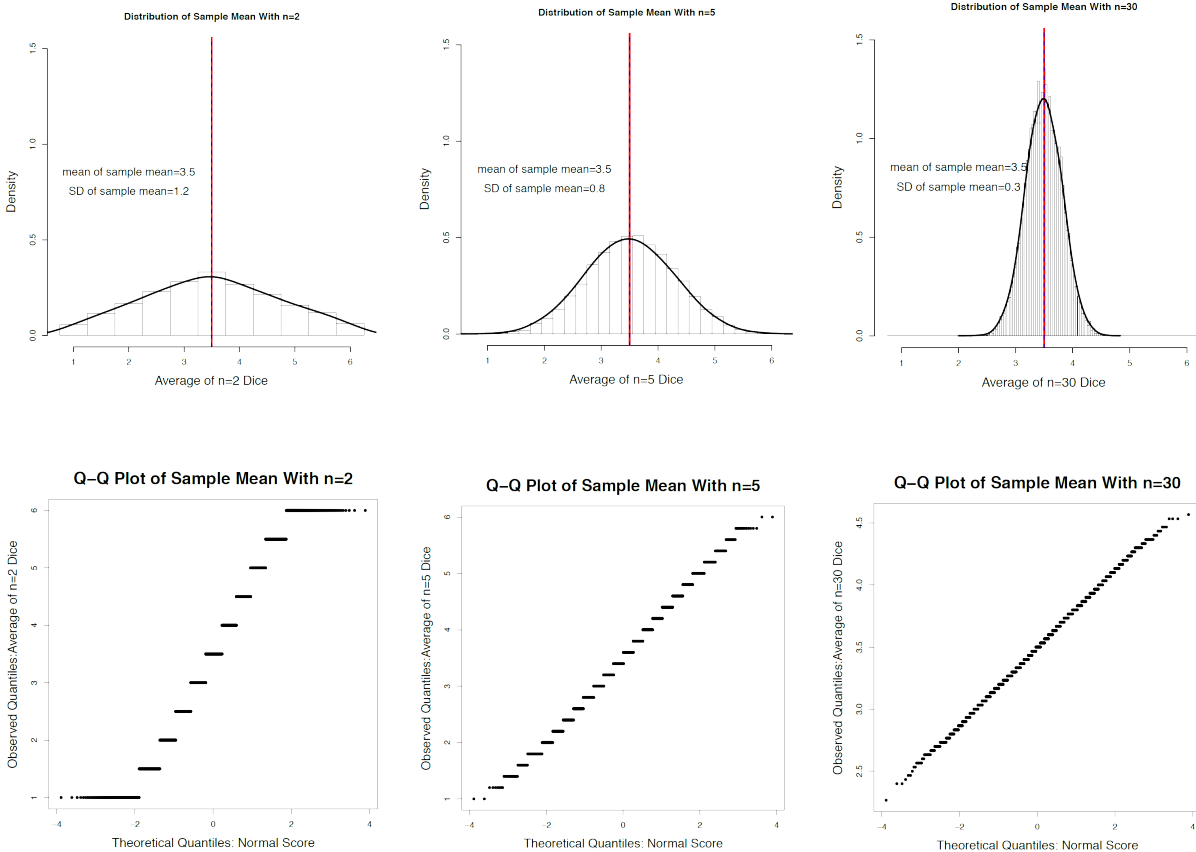


Figure 6.5: Density and Normal Probability Plot of the Average of $n=2, 5, 30$ Rolls (Sample Mean). [\[Image Description \(See Appendix D Figure 6.5\)\]](#) Click on the image to enlarge it.

Here are the findings regarding the distribution of the sample mean \bar{X} :

- The mean of the sample mean is 3.5, which equals the population mean regardless of the sample size n ; the standard deviation roughly equals the population standard deviation divided by the square root of the sample size.
- Notice that for $n = 2$, the distribution of the sample mean appears triangular (not normal), but it becomes increasingly normal for $n = 5$ and $n = 30$.

Example: Population Distribution is Exponential (Extremely Right Skewed)

The exponential distribution is an extremely right-skewed distribution that appears in a variety of real-world applications, including survival times. Suppose X = survival time of liver cancer patients, and that X follows an exponential distribution with a mean and standard deviation of 5 years.

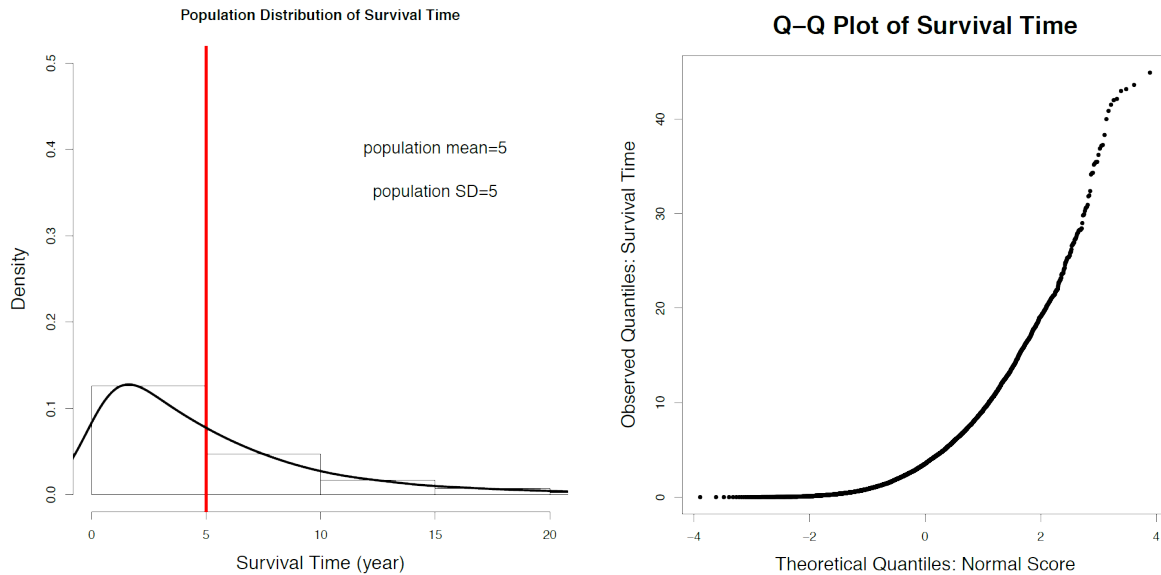
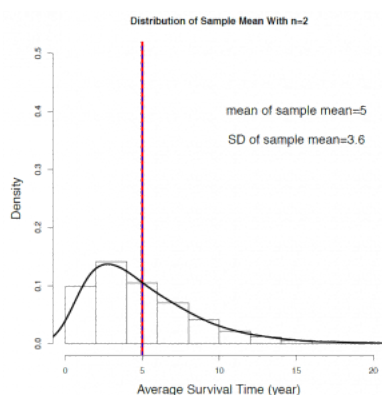


Figure 6.6: Density and Normal Probability Plot of Survival Time (Population). [[Image Description \(See Appendix D Figure 6.6\)](#)]

Let's examine the distribution of the sample mean with sample sizes $n = 2, 5, 30$. That is, the distribution of the average survival time of n randomly selected patients. Once again, note that the mean and standard deviation of the sample mean are: $\mu_{\bar{X}} = \mu = 5$; $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{n}}$

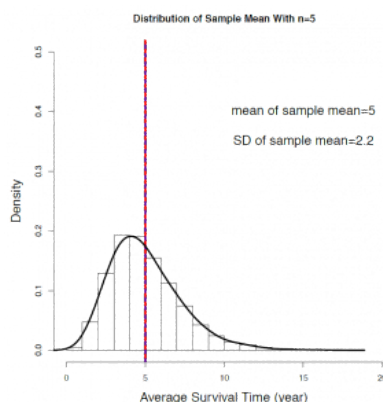
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{2}} = 3.54$$

shape: right skewed



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{5}} = 2.24$$

shape: right skewed



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{30}} = 0.91$$

shape: approximately normal

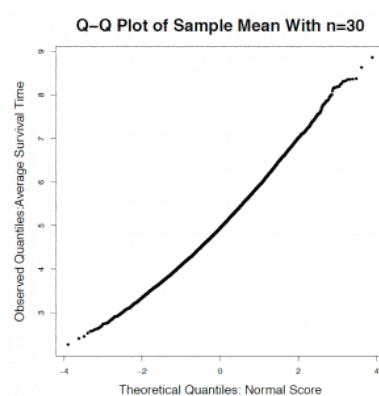
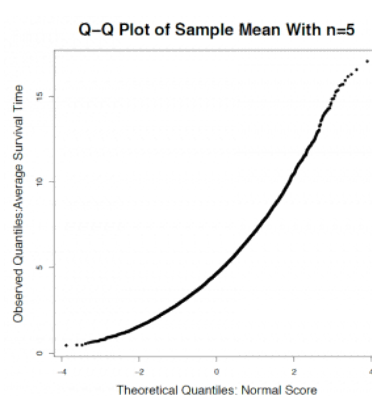
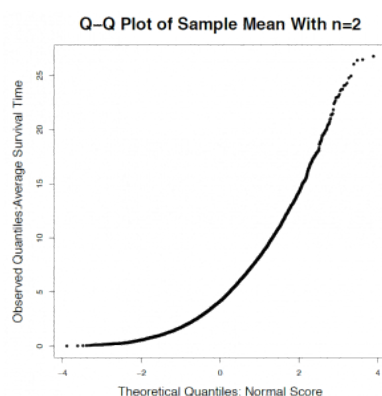
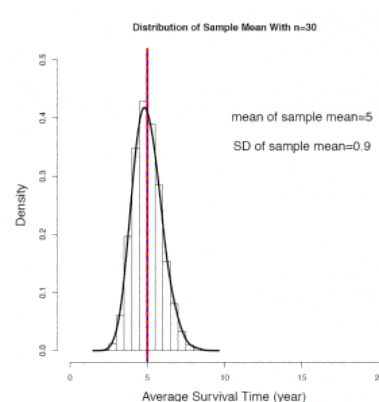


Figure 6.7: Density and Normal Probability Plot of the Average Survival Time of $n=2, 5, 30$ Patients (Sample Mean). [\[Image Description \(See Appendix D Figure 6.7\)\]](#) Click on the image to enlarge it.

Here are the findings:

- The mean of the sample mean is 5, which equals the population mean regardless of the sample size n ; the standard deviation roughly equals the population standard deviation divided by the square root of the sample size.
- The distribution of the sample mean inherits the right skewness of the parent population for relatively small sample sizes $n = 2, 5$, but it is roughly normal when $n = 30$ (note that this trend towards normality increases as n grows beyond 30).

These two examples illustrate that the shape of the distribution of the sample mean \bar{X} is approximately normal when the sample size n is sufficiently large, even if the population distribution is not normal. The more “non-normal” the parent population is, the larger n must be. This is the result of the central limit theorem, which will be discussed in the next section.

6.3 Central Limit Theorem (CLT)

The central limit theorem is one of the most important theorems in statistics.

Key Fact: The Central Limit Theorem

When a random sample of size n is drawn from any population with mean μ and standard deviation σ , the distribution of the sample mean \bar{X} will be (approximately) normally distributed if the sample size n is large enough. In general, $n \geq 30$ is large enough if the population distribution is not too extremely skewed.

Note that:

- The central limit theorem is about the **shape of the distribution of the sample mean** \bar{X} . It is the distribution of the random variable X that will be normally distributed if the sample size n is large enough.
- The required sample size n depends on how skewed the population distribution is. If the population distribution, the distribution of X , is symmetric, $n \geq 5$ might be large enough to claim that the sample mean \bar{X} is approximately normally distributed; if the distribution of X is not too extremely skewed, $n \geq 30$ should be enough; if the population is very skewed, we might need $n \geq 100$ (see the central limit theorem for proportion in Chapter 10).

In addition to the results on the mean and standard deviation of \bar{X} , we can claim that:

Key Fact: The Distribution of the Sample Mean \bar{X}

For a normal population or large sample, the sample mean \bar{X} follows a normal distribution with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. That is $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

1. Let X denote student grades in a particular class, and suppose X is normally distributed with a mean of 70 and a standard deviation of 10, i.e., $X \sim N(\mu = 70, \sigma = 10)$.
 - a. If I randomly pick four students, determine the distribution of their average grade. Indicate the mean, standard deviation, and shape.
Mean: $\mu_{\bar{X}} = \mu = 70$.
Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{4}} = 5$.
Shape: normal since the population is normal. Recall that when the population distribution is normal, the distribution of the sample mean \bar{X} is also normal regardless of the sample size.
Therefore, the average grade of four randomly selected students $\bar{X} \sim N(\mu_{\bar{X}} = 70, \sigma_{\bar{X}} = 5)$.
 - b. If I randomly pick 100 students, determine the distribution of their average grade. Indicate the mean, standard deviation, and shape.
Mean: $\mu_{\bar{X}} = \mu = 70$.
Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$.
Shape: normal since the population is normal. Recall that when the population distribution is normal, the distribution of the sample mean \bar{X} is also normal regardless of the sample size.
Therefore, the average grade of 100 randomly selected students $\bar{X} \sim N(\mu_{\bar{X}} = 70, \sigma_{\bar{X}} = 1)$.
 - c. If I randomly pick four students, find the probability that their average is between 60 and 90.
By part (a), for $n = 4$, average grade $\bar{X} \sim N(\mu_{\bar{X}} = 70, \sigma_{\bar{X}} = 5)$. Therefore,
$$\begin{aligned} P(60 \leq \bar{X} \leq 90) &= P\left(\frac{60 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{90 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(\frac{60 - 70}{5} \leq Z \leq \frac{90 - 70}{5}\right) \\ &= P(-2 \leq Z \leq 4) \\ &= P(Z \leq 4) - P(Z \leq -2) \\ &= 1 - 0.0228 = 0.9772. \end{aligned}$$
2. Suppose the lifetime of a brand of laptops follows an extremely right-skewed distribution with a mean $\mu = 5$ years and a standard deviation $\sigma = 5$.
 - a. If I randomly pick four laptops, determine the distribution of their average lifetime. Indicate the mean, standard deviation, and shape.
Mean: $\mu_{\bar{X}} = \mu = 5$ years.
Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{4}} = 2.5$ years.
Shape: Not normal, still right-skewed. The population is extremely right-skewed, and the sample size $n = 4$ is too small to apply the central limit theorem.
 - b. If I randomly pick 100 laptops, determine the distribution of their average lifetime. Indicate the mean, standard deviation, and shape.
Mean: $\mu_{\bar{X}} = \mu = 5$ years.

Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = 0.5$ years.

Shape: approximately normal. The population is extremely right-skewed, but the sample size $n = 100 > 30$ is large enough to apply the central limit theorem. Therefore,

$\bar{X} \sim N(\mu_{\bar{X}} = 5, \sigma_{\bar{X}} = 0.5)$.

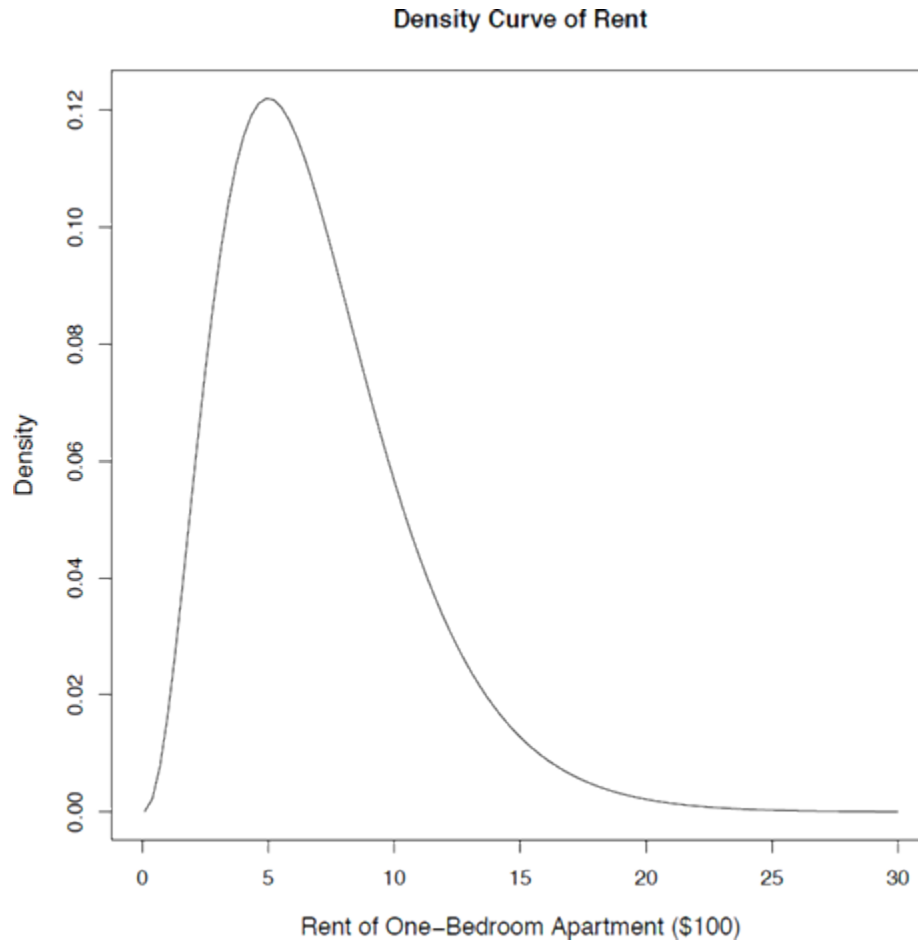
- c. If I randomly pick 100 laptops, find the probability that their average lifetime is at least four years.

By part (b), for $n = 100$, the average lifetime $\bar{X} \sim N(\mu_{\bar{X}} = 5, \sigma_{\bar{X}} = 0.5)$. Hence,

$$\begin{aligned} P(\bar{X} \geq 4) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{4 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(Z \geq \frac{4 - 5}{0.5}\right) = P(Z \geq -2) \\ &= P(Z \leq 2) = 0.9722. \end{aligned}$$

Exercise: Distribution of the Sample Mean

Let X = the rent of a one-bedroom apartment in Edmonton, and suppose that X follows a distribution with a mean of \$700 and a standard deviation of \$400. The distribution of X (the population distribution) is given by the density curve below.



Exercise 6.1 [[Image Description \(See Appendix D Exercise 6.1\)](#)]

- Describe the population distribution of the rent of a one-bedroom apartment in Edmonton, i.e., the distribution of X . Comment on modality, center, spread, and shape.
- If you randomly pick four one-bedroom apartments, describe the sampling distribution of their average rent. Indicate the mean, standard deviation, and shape.
- If you randomly pick 100 one-bedroom apartments, describe the sampling distribution of their average rent. Indicate the mean, standard deviation, and shape.
- If you randomly pick 100 one-bedroom apartments, find the probability that their average rent is above \$800.

Show/Hide Answer

- Unimodal, right skewed, centered at the mean 700 with a spread of 400 as the standard deviation.
- Mean: $\mu_{\bar{X}} = \mu = 700$.
Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{400}{\sqrt{4}} = 200$.
Shape: Not normal, still right-skewed. The population is right-skewed, and the sample size $n = 4$ is too small to apply the central limit theorem.
- Mean: $\mu_{\bar{X}} = \mu = 700$.

Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{400}{\sqrt{100}} = 40$.

Shape: approximately normal. The population is right-skewed, but the sample size $n = 100 > 30$, so it is large enough to apply the central limit theorem.

- d. By part (c), $n = 100$ for, the average rent $\bar{X} \sim N(\mu_{\bar{X}} = 700, \sigma_{\bar{X}} = 40)$. Hence,

$$\begin{aligned} P(\bar{X} \geq 800) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{800 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(Z \geq \frac{800 - 700}{40}\right) \\ &= P(Z \geq 2.5) \\ &= P(Z \leq -2.5) \\ &= 0.0062. \end{aligned}$$

6.4 Learning Objectives Revisited

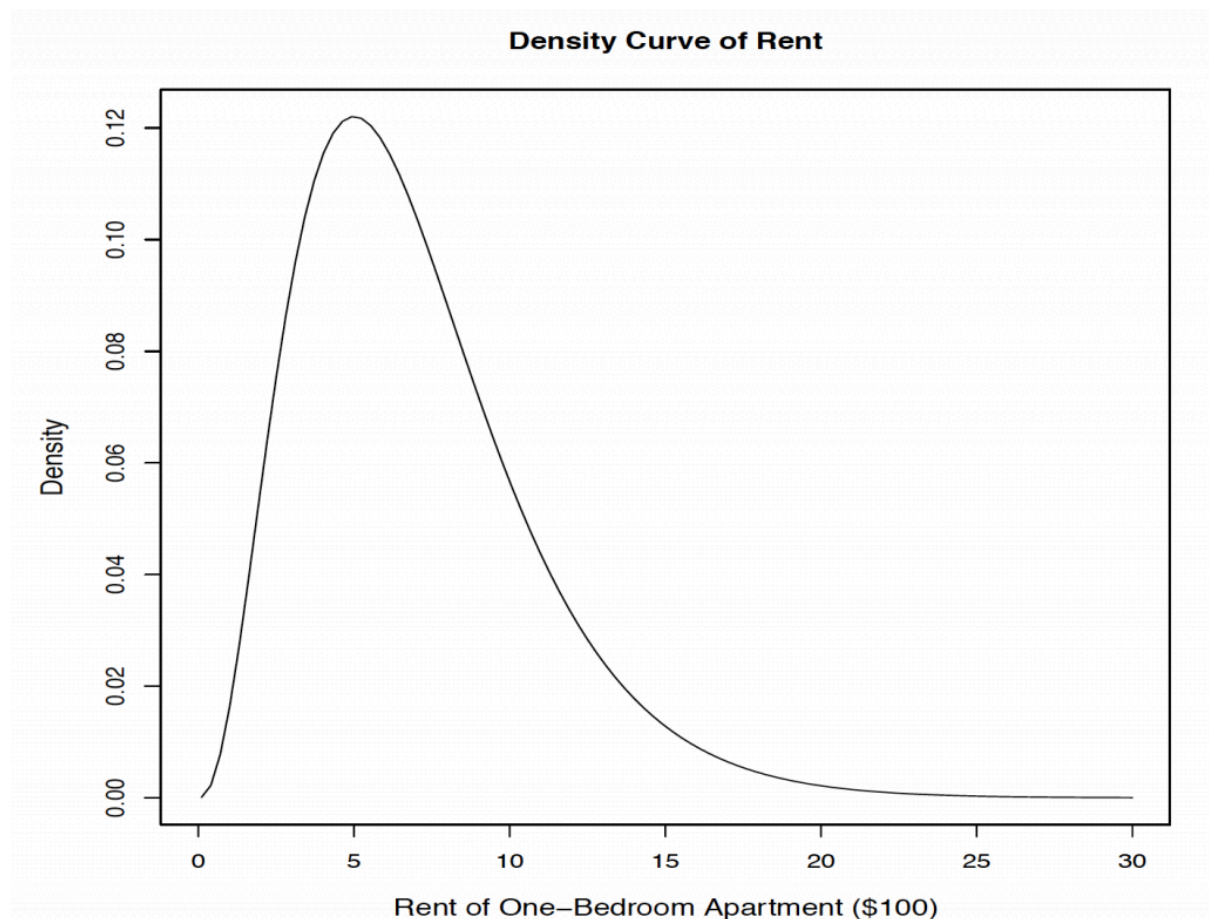
Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the difference between a statistic and a parameter (Section 6.1).
- Explain the relationships among a parameter, an estimator, and a point estimate (Section 6.1).
- Describe how to obtain the sampling distribution of the sample mean (Section 6.2).
- Explain and describe the distribution of sample mean in three aspects: mean, standard deviation, and shape (Section 6.2).
- Explain the central limit theorem in plain English (Section 6.3).
- Apply the central limit theorem to answer questions about the sample mean (Section 6.3).

6.5 Review Questions

1. Explain the relationship between the population mean μ , sample mean \bar{X} , and a value of the sample mean x . Which is the population parameter, which is an estimate and which is an estimator?
2. $P(Z < -4) = \underline{\hspace{2cm}}$; $P(Z < 5) = \underline{\hspace{2cm}}$; $P(Z > 5) = \underline{\hspace{2cm}}$.
3. Which of the following statements about the distribution of the sample mean are true?
 - a. The distribution of the sample mean is never exactly normal.
 - b. The sample mean will be approximately normally distributed when the population is large enough.
 - c. When the sample size is at least 30, the sample mean is always approximately normally distributed. (just a rule of thumb, when the population is NOT extremely skewed))
 - d. When the sample size is large enough, the sample mean will be approximately normally distributed.
 - e. When the population distribution is normal, the sample mean will be exactly normally distributed regardless of the sample size.
 - f. When the sample size is small, the sample mean will never be normally distributed.
 - g. When the sample size is large enough, the population distribution will be approximately normally distributed.
4. Is the following statement about the central limit theorem correct? If it is not correct, could you modify it to make it correct? "When the sample size is greater than 30, it follows a normal distribution".
5. Suppose random variable X is normally distributed with mean 40 and standard deviation 4. Consider all possible random samples of size $n = 4$ from this population, and find the mean, standard deviation, and shape of the sample mean \bar{X} .
6. The time that a technician requires to perform preventive maintenance on an air-conditioning unit is known to have an extremely right-skewed distribution with a mean of 1.2 hours and a standard deviation of 1.2 hours. Consider all possible random samples of size $n = 4$ from this population, and find the mean, standard deviation, and shape of the sample mean \bar{X} .
7. Suppose that X , the rent of a one-bedroom apartment in Edmonton, follows a distribution with a mean of \$800 and a standard deviation of \$400. Its density curve is given below.



[\[Image Description \(See Appendix D Chapter 6 Question 7\)\]](#)

- If you randomly pick four one-bedroom apartments, describe the sampling distribution of their average rent. Indicate the mean, standard deviation, and shape.
- If you randomly pick 64 one-bedroom apartments, describe the sampling distribution of their average rent. Indicate the mean, standard deviation, and shape.
- If randomly pick 64 one-bedroom apartments, find the probability that their average rent is below \$980.
- Fill in the following table about the mean, the standard deviation, and the shape of X (the variable of interest, the population) and sample mean \bar{X} with different sample size n .

Variable	Mean	Standard Deviation	Shape
X (population)	$\mu =$	$\sigma =$	Not normal, right skewed
\bar{X} (sample mean) with $n = 4$	$\mu_{\bar{X}} =$	$\sigma_{\bar{X}} =$	
\bar{X} with $n = 64$	$\mu_{\bar{X}} =$	$\sigma_{\bar{X}} =$	

Show/Hide Answer

- The population mean μ is used to describe the population, μ is a constant but in general unknown. The sample mean \bar{X} is a random variable, and it is an estimator of μ ; its value depends on the sample; the sample mean \bar{x} is a possible value of \bar{X} based on one sample, \bar{x} is also a point estimate of μ .
- $P(Z < -4) = 0$; $P(Z < 5) = 1$; $P(Z > 5) = 0$.
 - False
If the population distribution is normal, the distribution of the sample mean \bar{X} is always exactly normal regardless of the sample size n .
 - False
When the **sample size n** is large enough, the sample mean \bar{X} will be approximately normally distributed.
 - False
Sample size $n \geq 30$ is just a rule of thumb. It is true only when the population is NOT extremely skewed.
 - True, CLT
 - True
 - False
If the population distribution is normal, the distribution of the sample mean \bar{X} is always exactly normal regardless of the sample size n .
 - False
The CLT is about the shape of the distribution of the sample mean \bar{X} ; it has nothing to do with the population distribution.
- The problem of this statement is what “it” refers to. We can modify it to “When the sample size is large enough, the sample mean \bar{X} is approximately normal.”
- The mean of the sample mean equals the population mean: $\mu_{\bar{X}} = \mu = 40$.
The standard deviation of the sample mean equals the population standard deviation divided by the square root of the sample size n : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{4}} = 2$.

The shape of the sample mean \bar{X} is normal, exactly normal, since the population is normal.

5. The mean of the sample mean equals the population mean: $\mu_{\bar{X}} = \mu = 1.2$ hours.

The standard deviation of the sample mean equals the population standard deviation divided by the square root of the sample size n : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{4}} = 0.6$.

The shape of the sample mean \bar{X} is not normal, still right skewed, because the population is right skewed and the sample size $n = 4$ is too small to apply the central limit theorem.

6. This question is almost identical to the Exercise: Distribution of the Sample Mean in Section 6.3.

- a. Mean: $\mu_{\bar{X}} = \mu = 800$.

Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{400}{\sqrt{4}} = 200$.

Shape: Not normal, still right-skewed. The population is right-skewed, and the sample size $n = 4$ is too small to apply the central limit theorem.

- b. Mean: $\mu_{\bar{X}} = \mu = 800$.

Standard deviation: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{400}{\sqrt{64}} = 50$.

Shape: approximately normal. The population is right-skewed, but the sample size $n = 64 > 30$, so it is large enough to apply the central limit theorem.

- c. By part (b), $n = 64$ for, the average rent $\bar{X} \sim N(\mu_{\bar{X}} = 800, \sigma_{\bar{X}} = 50)$. Hence,

$$\begin{aligned} P(\bar{X} < 980) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{980 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(Z \geq \frac{980 - 800}{50}\right) \\ &= P(Z < 3.6) \\ &= 0.9998. \end{aligned}$$

- d. The filled-in table is as follows.

Variable	Mean	Standard Deviation	Shape
X (population)	$\mu = 800$	$\sigma = 400$	Not normal, right skewed
\bar{X} (sample mean) with $n = 4$	$\mu_{\bar{X}} = 800$	$\sigma_{\bar{X}} = \frac{400}{\sqrt{4}} = 200$	Not normal right skewed
\bar{X} with $n = 100$	$\mu_{\bar{X}} = 800$	$\sigma_{\bar{X}} = \frac{400}{\sqrt{64}} = 50$	Approximately normal

6.6 Assignment 6

Purposes

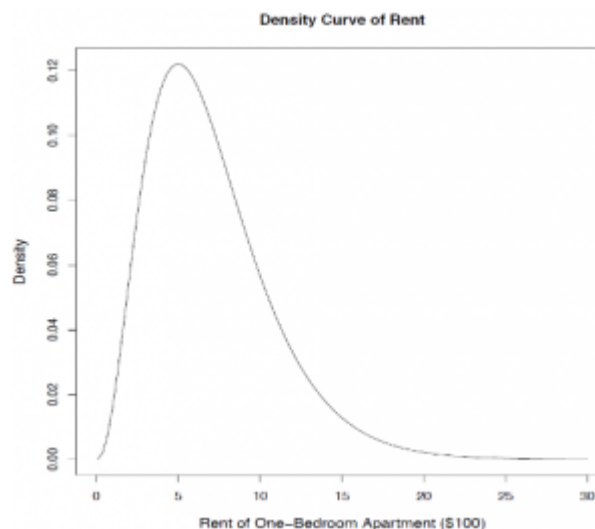
This assignment has two parts to assess your knowledge of the distribution of the sample mean (including the mean, standard deviation, and shape) and the central limit theorem.

Instructions

Part A

Complete the following:

1. Explain the relationship between the population mean μ , sample mean \bar{X} , and a value of the sample mean \bar{x} . Which is the population parameter, which is an estimate, and which is an estimator? (5 marks: 3+2)
2. Explain in plain English why the standard deviation of the sample mean $\sigma_{\bar{X}}$ becomes smaller when the sample size increases. (2 marks)
3. Does the sample size n affect the mean of all possible sample means? Explain your answer. (3 marks)
4. Does the sample size n affect the standard deviation of all possible sample means? Explain your answer. (3 marks)
5. What does the central limit theorem tell us? (3 marks)
6. Do we need the population to be normal or the sample size n to be large to claim that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$? Explain your answer. (3 marks)
7. Suppose that X , the rent of a one-bedroom apartment in Edmonton, follows a distribution with a mean of \$700 and a standard deviation of \$400. Its density curve is given below.



[\[Image Description \(See Appendix D Assignment 6 Question 7\)\]](#) Click on the image to enlarge it.

- Describe the population distribution of the rent of a one-bedroom apartment in Edmonton, i.e., the distribution of X . Comment on modality, centre, spread, and shape. (3 marks)
- If you randomly pick four one-bedroom apartments, describe the sampling distribution of their average rent. Indicate the mean, standard deviation, and shape. (3 marks)
- If you randomly pick 100 one-bedroom apartments, describe the sampling distribution of their average rent. Indicate the mean, standard deviation, and shape. (3 marks)
- If you randomly pick 100 one-bedroom apartments, find the probability that their average rent is above \$800. (4 marks)
- Fill in the following table about the mean, the standard deviation, and the shape of X (the variable of interest, the population) and the sample mean \bar{X} with different sample size n . (4 marks)

Variable	Mean	Standard Deviation	Shape
X (population)	$\mu =$	$\sigma =$	Not normal, right skewed
\bar{X} with $n = 4$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	
\bar{X} with $n = 100$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	

Part B

Finish the following questions using R and R commander. Make sure that you copy and

paste the computer outputs into the space below each question, and write down your answers in statements.

1. Use R commander to explore the distribution of the sample mean \bar{X} when the variable of interest X follows a normal distribution.
 - a. Generate 1,000 observations from a normal distribution with a mean of 70 and a standard deviation of 10. Set the random number seed to be 6,194. Draw a histogram, box plot, and a normal probability plot on these 1,000 observations. Comment on the centre, spread (variation), and shape of the distribution. We call this the population distribution, i.e., the distribution of $X \sim N(70, 10)$. (6 marks)
 - b. Generate $n = 2$ observations from a normal distribution with a mean of 70 and standard deviation of 10, and calculate the mean of these two observations. Repeat this procedure 500 times. Draw a histogram, box plot, and normal probability plot on these 500 sample means, comment on the centre, spread (variation), and shape of the distribution. We call this the sampling distribution of the sample mean \bar{X} with sample size $n = 2$. (6 marks)
 - c. Repeat part (b) with $n = 10$ to obtain the sampling distribution of the sample mean \bar{X} with sample size $n = 10$. (6 marks)
 - d. Repeat part (b) with $n = 36$ to obtain the sampling distribution of the sample mean \bar{X} with sample size $n = 36$. (6 marks)
 - e. Complete the following table about the mean, the standard deviation, and the shape of X (the variable of interest, the population) and the sample mean \bar{X} with different sample size n . (6 marks)

Variable	Mean	Standard Deviation	Shape
X (population)	$\mu =$	$\sigma =$	Normal
\bar{X} with $n = 2$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	
\bar{X} with $n = 10$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	
\bar{X} with $n = 36$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	

2. Use R commander to explore the distribution of the sample mean \bar{X} when the variable of interest X follows an extremely skewed distribution.
 - a. Generate 1,000 observations from an exponential distribution with mean 5 (or rate = $1/5 = 0.2$). Set the random number seed to be 4,067. Draw a histogram, box plot, and a normal probability plot on these 1,000 observations. Comment on the centre, spread (variation), and shape of the distribution. We call this the population

distribution, i.e., the distribution of $X \sim \text{Exponential}(5)$. Note that the standard deviation of an exponential distribution is the same as the mean, i.e., $\sigma = \mu = 5$ in this question. (6 marks)

- b. Generate $n = 2$ observations from an exponential distribution with mean 5, and calculate the mean of these two observations. Repeat this procedure 500 times. Draw a histogram, box plot, and a normal probability plot on these 500 sample means, comment on the centre, spread (variation), and shape of the distribution. We call this the sampling distribution of the sample mean \bar{X} with sample size $n = 2$. (6 marks)
- c. Repeat part (b) with $n = 10$ to obtain the sampling distribution of the sample mean \bar{X} with sample size $n = 10$. (6 marks)
- d. Repeat part (b) with $n = 36$ to obtain the sampling distribution of the sample mean \bar{X} with sample size $n = 36$. (6 marks)
- e. Complete the following table about the mean, the standard deviation, and the shape of X (the variable of interest, the population) and the sample mean \bar{X} with different sample size n . (6 marks)

Variable	Mean	Standard Deviation	Shape
X (population)	$\mu =$	$\sigma =$	Not normal, right skewed
\bar{X} with $n = 2$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	
\bar{X} with $n = 10$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	
\bar{X} with $n = 36$	$\mu_{\bar{X}} = \mu =$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} =$	

Quiz 6



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2628#h5p-8>

CHAPTER 7: CONFIDENCE INTERVAL FOR ONE POPULATION MEAN

We will focus on inferential statistics hereafter. Inferential statistics includes estimation and hypothesis testing: estimation is to estimate the value of a population parameter; hypothesis testing is to test the plausibility of a statement about the value of a population parameter. Estimation consists of point estimates and confidence interval estimates. This chapter introduces how to obtain a point estimate and a confidence interval for the population mean μ .

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the difference between an estimator, a point estimate, and a parameter.
- Explain the practical importance of confidence intervals.
- Distinguish between the standard normal distribution and the t distribution.
- Construct a $(1 - \alpha) \times 100\%$ z and t confidence interval.
- Interpret a confidence interval.
- Explain the relationship between confidence level, precision, length of the interval, and margin of error.
- Calculate the required sample size given the margin of error and confidence level.

7.1 Confidence Interval When σ is Known

Recall that a statistic is a function of the sample data. A statistic is an estimator when it is used to estimate the value of a population parameter. Each possible value of the estimator provides a point estimate of the parameter. For example, we use the sample mean \bar{X} to estimate the population mean μ ; therefore, \bar{X} is an estimator of μ . Given a sample of size n , the value of the sample mean \bar{X} , denoted as \bar{x} , is a point estimate of the population parameter μ . Similarly, the sample standard deviation s provides a point estimate of the population standard σ . Recall that \bar{x} and s are numbers derived from the observed sample data and that different samples tend to yield different values of \bar{x} and s .

Although a point estimate may give us an idea of the true value of the population parameter, the point estimate alone is insufficient. The value of a point estimate is usually not equal to the parameter of interest and error exists when estimating a parameter with a point estimate. In order to improve our estimation, we also need information about the precision of a point estimate, which is provided in the form of an interval (estimate – error, estimate + error). We call this kind of interval a confidence interval in the sense that by adjusting the error, we are able to claim that the parameter is within the interval with a certain confidence level.

7.1.1 One-Sample Z Interval When σ is Known

Recall the distribution of the sample mean \bar{X} : for a normal population or a large sample size, the sample mean $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. According to the 68.26-95.44-99.74 empirical rule for a normal distribution, 95.44% of the \bar{x} values are within two standard deviations ($2\frac{\sigma}{\sqrt{n}}$) away from the population mean μ . That is, 95.44% of the \bar{x} values are within the interval $(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}})$, if we consider all samples of size n . Therefore, for each of those \bar{x} values, the distance between \bar{x} and μ is at most $2\frac{\sigma}{\sqrt{n}}$. That is 95.44% of the \bar{x} values satisfy

$$-2\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < 2\frac{\sigma}{\sqrt{n}}.$$

In other words, 95.44% of the intervals in the form of $(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}})$, contain the population mean μ . Similarly, 95% of the intervals in the form of $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$,

contain the population mean μ . We call the interval $(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$ a 95% confidence interval for μ , which means we are 95% confident that the interval contains the population mean μ . In general, $(1 - \alpha) \times 100\%$ intervals in the form of $(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$ contain the population mean μ and the interval $(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$ is called a $(1 - \alpha) \times 100\%$ confidence interval. The number $(1 - \alpha) \times 100\%$ is called the **confidence level**. Recall that $z_{\alpha/2}$ is the z-score with area of $\frac{\alpha}{2}$ to its right. For example, suppose we wish to construct a 95% confidence interval for μ , then $1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \frac{\alpha}{2} = 0.025 \implies z_{\alpha/2} = z_{0.025} = 1.96$. Hence, a 95% confidence interval for μ is $(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$.

Obtain a $(1 - \alpha) \times 100\%$ Z-Interval When σ is Known

Assumptions:

- A simple random sample (SRS)
- Normal population or large sample size ($n \geq 30$)
- The population standard deviation σ is known

Formula: $(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$ or $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Interpretation: We are $(1 - \alpha) \times 100\%$ confident that the interval contains μ .

Example: One-Sample Z-Interval

A machine fills beer into bottles whose volume is supposed to be 341 ml, but the exact amount varies from bottle to bottle. We randomly picked 100 bottles and obtained the sample mean volume of 339 ml. Assume the population standard deviation $\sigma = 5$ ml. Obtain a 95% confidence interval for the population mean volume μ .

Check the assumptions:

- We have a simple random sample (SRS).
- We do not know whether the population is normal or not, but we have a large sample size with $n = 100 \geq 30$.
- $\sigma = 5$ ml is known.

Steps:

- Find $z_{\alpha/2}$: $1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \frac{\alpha}{2} = 0.025 \implies z_{\alpha/2} = z_{0.025} = 1.96$.
- Information: $n = 100, \bar{x} = 339, \sigma = 5$.

- Interval: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 339 \pm 1.96 \times \frac{5}{\sqrt{100}} = (339 - 0.98, 339 + 0.98) = (338.02, 339.98)$.

Interpretation: we are 95% confident that the interval (338.02, 339.98) contains the population mean volume μ . In other words, we are 95% confident that the mean volume among all bottles filled with this machine is somewhere between 338.02 ml and 339.98 ml.

7.1.2 Interpretation of a Confidence Interval

The interpretation of a confidence interval should be based on repeated samples. That is, suppose we repeatedly draw samples from the population of interest, and, for each sample, we calculate the confidence interval. If we continue this process indefinitely, then the confidence level is the relative frequency of intervals that contain the true value of the parameter of interest. For example, recall that the 95% Z-interval for μ is of the form $(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$; this means that 95% of such intervals contain the true value of μ , while 5% of such intervals fail to capture μ . This is because $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, from which we are able to conclude that there is a 0.95 probability that the random variable \bar{X} will be at most $1.96 \frac{\sigma}{\sqrt{n}}$ away from μ . However, once we obtain our random sample and compute the value of \bar{x} , it is no longer correct to say that there is a 0.95 probability the interval captures μ . Instead, we are 95% confident the interval captures μ (since 95% of all such intervals capture μ , and we hope that the one we obtained is one of those 95% that contain μ).

Let's consider an analogous example. If 95% of students will pass the final exam, then if we randomly choose a student, we should have 95% confidence that this student will pass the exam with the hope that he/she will be one of those 95% who pass the exam. Our confidence comes from the fact that 95% of students will pass the exam.

Example: Interpretation of Confidence Interval

Recall the beer example, wherein the population mean volume is $\mu = 341$ ml with a population standard deviation $\sigma = 5$ ml.

1. Suppose we obtain a random sample of 100 bottles, from which we obtain a sample mean of $\bar{x} = 341.55$. The 95% confidence interval is:

$$(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) = (341.55 - 1.96 \frac{5}{\sqrt{100}}, 341.55 + 1.96 \frac{5}{\sqrt{100}}) = (340.57, 342.53).$$

The population mean, $\mu = 341$, is within the interval since the sample mean (red diamond) is within 1.96 standard deviations of μ (blue lines).

2. We repeat the previous step 20 times and observe that only one sample mean is outside the blue lines and hence, only one interval does not contain μ ; the other nineteen intervals all contain μ , which is 95% out of 20.
3. For each interval, we hope it is one of those 95% contain μ . Therefore, we have 95% confidence that each of those intervals contains the population mean $\mu = 341$.

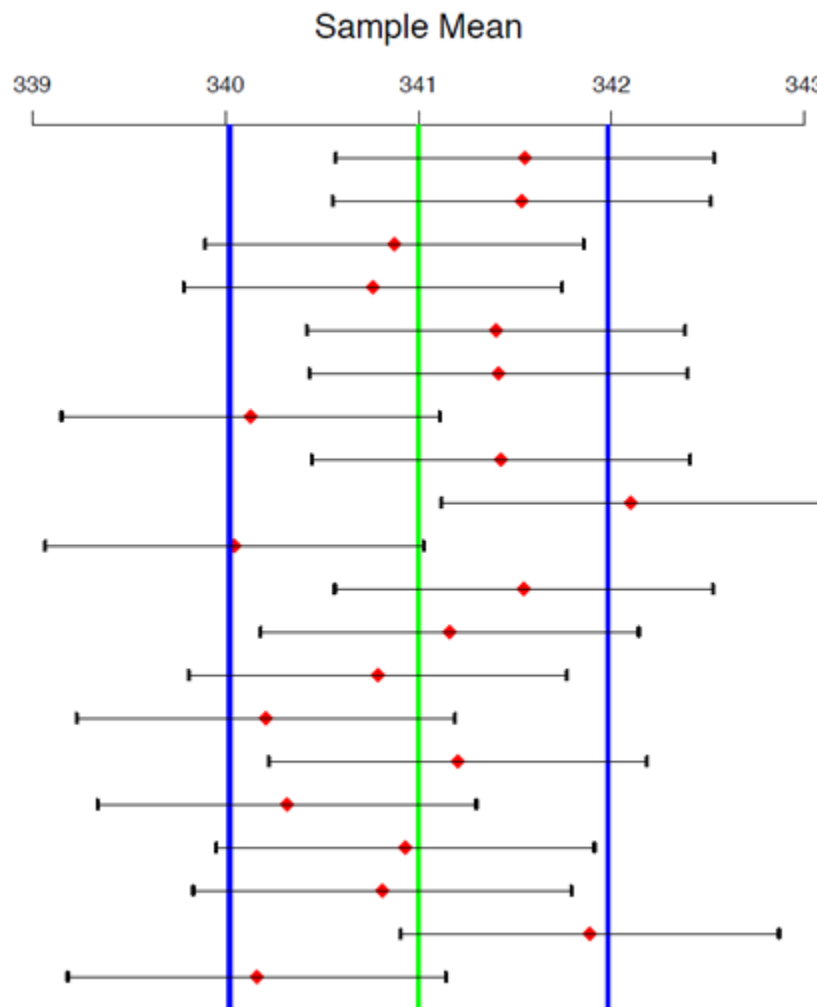


Figure 7.1: Interpretation of Confidence Interval. [Image Description (See Appendix D Figure 7.1)]

Note: In the above example, there is no guarantee that exactly 95% of the samples will capture μ , but rather, this is the expected percentage. As we obtain more samples, the percentage of intervals containing μ will converge to 95%.

Key Fact: Interpretation of a 95% Confidence Interval

A 95% confidence interval for the population mean given by a sample of size n means:

- 95% of samples of size n will produce confidence intervals that contain μ .
- We are 95% confident that the interval will contain μ .
- It is the intervals that vary from sample to sample.
- The population mean μ is fixed. It is usually an unknown constant.

Example: Interpretation of One-sample Z Interval

A machine fills beer into bottles whose volume is supposed to be 341 ml, but the exact amount varies from bottle to bottle. We randomly pick 100 bottles and obtain the sample mean volume is 339 ml. Assume the population standard deviation $\sigma = 5$ ml. A 95% confidence interval for the population mean volume μ is

$$\begin{aligned} & (\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \\ &= 339 \pm 1.96 \times \frac{5}{\sqrt{100}} \\ &= (339 - 0.98, 339 + 0.98) \\ &= (338.02, 339.98). \end{aligned}$$

Interpret this interval. Does it provide evidence to suggest the machine is not working properly?

Interpretation: We are 95% confident that the interval (338.02, 339.98) contains the population mean volume μ . In other words, we are 95% confident that the population mean volume μ is somewhere between 338.02 ml and 339.98 ml.

Since the interval does not contain 341, we are 95% confident that the true mean μ is not equal to 341 ml. As a result, we have evidence to suggest that the machine is NOT working properly. Moreover, since the entire interval is below 341 ml, the data provide evidence that the true mean volume μ is less than 341 ml.

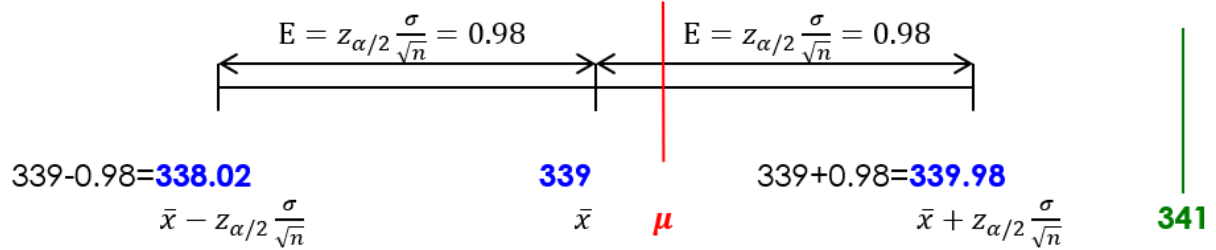


Figure 7.2: Interpretation of Confidence Interval for Population Mean Volume. [[Image Description \(See Appendix D Figure 7.2\)](#)]

Note: The population mean μ is a constant; however, we don't know its value. It can be anywhere within the interval or outside the interval. We are 95% confident that μ is within the resulting confidence interval.

7.1.3 Margin of Error and Sample Size Calculation

The length of a confidence interval $(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$ is twice of the quantity $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. The term $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is called the **margin of error**, denoted as E and it quantifies the accuracy/precision of an estimate. The margin of error is the largest distance between the estimate \bar{x} and the parameter μ given a certain confidence level. Figure 7.2 implies that a confidence interval's length equals twice the margin of error. The larger the margin of error, the wider the confidence interval is, and hence, the more confident we are that the interval contains the parameter of interest, but the estimate is less precise. This is the trade-off between confidence level and precision. If we would like to have a higher level of precision, we have to reduce the confidence level and, therefore, reduce the length of the interval. The formula of the margin of error (i.e., $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$) implies that three things affect the margin of error: the confidence level $1 - \alpha$, the population standard deviation, and the sample size n . If we want to improve precision (i.e., reduce the margin of error E) and maintain the same level of confidence (i.e., fix $1 - \alpha$ or α and hence the z-score $z_{\alpha/2}$), we have to increase the sample size n .

It is often useful to know the minimum sample size n needed in order to ensure the margin is at most E when estimating the population mean μ with the sample mean \bar{x} . More specifically, suppose we have a fixed confidence level $1 - \alpha$, and we want the margin of error to be at most E . Then, recalling that margin of error is defined as $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, we solve for n in order to obtain:

$$n = \left(\frac{\sigma \times z_{\alpha/2}}{E} \right)^2.$$

Since the sample size is an integer, we should round the final result **up to the next integer** (rounding down will give us a margin of error that is slightly larger than E , while rounding up will give us a margin of error that is slightly smaller than E).

Example: Sample Size Calculation

A machine fills beer into bottles whose volume is supposed to be 341 ml.

- a. Determine the number of bottles n we should pick such that we will have 95% confidence that the error is at most 2 ml when the sample mean \bar{x} is used to estimate the population mean volume μ . Assume the population standard deviation of the volume of the bottles is $\sigma = 10$ ml.

Steps:

1. Find the z-score $z_{\alpha/2}$:

$$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \frac{\alpha}{2} = 0.025 \implies z_{\alpha/2} = z_{0.025} = 1.96.$$

2. Apply the formula with: $\sigma = 10$, $E = 2$

$$n = \left(\frac{\sigma \times z_{\alpha/2}}{E} \right)^2 = \left(\frac{10 \times 1.96}{2} \right)^2 = 96.04.$$

Round up to 97. Therefore, the required sample size is $n = 97$.

- b. Determine the sample size n such that the length of a 95% confidence interval is at most 1 ml. Assume $\sigma = 10$ ml.

Steps:

1. Find the z-score $z_{\alpha/2}$:

$$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \frac{\alpha}{2} = 0.025 \implies z_{\alpha/2} = z_{0.025} = 1.96.$$

2. Apply the formula with $\sigma = 10$, $E = \frac{\text{length}}{2} = \frac{1}{2} = 0.5$ (recall that the length of a confidence interval = $2E$):

$$n = \left(\frac{\sigma \times z_{\alpha/2}}{E} \right)^2 = \left(\frac{10 \times 1.96}{0.5} \right)^2 = 1536.64.$$

Round up to 1537. Therefore, the required sample size is $n = 1537$.



Activity

Exercise: Confidence Interval and Sample Size Calculation

Suppose the average birth weight of newborn babies was eight pounds in Edmonton in 2000. I want to investigate whether the average birth weight in 2010 had changed. Assume the birth weight of newborn babies in Edmonton in 2010 has a population standard deviation of $\sigma = 2$ pounds. Suppose that a simple random sample of 100 babies in 2010 gives a mean birth weight of $\bar{x} = 8.6$ pounds.

- Obtain a 90% confidence interval for the mean birth weight of newborn babies in Edmonton in 2010.
- Interpret the confidence interval obtained in part a).
- According to the confidence interval obtained in part b), could you claim that the average birth weight in 2010 is different from that in 2000? If yes, how did it change? Did it seem to increase or decrease?
- Determine the number of babies I should sample in order to be 96% confident that the sampling error is at most 0.5 pounds when estimating μ with \bar{x} .
- Recalculate a 90% confidence interval using the sample size obtained in part d) and compare it with the 90% confidence interval in part a).

Show/Hide Answer

- Check the assumptions:

- The SRS assumption is met since we have a simple random sample of 100 babies.
- Normal population or large sample size assumption is satisfied since the sample size $n = 100 > 30$.
- $\sigma = 2$ is known.

All assumptions are met; we can use a one-sample z interval.

$1 - \alpha = 0.9 \implies \alpha = 0.1 \implies \frac{\alpha}{2} = 0.05 \implies z_{\alpha/2} = z_{0.05} = 1.645$. A one-sample 90% z interval is given by $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 8.6 \pm 1.645 \times \frac{2}{\sqrt{100}} = [8.6 - 0.329, 8.6 + 0.329] = [8.271, 8.929]$.

b) **Interpretation:** We are 90% confident that the mean birth weight of newborn babies in Edmonton in 2010 was somewhere between 8.271 and 8.929 pounds.

c) Since we are 90% confident that the mean birth weight of newborn babies in Edmonton in 2010 is somewhere between 8.271 and 8.929 pounds and the interval does not contain 8, we can claim that the average birth weight in 2010 is different from that in 2000. Moreover, since the entire interval is above 8 pounds, we have evidence that the mean birth weight of newborn babies increased in 2010.

d) Find the z-score $z_{\alpha/2}$:

$$1 - \alpha = 0.96 \implies \alpha = 0.04, \frac{\alpha}{2} = 0.02 \implies z_{\alpha/2} = z_{0.02} = 2.05,$$

$$n = \left(\frac{\sigma \times z_{\alpha/2}}{E} \right)^2 = 67.24.$$

Round up to $n = 68$.

e) Using $n = 68$, the 90% confidence interval is

$$\begin{aligned}\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 8.6 \pm 1.645 \times \frac{2}{\sqrt{68}} \\ &= (8.6 - 0.399, 8.6 + 0.399) \\ &= (8.201, 8.999).\end{aligned}$$

The interval is wider than the one obtained in part a), since the sample size $n = 68$ is smaller ($n = 100$ in part a) which results in a larger margin of error and hence a wider interval.

7.2 Confidence Interval When σ is Unknown

In practice, the population standard deviation is usually unknown. It is often estimated by the sample standard deviation

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

7.2.1 t Distribution and t -Score Table

Recall the distribution of the sample mean \bar{X} : if the population from which we sample is normally distributed or if the sample size is large, it follows that $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. For computational simplicity, we often transform \bar{X} into the standardized variable $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, which follows the standard normal distribution. However, when σ is unknown, it is estimated with the sample standard deviation s , and this leads to a different random variable $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$, which follows the t distribution with a parameter called degrees of freedom $df = n - 1$.

In general, degrees of freedom are the number of independent variables that can take arbitrary values; it equals the number of variables minus the number of relationships among the variables. For example, if two random variables, X and Y , are independent, we have $df = 2$. However, if they satisfy the relationship $X+Y=5$, then $df = 2 - 1 = 1$. The random variable $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ is based on n random variables X_1, X_2, \dots, X_n with $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$; therefore, we have n independent variables with one relationship. As a result, the degree of freedom is $df = n - 1$.

The t density curve is very similar to the standard normal density curve. The following figure shows several t density curves with different degrees of freedom and the standard normal density curve.

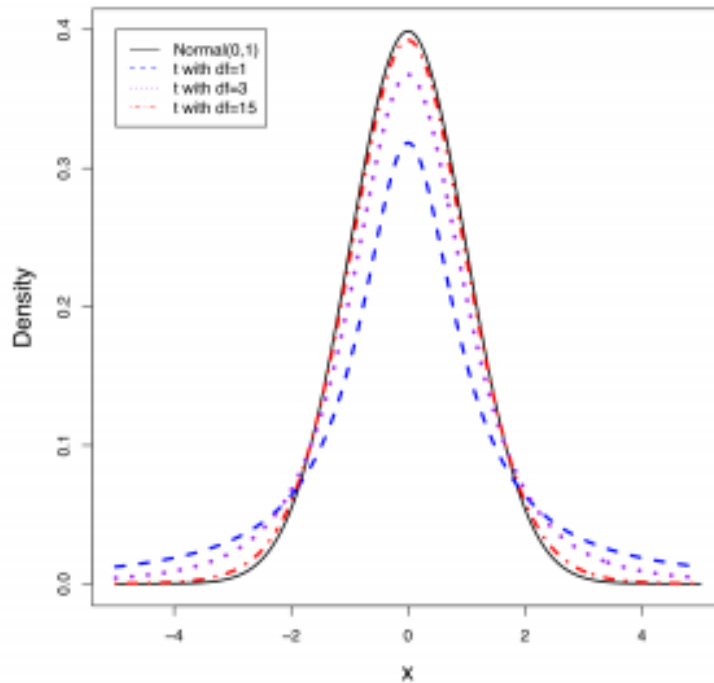


Figure 7.3: Standard Normal Versus t Distributions. [\[Image Description \(See Appendix D Figure 7.3\)\]](#) Click on the image to enlarge it.

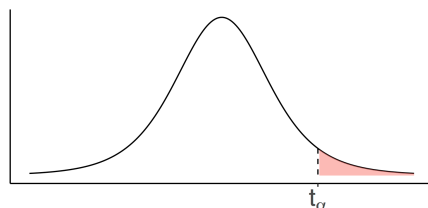
Here are the properties of a t distribution:

Key Facts: Properties of t Density Curve

- The total area under the curve is 1.
- Bell-shaped and symmetric at 0, that is, the area to the right of any given t -score is the same as the area to the left of its negative counterpart: $P(t > t_\alpha) = P(t < -t_\alpha)$. For example, $P(t > 2) = P(t < -2)$.
- When the degrees of freedom $df = n - 1$ increases, the t distribution approaches the standard normal distribution. When $df = \infty$, the t distribution becomes the standard normal.
- The standard normal curve is taller and slimmer, and the t distribution has a fatter and wider tail.

Unlike the standard normal table (Table II) whose main body gives left-tailed areas under the standard normal density curve, the main body of the t -score table (Table IV) gives t -scores, t_α , which are defined in a manner analogous to z_α . That is, the t -scores t_α is the value such that the area to its **right** is α , under the density curve of the t distribution with a given degree of freedom.

Table IV: Values of t_α of t -distribution



df	α : Area to the Right of t_α											
	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.010	0.0075	0.0050	0.0025	0.0005
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	31.821	42.433	63.657	127.321	636.619
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	6.965	8.073	9.925	14.089	31.599
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	3.365	3.634	4.032	4.773	6.869
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	3.143	3.372	3.707	4.317	5.959
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.998	3.203	3.499	4.029	5.408
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.896	3.085	3.355	3.833	5.041
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.821	2.998	3.250	3.690	4.781
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.764	2.932	3.169	3.581	4.587
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.718	2.879	3.106	3.497	4.437
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.681	2.836	3.055	3.428	4.318
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.650	2.801	3.012	3.372	4.221
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.624	2.771	2.977	3.326	4.140
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.602	2.746	2.947	3.286	4.073

Figure 7.4: Usage of t -score Table (Table IV). [Image Description (See Appendix D Figure 7.4)]

For example, if $n = 10$ and $df = n - 1 = 9$, then $t_{0.025} = 2.262$. That is, the t -score 2.262 has an area of 0.025 to its right, under the t -density curve with 9 degrees of freedom. Notice that for each df , the t -table lists only 12 t -scores. For this reason, we are often required to approximate the area to the right of a given t -score. For example, to find the area to the right of the t -score 1.5 under the t density curve with $df = 9$, we first locate the t -score 1.5, which is between 1.383 and 1.833; then, if we look at the top of the table, we see that the area to the right of 1.5 is between 0.1 and 0.05. If we use technology, for example, the R commander, we determine that the t -score of 1.5 has a right-tailed area of 0.0839. That is, when $df = 9$, $t_{0.0839} = 1.5$.



Activity

Exercise: Use of the t -Score Table

Given that $n = 15$, use the t -score table (Table IV) to find

- $t_{0.025}$
- $t_{0.005}$
- $P(t \geq 2.145)$, which is the area to the right of 2.145 under the t density curve.
- $P(t \leq -2.145)$, which is the area to the left of -2.145 under the t density curve.
- $P(t \geq 2.5)$, which is the area to the right of 2.5 under the t density curve.

Show/Hide Answer

For $n = 15$, $df = n - 1 = 14$. Hence, we may refer to the bottom row of the table in Figure 7.4 and Figure 7.5.

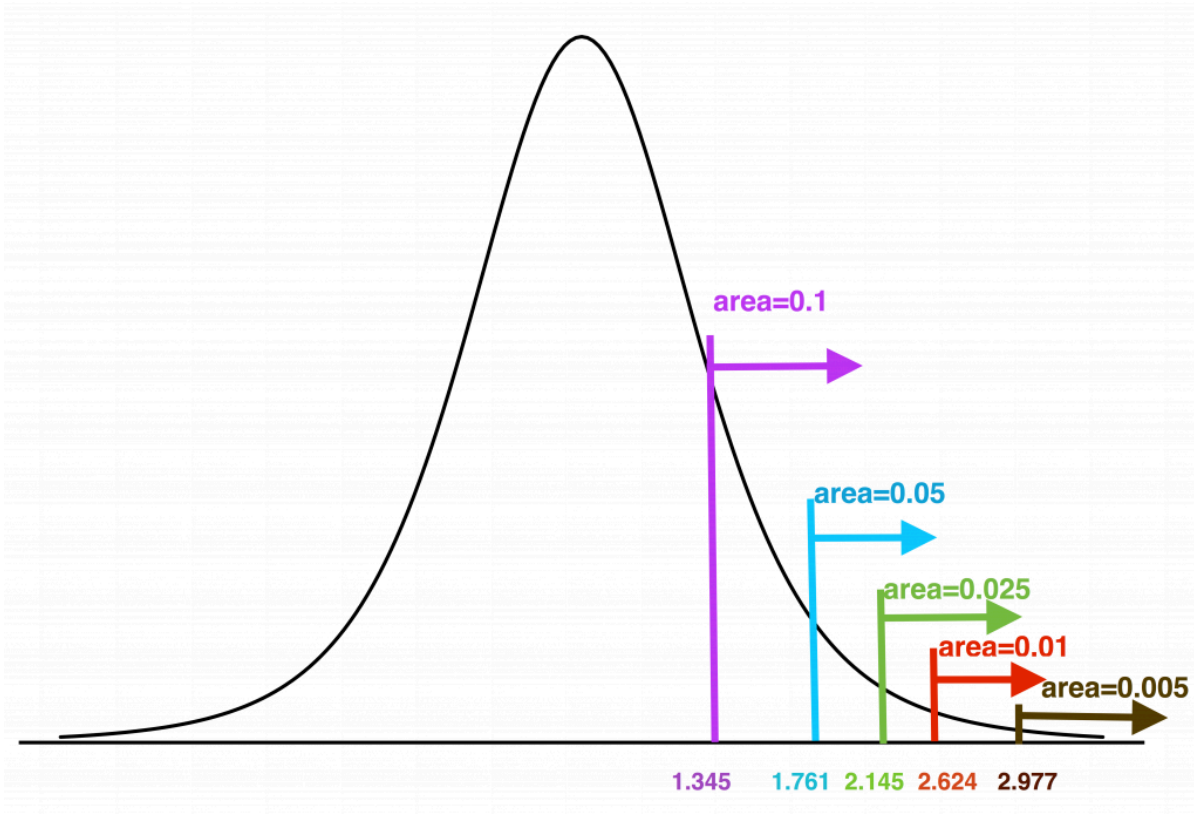


Figure 7.5: Critical Values of t Distribution with $df=14$. [Image Description (See Appendix D Figure 7.5)]

- $t_{0.025} = 2.145$
- $t_{0.005} = 2.977$
- Since $t_{0.025} = 2.145$, it follows that $P(t \geq 2.145) = 0.025$.
- First note that the t distribution is symmetric at 0, so the area to the left of -2.145 is the same as the area to the right of 2.145. Therefore, $P(t \leq -2.145) = P(t \geq 2.145) = 0.025$, which is the area under the t density curve to the left of -2.145.
- Since 2.145 (which is $t_{0.025}$) $< 2.5 < 2.624$ (which is $t_{0.01}$), the area to the right of 2.5 should be somewhere between 0.025 and 0.01. That is, $0.01 < P(t \geq 2.5) < 0.025$.

7.2.2 One-Sample t Interval When σ is Unknown

When the population standard deviation σ is unknown and estimated by the sample standard deviation s , a $(1 - \alpha) \times 100\%$ confidence interval is given by a one-sample t interval:

Assumptions:

1. A simple random sample (SRS)
2. Normal population or large sample size (rule of thumb: $n \geq 30$)
3. The population standard deviation σ is unknown

Formula: $(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}})$ or $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

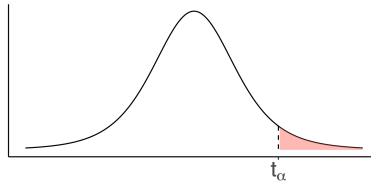
Interpretation: We are $(1 - \alpha) \times 100\%$ confident that the interval contains the population mean μ .

Example: One-Sample t Interval

A computer company claims that the average lifetime of its laptops is about 4 years. A simple random sample of 36 laptops yields an average lifetime of 3.5 years with a sample standard deviation of 4.2 years.

You could use the following truncated Table IV to obtain the t -scores.

Table IV: Values of t_α of t -distribution



df	α : Area to the Right of t_α											
	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.010	0.0075	0.0050	0.0025	0.0005
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	31.821	42.433	63.657	127.321	636.619
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	6.965	8.073	9.925	14.089	31.599
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	3.365	3.634	4.032	4.773	6.869
31	0.256	0.530	0.853	1.054	1.309	1.696	2.040	2.453	2.576	2.744	3.022	3.633
32	0.255	0.530	0.853	1.054	1.309	1.694	2.037	2.449	2.571	2.738	3.015	3.622
33	0.255	0.530	0.853	1.053	1.308	1.692	2.035	2.445	2.566	2.733	3.008	3.611
34	0.255	0.529	0.852	1.052	1.307	1.691	2.032	2.441	2.562	2.728	3.002	3.601
35	0.255	0.529	0.852	1.052	1.306	1.690	2.030	2.438	2.558	2.724	2.996	3.591
36	0.255	0.529	0.852	1.052	1.306	1.688	2.028	2.434	2.555	2.719	2.990	3.582
37	0.255	0.529	0.851	1.051	1.305	1.687	2.026	2.431	2.551	2.715	2.985	3.574
38	0.255	0.529	0.851	1.051	1.304	1.686	2.024	2.429	2.548	2.712	2.980	3.566
39	0.255	0.529	0.851	1.050	1.304	1.685	2.023	2.426	2.545	2.708	2.976	3.558
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.423	2.542	2.704	2.971	3.551
41	0.255	0.529	0.850	1.050	1.303	1.683	2.020	2.421	2.539	2.701	2.967	3.544
42	0.255	0.528	0.850	1.049	1.302	1.682	2.018	2.418	2.537	2.698	2.963	3.538
43	0.255	0.528	0.850	1.049	1.302	1.681	2.017	2.416	2.534	2.695	2.959	3.532
44	0.255	0.528	0.850	1.049	1.301	1.680	2.015	2.414	2.532	2.692	2.956	3.526
45	0.255	0.528	0.850	1.049	1.301	1.679	2.014	2.412	2.529	2.690	2.952	3.520
46	0.255	0.528	0.850	1.048	1.300	1.679	2.013	2.410	2.527	2.687	2.949	3.515
47	0.255	0.528	0.849	1.048	1.300	1.678	2.012	2.408	2.525	2.685	2.946	3.510
48	0.255	0.528	0.849	1.048	1.299	1.677	2.011	2.407	2.523	2.682	2.943	3.505
49	0.255	0.528	0.849	1.048	1.299	1.677	2.010	2.405	2.521	2.680	2.940	3.500
50	0.255	0.528	0.849	1.047	1.299	1.676	2.009	2.403	2.519	2.678	2.937	3.496

[Image Description (See Appendix D Example 7.1)]

- a. Obtain a 99% confidence interval for the population mean lifetime μ .

Check the assumptions:

1. We have a simple random sample (SRS).
2. We do not know whether the population is normal or not, but we have a large sample size $n = 36 > 30$.
3. σ is unknown and estimated by $s = 4.2$.

Steps:

- Find $t_{\alpha/2}$: $n = 36$, $df = n - 1 = 36 - 1 = 35$
 $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \frac{\alpha}{2} = 0.005 \Rightarrow t_{\alpha/2} = t_{0.005} = 2.724$ (using Table IV).
- Information: $n = 36$, $\bar{x} = 3.5$, $s = 4.2$.
- Interval:

$$\begin{aligned}\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} &= 3.5 \pm 2.724 \times \frac{4.2}{\sqrt{36}} = (3.5 - 1.9068, 3.5 + 1.9068) \\ &= (1.5932, 5.4068).\end{aligned}$$

Interpretation: We are 99% confident that the interval (1.5932, 5.4068) contains the population mean lifetime. In other words, we are 99% confident that this computer company produces laptops with a mean lifetime somewhere between 1.5932 and 5.4068 years.

- b. Obtain an 80% confidence interval for the population mean lifetime.

Steps:

- Find $t_{\alpha/2}$: $n = 36$, $df = n - 1 = 36 - 1 = 35$
 $1 - \alpha = 0.8 \implies \alpha = 0.2 \implies \frac{\alpha}{2} = 0.1 \implies t_{\alpha/2} = t_{0.1} = 1.306$ (using Table IV).
- Information: $n = 36$, $\bar{x} = 3.5$, $s = 4.2$.
- Interval:

$$\begin{aligned}\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} &= 3.5 \pm 1.306 \times \frac{4.2}{\sqrt{36}} = (3.5 - 0.9142, 3.5 + 0.9142) \\ &= (2.5858, 4.4142).\end{aligned}$$

Interpretation: We are 80% confident that the interval (2.5858, 4.4142) contains the population mean life μ . In other words, we are 80% confident that this computer company produces laptops with a mean lifetime somewhere between 2.5858 and 4.4142 years.

- c. Does the confidence interval in part a) provide any evidence against the company's claim that the average lifetime of this brand of laptops is about 4 years?
 No. Since the interval (1.5932, 5.4068) contains 4, we can not reject the claim that the average lifetime is about 4 years.



Activity

Exercise: One-Sample t Interval

A nutrition laboratory tests 50 “reduced sodium” hot dogs and finds the sample mean sodium content is 300 mg, with a sample standard deviation of 36 mg.

- a. Obtain a 90% confidence interval for the mean sodium content of this brand of hot dog.
- b. Interpret the confidence interval obtained in part (a).
- c. Suppose that the mean sodium content of all brands of hot dogs on the market is 320 mg. Can

we claim that this brand of “reduced sodium” hot dogs has a lower average sodium content?

Show/Hide Answer

Answers:

a. **Steps:**

- Find $t_{\alpha/2}$: $n = 50, df = n - 1 = 50 - 1 = 49$
 $1 - \alpha = 0.9 \implies \alpha = 0.1 \implies \frac{\alpha}{2} = 0.05 \implies t_{\alpha/2} = t_{0.05} = 1.677$ (using Table IV).
- Information: $n = 50, \bar{x} = 300, s = 36$.
- Interval:

$$\begin{aligned}\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} &= 300 \pm 1.677 \times \frac{36}{\sqrt{50}} = (300 - 8.538, 300 + 8.538) \\ &= (291.462, 308.538).\end{aligned}$$

- b. **Interpretation:** We are 90% confident that this brand of “reduced sodium” hot dogs has a mean sodium content somewhere between 291.462 mg and 308.538 mg.
- c. Since the entire interval $(291.462, 308.538)$ is below 320 mg, we have evidence that this brand of “reduced sodium” hot dog has a lower average sodium content than 320 mg.

7.3 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the difference between an estimator, a point estimate, and a parameter (Section 7.1).
- Explain the practical importance of confidence intervals (Section 7.1).
- Distinguish between the standard normal distribution and the t distribution (Section 7.2).
- Construct a $(1 - \alpha) \times 100\%$ z and t confidence interval (Section 7.2).
- Interpret a confidence interval (Section 7.1).
- Explain the relationship between confidence level, precision, length of the interval, and margin of error (Section 7.1).
- Calculate the required sample size given the margin of error and confidence level (Section 7.1).

7.4 Review Questions

1. The value of a statistic used to estimate a parameter is called a _____ of the parameter.
2. Explain the relationship between the population mean μ , sample mean \bar{X} , and a value of the sample mean \bar{x} . Which is the population parameter, which is a statistic, which is a point estimate, and which is an estimator?
3. Suppose the average birth weight of newborn babies in Edmonton was 8 pounds in 2000. I want to know whether the average birth weight in 2010 had changed or not. Assume the population standard deviation of birth weight of newborn babies in Edmonton is $\sigma = 2$ pounds. Suppose a simple random sample of 100 babies in 2010 gives a mean birth weight of $\bar{x} = 8.6$ pounds.
 - a. Obtain a 90% confidence interval for the mean birth weight of newborn babies in Edmonton in 2010.
 - b. Interpret the confidence interval obtained in part (a).
 - c. According to the confidence interval obtained in part (b), could you claim that the average birth weight in 2010 is different from that in 2000? If yes, how did it change?
 - d. Determine the number of babies I should pick to guarantee that the error of \bar{x} in estimating μ is at most 0.5 pounds with 96% confidence.
 - e. Re-calculate a 90% confidence interval using the sample size obtained in part (d) and compare it with the 90% confidence interval in part (a).
4. Given that $n = 15$, use the t-score Table (Table IV) to find
 - a. $t_{0.025} =$ _____.
 - b. $t_{0.005} =$ _____.
 - c. $P(t \geq 2.145)$ that is the area under the t density curve to the right of 2.145.
 - d. $P(t \leq -2.145)$ that is the area under the t density curve to the left of -2.145.
 - e. $P(t \geq 2.5)$ that is the area under the t density curve to the right of 2.5.
5. A computer company claims that the average lifetime of its laptop is about 4 years. A simple random sample of 36 laptops yields an average lifetime of 3.5 years with a sample standard deviation of 4.2 years.
 - a. Obtain a 99% confidence interval for the population mean lifetime μ .
 - b. Obtain an 80% confidence interval for the population mean lifetime μ .
 - c. Does the confidence interval in part (a) support the claim that the average lifetime of this brand of laptops is about 4 years? Explain why.
6. A nutrition laboratory tests 50 "reduced sodium" hot dogs, finding that the sample

mean sodium content is 300 mg, with a sample standard deviation of 36 mg.

- a. Obtain a 90% confidence interval for the mean sodium content of this brand of hot dog.
 - b. Interpret the confidence interval obtained in part (a).
 - c. Suppose that the mean sodium content of all brands of hot dogs on the market is 320 mg. Can we claim that this brand of "reduced sodium" hot dog has a lower sodium content? Explain why.
7. Suppose a 90% confidence interval for the mean weight of all newborn babies in Canada in 2015 was (5.5, 9.5) lb, which of the following statements about the interpretation of this confidence interval are correct.
- a. The chance that the true population mean falls in the interval (5.5, 9.5) is 0.9. (False, μ is fixed, and the interval is also fixed, with no randomness, no probability).
 - b. The chance that the sample mean falls in the interval (5.5, 9.5) is 0.9. (False)
 - c. We can be 90% confident that the interval (5.5, 9.5) contains the true population mean. (True)
 - d. We can be 90% confident that the true population mean is somewhere between 5.5 and 9.5 lb. (True, standard way of the textbook)
 - e. We can be 90% confident that the sample mean is somewhere between 5.5 and 9.5 lb. (False)
 - f. Consider all samples of the same size; 90% of the sample means fall in the interval (5.5, 9.5). (False, 90% of the sample means with the same sample size are within the interval $\mu \pm t_{0.05} \times \frac{s}{\sqrt{n}}$.)
 - g. Consider all samples of the same size and obtain a 90% confidence interval from each sample; 90% of those intervals contain the true population mean. (True)

Show/Hide Answer

1. The value of a statistic used to estimate a parameter is called a **point estimate** of the parameter.
2. μ is a population parameter that is a constant and normally unknown. \bar{X} is a statistic and is a random variable; its value varies from sample to sample. \bar{x} is the average of one sample; it is one value for \bar{X} and is a point estimate of μ .
3.
 - a. $1 - \alpha = 0.9 \Rightarrow \alpha = 1 - 0.9 = 0.1, z_{\alpha/2} = z_{0.05} = 1.645$
a 90% CI for mean birth weight is
 $\left(8.6 - 1.645 \times \frac{2}{\sqrt{100}}, 8.6 + 1.645 \times \frac{2}{\sqrt{100}}\right) = (8.271, 8.929)$
Note: the z-score $z_{\alpha/2}$ can be obtained at the bottom of the second page of the t -score table (Table IV).

- b. We can be 90% confident that the mean birth weight of newborn babies in Edmonton in 2010 was between 8.271 lb and 8.929 lb.
- c. Yes, the entire interval is above 8lb. We can claim that the average birth weight in 2010 was greater than 8lb. The average birth weight in 2010 has increased compared to that in 2000.
- d. $1 - \alpha = 0.96, \alpha = 0.04, z_{\alpha/2} = z_{0.02} = 2.05, E = 0.5, n = \left(\frac{z_{\alpha/2} \times \sigma}{E} \right)^2 = \left(\frac{2.05 \times 2}{0.5} \right)^2 = 67.24$
Round up to $n = 68$.
- e. $1 - \alpha = 0.9, \alpha = 0.1, z_{\alpha/2} = 2.05$
given $n=68$, a 90% CI is

$$\left(8.6 - 1.645 \times \frac{2}{\sqrt{68}}, 8.6 + 1.645 \times \frac{2}{\sqrt{68}} \right) = (8.103, 9.097)$$
which is wider than the 90% interval obtained in part (a). This 90% CI is not as precise as the one in part (a) since it has a larger margin of error. Note that the sample size in part (e) is $n = 68$, which $n = 100$ in part (a), so the 90% CI in part (a) is more accurate.

4. Given that $n = 15$, use the t-score Table (Table IV) to find. $df = n - 1 = 15 - 1 = 14$

- a. $t_{0.025} = 2.145$.
- b. $t_{0.005} = 2.977$.
- c. $P(t \geq 2.145)$ that is the area under the t density curve to the right of 2.145.
area=0.025 since $t_{0.025} = 2.145$, and hence $P(t \geq 2.145) = 0.025$.
- d. $P(t \leq -2.145)$ that is the area under the t density curve to the left of -2.145.
Since t density curve is symmetric at 0, $P(t \leq -2.145) = P(t \geq 2.145) = 0.025$.
- e. $P(t \geq 2.5)$ that is the area under the t density curve to the right of 2.5.
Since 2.415(area to its right is 0.025) $< 2.5 < 2.624$ (area to its right is 0.01),
 $0.01 < P(t \geq 2.5) < 0.025$.

5.

$$n = 36, df = n - 1 = 35, 1 - \alpha = 0.99, \alpha = 0.01, t_{\alpha/2} = t_{0.005} = 2.724, \bar{x} =$$

- a. 3.5, $s = 4.2$

a 99% CI is

$$\left(3.5 - 2.724 \times \frac{4.2}{\sqrt{36}}, 3.5 + 2.724 \times \frac{4.2}{\sqrt{36}} \right) = (1.593, 5.407)$$

Interpretation: we can be 99% confident that the average lifetime is between 1.593 and 5.407 years.

$$n = 36, df = n - 1 = 35, 1 - \alpha = 0.80, \alpha = 0.2, t_{\alpha/2} = t_{0.1} = 1.306, \bar{x} =$$

- b. 3.5, $s = 4.2$

an 80% CI is

$$\left(3.5 - 1.306 \times \frac{4.2}{\sqrt{36}}, 3.5 + 1.306 \times \frac{4.2}{\sqrt{36}} \right) = (2.5858, 4.4142)$$

Interpretation: we can be 80% confident that the average lifetime is between 2.5858 and 4.4142 years.

- c. Yes, since the interval contains 4. Therefore, we don't have sufficient evidence that the population mean differs from 4.

6.

a. $n = 50, df = n - 1 = 49, 1 - \alpha = 0.9, \alpha = 0.1, t_{\alpha/2} = t_{0.05} = 1.677, \bar{x} = 300, s = 36$

a 90% CI is

$$\left(300 - 1.677 \times \frac{36}{\sqrt{50}}, 300 + 1.677 \times \frac{36}{\sqrt{50}} \right) = (291.462, 308.538)$$

b. Interpretation: We can be 90% confident that the population mean sodium content is somewhere between 291.462 mg to 308.538 mg

c. Yes, since the entire interval is below 320 mg, which means $\mu < 320$. Therefore, we can claim that this brand of “reduced sodium” hot dog has a lower sodium content.

7. Suppose a 90% confidence interval for the mean weight of all newborn babies in Canada in 2015 was (5.5, 9.5) lb, which of the following statements about the interpretation of this confidence interval are correct.

a. False, μ is fixed, and the interval is also fixed, with no randomness; there is no probability concept here.

b. False, similar to (a). μ is fixed, and the interval is also fixed; there is no randomness.

c. True. If we consider all possible 90% intervals, 90% of the intervals will contain the population mean. We have 90% confidence that each of these intervals will contain the population mean.

d. True, standard way of interpretation given in the textbook.

e. False. The sample mean \bar{x} is the center of the interval and hence should be in the interval for sure.

f. False. Consider the distribution of the sample mean \bar{X} . 90% of the sample means with the same sample size n are within the interval $\mu \pm t_{0.05} \times \frac{s}{\sqrt{n}}$ for one-sample t interval.

g. True.

7.5 Assignment 7

Purposes

This assignment has two parts. The first part assesses your knowledge of distinguishing a statistic and a parameter, an estimator and a point estimate, obtaining and interpreting a one-sample Z interval and a one-sample t interval and calculating the required sample size given the margin of error and confidence level. The second part assesses your skills in using R commander to obtain a one-sample confidence interval for the population mean μ .

Resources

[M07_Age_Millionaire_Q9.xlsx](#)

[M07_BloodPressure_Diabete_Q10.xlsx](#)

Instructions

Part A

Complete the following:

1. The value of a statistic used to estimate a parameter is called a _____ of the parameter. (2 marks)
2. Explain the relationship between the population mean μ , the sample mean \bar{X} , and a value of the sample mean \bar{x} . Which is the population parameter, which is a statistic, which is a point estimate, and which is an estimator? (4 marks)
3. What is a confidence interval estimate of a parameter? Why is such an estimate superior to a point estimate? (3 marks)
4. Explain the similarities and differences between a standard normal distribution and a t distribution. (3 marks)
5. Must the variable under consideration be normally distributed for you to use the z -interval procedure or t -interval procedure? Explain your answer. (3 marks)
6. Given that $n = 36$, use the t-score Table (Table IV; you can find a copy online or in the

Blackboard course) to find

- a. $t_{0.025} =$ _____ (1 mark)
 - b. $t_{0.005} =$ _____ (1 mark)
 - c. $P(t \geq 2.030) =$ _____, that is the area under the t density curve to the right of 2.030. (2 marks)
 - d. $P(t \leq -2.030) =$ _____, that is the area under the t density curve to the left of -2.030. (2 marks)
 - e. $P(t \geq 2.6) =$ _____, that is the area under the t density curve to the right of 2.6. (2 marks)
7. If you obtained one thousand 95% confidence intervals for a population mean, μ , roughly how many of the intervals would actually contain μ ? (2 marks)
 8. A confidence interval for a population mean has a margin of error of 10.7.
 - a. Obtain the length of the confidence interval. (2 marks)
 - b. If the mean of the sample is 75.2, determine the confidence interval. (2 marks)
 9. The following table gives the age (in years) of 36 randomly selected U.S. millionaires. The sample mean $\bar{x} = 58.53$ years. Assume that the standard deviation of ages of all U.S. millionaires is 13.0 years. (See data on file: **M07_Age_Millionaire_Q9.xlsx**)

31	45	79	64	48	38	39	68	52
59	68	79	42	49	53	74	66	66
71	61	52	47	39	54	67	55	71
77	64	60	75	42	69	48	57	48

- a. Obtain a 95% confidence interval for μ , the mean age of all U.S. millionaires. (4 marks)
 - b. Interpret the confidence interval obtained in part (a). (2 marks)
 - c. According to the confidence interval obtained in part (b), could you claim that the average age of all U.S. millionaires is above 55 years? Explain your answer. (3 marks)
 - d. Determine the number of millionaires who should be picked to guarantee that the error of \bar{x} in estimating μ is at most 0.5 years with 98% confidence. (4 marks)
10. Past studies showed that maternal diabetes results in obesity, blood pressure, and glucose tolerance complications in the offspring. Following are the arterial blood pressures, in millimetres of mercury (mm Hg), for a random sample of 16 children of diabetic mothers. The sample mean is $\bar{x} = 85.99$ mm Hg and the sample standard deviation is $s = 8.08$ mm Hg. (See data on file: **M07_BloodPressure_Diabete_Q10.xlsx**)

81.6	84.1	87.6	82.8	82.0	88.9	86.7	96.4
84.6	101.9	90.8	94.0	69.4	78.9	75.2	91.0

- Obtain a 95% confidence interval for the mean arterial blood pressure of all children of diabetic mothers. (4 marks)
- Interpret the confidence interval obtained in part (a). (2 marks)
- Obtain a 90% confidence interval for the mean arterial blood pressure of all children of diabetic mothers. (4 marks)
- Compare the confidence intervals obtained in parts (a) and (c). Which interval is wider? Write a sentence to summarize the relationship between the length of a confidence interval and the confidence level. (4 marks)

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs into the space below each question, and write down your answers in statements.

- For Question 9 in Part A (See data on file: **M07_Age_Millionaire_Q9.xlsx**), use the proper graphical tools in R and R commander to assess whether it is reasonable to apply the one-sample z interval procedure. Make sure to write down the assumptions of the procedure and address whether the assumptions are satisfied. (5 marks)
- Refer to the data in Question 10 in Part A; also see data on file: **M07_BloodPressure_Diabete_Q10.xlsx**.
 - Use the proper graphical tools in R and R commander to assess whether applying the one-sample t interval procedure is reasonable. Make sure to write down the assumptions of the procedure and address whether the assumptions are satisfied. (5 marks)
 - Obtain a 95% confidence interval for the mean arterial blood pressure of all children of diabetic mothers. Compare the answer you obtained by hand. (2 marks)
 - Obtain a 90% confidence interval for the mean arterial blood pressure of all children of diabetic mothers. Compare the answer you obtained by hand. (2 marks)

Quiz 7



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2630#h5p-9>

CHAPTER 8: HYPOTHESIS TESTS FOR ONE POPULATION MEAN

Overview

Recall that inferential statistics includes estimation and hypothesis testing. Chapter 7 introduces point estimates and confidence intervals (estimation) for the population mean μ . This chapter introduces hypothesis tests for the population mean μ . Hypothesis tests are used to test statements about the value of the population mean μ . For example:

- A certain brand of energy-saving light bulb advertises that its bulbs last at least 15,000 hours. A contractor suspects this claim is invalid, and he wishes to test if the mean lifespan of bulbs is less than 15,000 hours.
- A factory produces bottles of lotion that are intended to contain 100 ml of lotion. A factory worker performs regular tests to determine if the average volume of bottles differs from 100ml.
- A popular pizzeria wishes to ensure speedy service to customers, so the franchise strives to ensure that its average delivery time is below 30 minutes. After receiving a complaint regarding slow service, the owner decides to conduct a test to see if the average delivery time is above 30 minutes.

Hypothesis tests provide a formal tool that can be used to measure the strength of evidence supporting your suspicions.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Write down the null and alternative hypotheses for a study.
- Explain the difference between the type I and the type II errors and the relationship between these two types of errors.
- Define the P-value, both mathematically and in plain English.
- Distinguish between the one-mean Z and the one-mean t -test, and identify when each test should be used.
- Conduct a one-mean Z test and a one-mean t -test using the P-value or critical value approaches.

- Calculate the P-value or the range of the P-value for a hypothesis test.
- Explain the relationship between confidence intervals and hypothesis tests.

8.1 Hypotheses

The first step of a hypothesis test is the formulation of the hypotheses. A hypothesis is a statement about the value of a parameter. For example, $\mu \neq 100$ ml, $\mu < 100$ ml or $\mu > 100$ ml, etc.

There are two hypotheses in a hypothesis test: the null hypothesis, denoted as H_0 ; and the alternative hypothesis, denoted as H_a . The alternative hypothesis is a claim we wish to establish regarding a population parameter, and the null hypothesis is the opposite claim. More specifically, the alternative hypothesis represents the plausible values of the parameter if we reject the null hypothesis. Therefore, the null and alternative hypotheses must be complimentary statements about the population parameter.

There are two possible outcomes in a hypothesis test: reject the null H_0 in favour of the alternative H_a or do not reject H_0 . Rejecting a hypothesis means that the hypothesis is claimed to be false, and accepting a hypothesis implies that the hypothesis is claimed to be true. If sample data provide strong evidence against the null hypothesis, it is rejected in favor of the alternative hypothesis; if the data do not provide strong evidence against the null hypothesis, we fail to reject it. Failure to reject the null hypothesis does not imply the null hypothesis is true, but rather, it could be true.

For example, the legal systems of many countries act in accordance with the presumption of innocence, under which an individual is presumed innocent until proven guilty. In such countries, when an individual has been charged with a crime, the null and alternative hypotheses are H_0 the defendant is innocent versus H_a the defendant is guilty. The defendant will be found guilty (H_0 rejected in favor of H_a) only if there is strong evidence to indicate their guilt; if the evidence provided is not sufficient, the defendant is found not guilty (H_0 is not rejected). However, in the case where the defendant is found not guilty, it is not because there is strong evidence to establish their innocence but rather insufficient evidence to prove their guilt.

Steps to set up the hypotheses:

1. Look for the keywords and write down what we want to claim under the alternative H_a .
2. Write the opposite of the alternative H_a to obtain the null H_0 .
Depending on the purpose of the hypothesis test, there are three choices for H_a :

Table 8.1: Key Words and Choices of Alternative Hypothesis

Two tailed	Right (upper) tailed	Left (lower) tailed
$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$
"differ", "change"	"more than", "increase"	"less than", "decrease"

Note: The notation μ_0 represents a specific hypothesized value for the population mean μ . It will be replaced by a certain number in context. For example, $\mu_0 = 7$ in the following example.

Example: Set Up Hypotheses

Newborn babies in 2010 had an average weight of 7 pounds. Write the hypotheses to test whether:

- a. The average weight in 2015 has **changed**.

Steps:

1. Key word: "changed" → a two-tailed test. Write down the alternative: $H_a : \mu \neq 7 \text{ lb}$
2. Write the opposite of the alternative to obtain the null $H_0 : \mu = 7 \text{ lb}$

Therefore, the hypotheses are $H_0 : \mu = 7 \text{ lb}$ versus $H_a : \mu \neq 7 \text{ lb}$.

- b. The average weight in 2015 has **increased**.

Steps:

1. Key word: "increased" → a right-tailed test. Write down the alternative: $H_a : \mu > 7 \text{ lb}$.
2. Write the opposite of the alternative to obtain the null $H_0 : \mu \leq 7 \text{ lb}$.

Therefore, the hypotheses are $H_0 : \mu \leq 7 \text{ lb}$ versus $H_a : \mu > 7 \text{ lb}$

- c. The average weight in 2015 has **decreased**.

Steps:

1. Key word: "decreased" → a left-tailed test. Write down the alternative: $H_a : \mu < 7 \text{ lb}$
2. Write the opposite of the alternative to obtain the null $H_0 : \mu \geq 7 \text{ lb}$

Therefore, the hypotheses are $H_0 : \mu \geq 7 \text{ lb}$ versus $H_a : \mu < 7 \text{ lb}$.



Activity

Exercises: Set up Hypotheses for the Following Studies

- a. A machine fills beer into bottles whose volume is supposed to be 341 ml, but the exact amount varies from bottle to bottle. We randomly picked 100 bottles and obtained the sample mean volume of 339 ml. Assume the population standard deviation $\sigma = 5$ ml. Test at the 5% significance level whether the machine is NOT working properly.
- b. A computer company claims that its laptops have an average lifetime of about 4 years. A simple random sample of 36 laptops yields an average lifetime of 3.5 years with a sample standard deviation of 4.2 years. Test at the 1% significance level whether this brand of laptops has a mean lifetime of less than 4 years.
- c. The number of cell phone users has increased dramatically since 1997. Suppose the mean local monthly bill was \$50 for cell phone users in the United States in 2006. A simple random sample of 50 cell phone users was obtained in 2019, and the sample mean local monthly bill was $\bar{x} = 55$ with a sample standard deviation $s = \$25$. At the 10% significance level, do the data provide sufficient evidence to conclude that the 2019 mean local monthly bill for cell phone users has increased from the 2006 mean of \$50?

Show/Hide Answer

Answers:

1. $H_0 : \mu = 341$ ml versus $H_a : \mu \neq 341$ ml
2. $H_0 : \mu \geq 4$ years versus $H_a : \mu < 4$ years
3. $H_0 : \mu \leq 50$ versus $H_a : \mu > 50$

8.2 Type I and Type II Errors

In testing hypotheses, there are only two possible outcomes: either reject H_0 or do not reject H_0 ; in reality, there are only two possible scenarios: either H_0 is true or H_0 is false. Hence, regardless of which conclusion we make, we have a chance to make an error. There are two types of errors: Type I and Type II.

Type I error: reject the null H_0 when H_0 is in fact true.

Type II error: do not reject the null H_0 when H_0 is false.

Table 8.2: Type I and Type II Error

	H_0 is True	H_0 is False
Decision: Do not reject H_0	Correct decision	Type II error
Decision: Reject H_0	Type I error	Correct decision

The probability of type I error is denoted as α , and the probability of type II error is denoted as β . That is:

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true})$$

$$\beta = P(\text{Type II error}) = P(\text{Do not reject } H_0 | H_0 \text{ is false})$$

The type I error rate α is also called the **significance level** of a hypothesis test.

Example: Type I and Type II Errors

In a diabetes blood test, a patient is diagnosed with the disease if the sugar level in their bloodstream is larger than the threshold $C=130$ mg/dL. Suppose the distributions of sugar levels for the two populations (diabetes-free and having diabetes) are the two bell-shaped curves shown in the following figure.

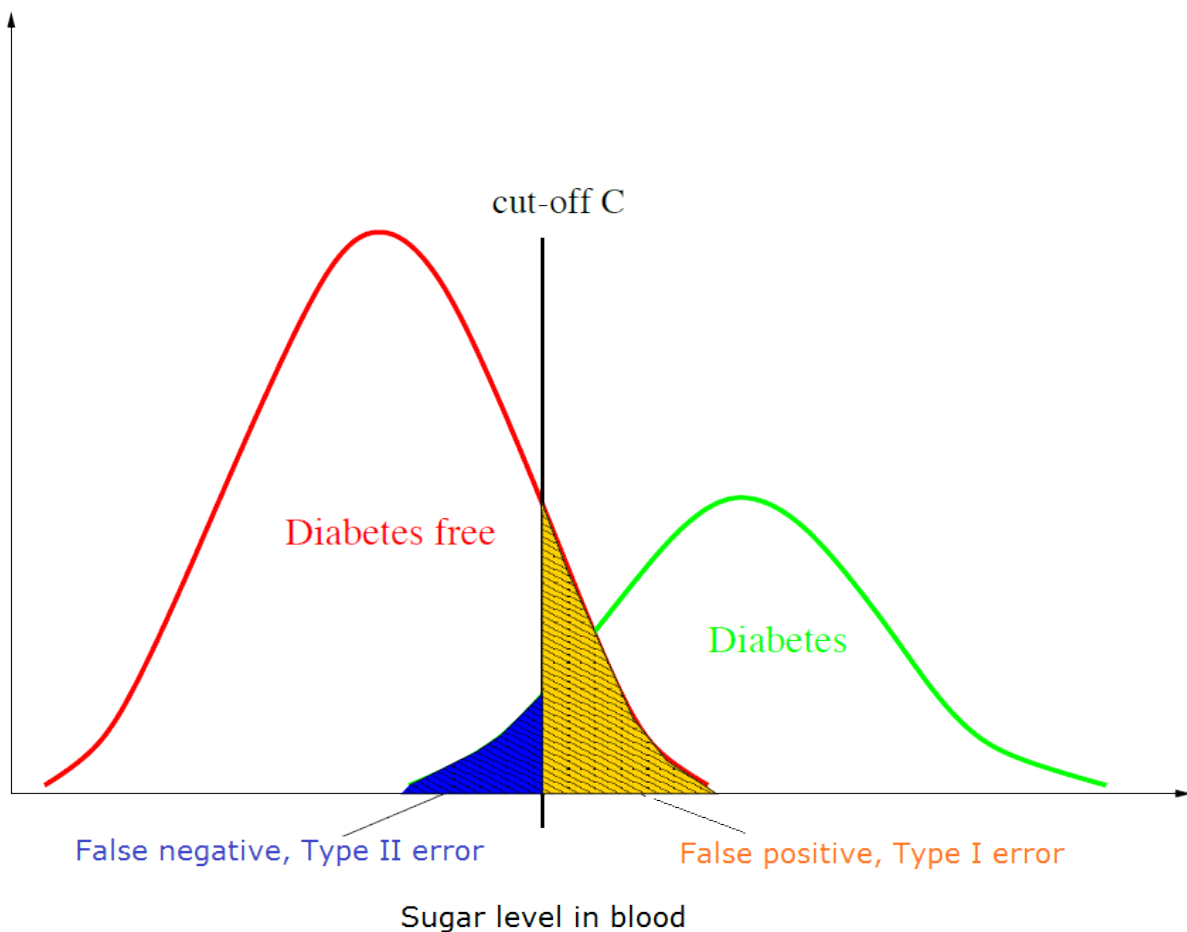


Figure 8.1: Trade-Off Between Type I and Type II Errors. [[Image Description \(See Appendix D Figure 8.1\)](#)]

Define the hypotheses: H_0 (a patient is disease free) vs. H_a (a patient has diabetes). What are the type I and type II errors in this example?

Type I error: claim the person has diabetes (reject the null H_0) but actually the person does not have diabetes (H_0 is in fact true). This is often referred to as a false positive. Type II error: claim the person does not have diabetes (do not reject the null H_0), but actually the person has diabetes (H_0 is false). This is often referred to as a false negative.

The figure in the above example shows the trade-off between type I and type II errors. The gold area gives α , the probability of the type I error; and the blue area gives β , the probability of the type II error. If we increase the threshold C (move the cut-off to the right), the gold area will reduce and the blue area will increase. That is the type I error rate α

will decrease and the type II error rate β will increase. On the other hand, if we reduce the threshold C (move the cut-off to the left), the type I error rate α will increase and the type II error rate will decrease. This is the trade-off between the type I and type II errors α and β . It is not a good idea to set either α or β to be too close to 0; otherwise, the other error rate will be huge. For example, if we set the threshold C very large, few individuals will be diagnosed as diabetic; as a result, many diabetic individuals will be misclassified as not having the disease (meaning we have a high probability of committing a type I error). On the other hand, if we set the threshold C very small, most individuals will be diagnosed as diabetic; consequently, many individuals who are free of diabetes will be misclassified as diabetic (meaning we have a high probability of committing a type II error). In general, we can set α (or β) to be relatively small if the consequence of the type I (or type II) error is more serious. The **power** of a test is defined as

$$1 - \beta = 1 - P(\text{Type II error}) = 1 - P(\text{Do not reject } H_0 | H_0 \text{ false}) = P(\text{Reject } H_0 | H_0 \text{ false}).$$

This is the probability that we reject H_0 when H_0 is false. Thus, it is of interest for a statistical test to have a high level of power.



Activity

Exercise: Type I and Type II Errors

Suppose you are performing a statistical test to decide whether a nuclear reactor should be approved. The null hypothesis is that the reactor is safe to use, and so failing to reject the null hypothesis corresponds to approval.

- Write down the null and alternative hypotheses.
- What are the type I and type II errors in this example?
- Which error has a more serious consequence, type I or type II? Which of α or β should be smaller?

Show/Hide Answer

Answers:

- H_0 : the nuclear reactor is safe versus H_a : the nuclear reactor is not safe.
- Type I error: disapprove the nuclear reactor for use given that the nuclear reactor is actually safe.
Type II error: approve the nuclear reactor for use given that the nuclear reactor is not safe.
- The type II error is more serious than the type I. Disapproving a safe reactor would waste time and money, but approving an unsafe reactor could lead to a nuclear meltdown, which is a catastrophic event. For this reason, we should set the type II error rate β to be relatively small.

8.3 Main Idea Behind Hypothesis Tests for μ

The main idea of a hypothesis test is to use the data as evidence to disprove the null H_0 and thus prove that the alternative H_a is true. The idea behind a hypothesis test for the population mean is as follows:

Collect from the population a simple random sample: x_1, x_2, \dots, x_n and calculate the sample mean $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$. Our “evidence” stems from the discrepancy between the point estimate \bar{x} and the hypothesized population mean μ_0 .

Reject the null hypothesis H_0 if the sample mean \bar{x} does not support the null H_0 . That is, we should reject H_0 if \bar{x} is too extreme. The word “extreme” means contradicting the null H_0 in favour of the alternative H_a .

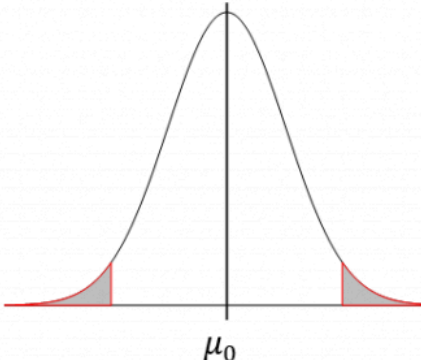
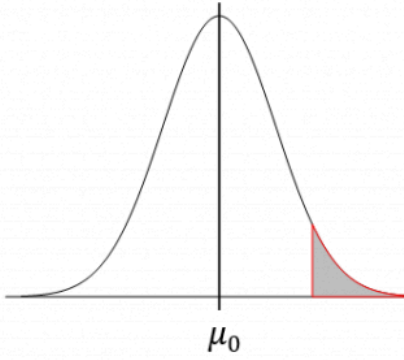
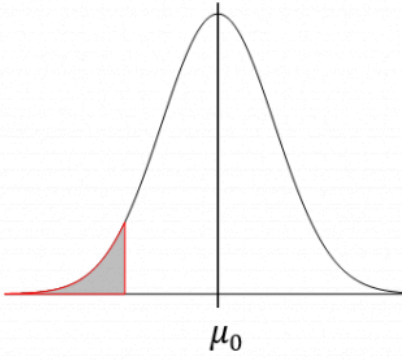
$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
		
Reject H_0 if \bar{x} is either too large or too small, i.e., falls in the shaded area.	Reject H_0 if \bar{x} is too large, i.e., falls in the shaded area.	Reject H_0 if \bar{x} is too small, i.e., falls in the shaded area.

Figure 8.2: Rejection Region Based on Sample Mean. [[Image Description \(See Appendix D Figure 8.2\)](#)]

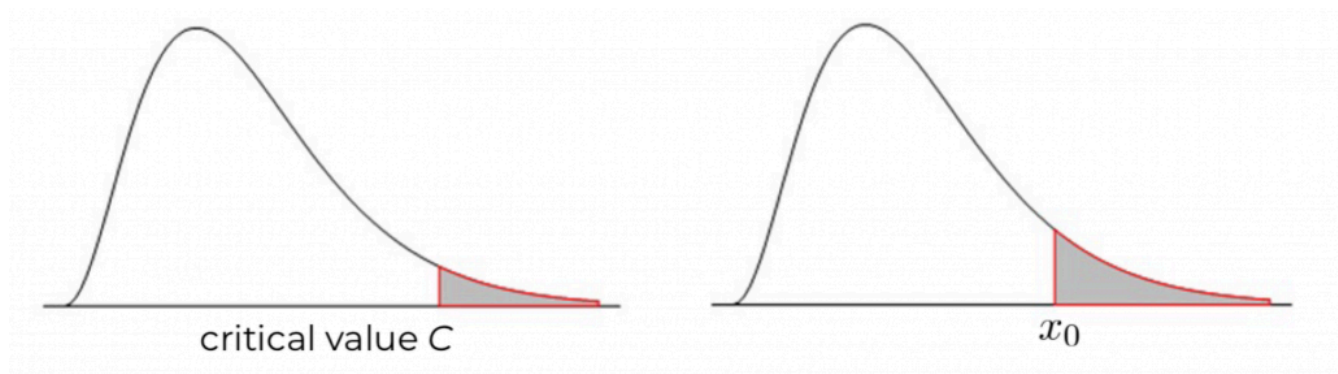
In order to quantify how the data (our evidence) contradict the null hypothesis, we first assume the null hypothesis H_0 is true and **calculate the chance of observing a sample mean at least as extreme as the observed \bar{x}** . Reject the null H_0 if the chance is small; otherwise, fail to reject H_0 . Recall that for a normal population or a large sample size,

the sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ and $t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t$ distribution with $df = n - 1$. We call the variables $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ or $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ the test statistics. We should reject the null hypothesis H_0 if the observed test statistic $z_o = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ or $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is too extreme.

8.4 Quantify the “Extremeness”

There are two ways to quantify the extremeness of the data under the assumption that the null hypothesis H_0 is true: the critical value approach and the P-value approach. These two methods will give the same conclusion. For example, Bill claims he is not rich, and we want to prove that he is lying. The steps are:

1. Write down the hypotheses. H_0 : Bill is not rich versus H_a : Bill is rich.
2. Collect the evidence and conclude. Suppose Bill has total wealth x_0 , and we know the total wealth for every adult in the world; then we can draw the population distribution of the wealth, which is assumed to be the following graph.
 - a. We can define the so-called **rejection region** by a cut-off C . Those with a total wealth at least C are defined as “rich” people, say the top 5%, meaning the shaded area (the left panel) is 0.05. Reject the null hypothesis H_0 if x_0 falls in the rejection region, i.e., $x_0 \geq C$, meaning Bill is one of those top 5% rich people. Note that the null hypothesis is H_0 Bill is not rich. Rejecting H_0 implies Bill is rich.
 - b. We can also find the percentage of people at least as rich as Bill; that is the area to the right of x_0 , the shaded area in the right panel. We call this area the **P-value**. We should reject the null hypothesis H_0 if the P-value is small. The smaller the area (P-value), the fewer people richer than Bill, the stronger the evidence that Bill is rich.



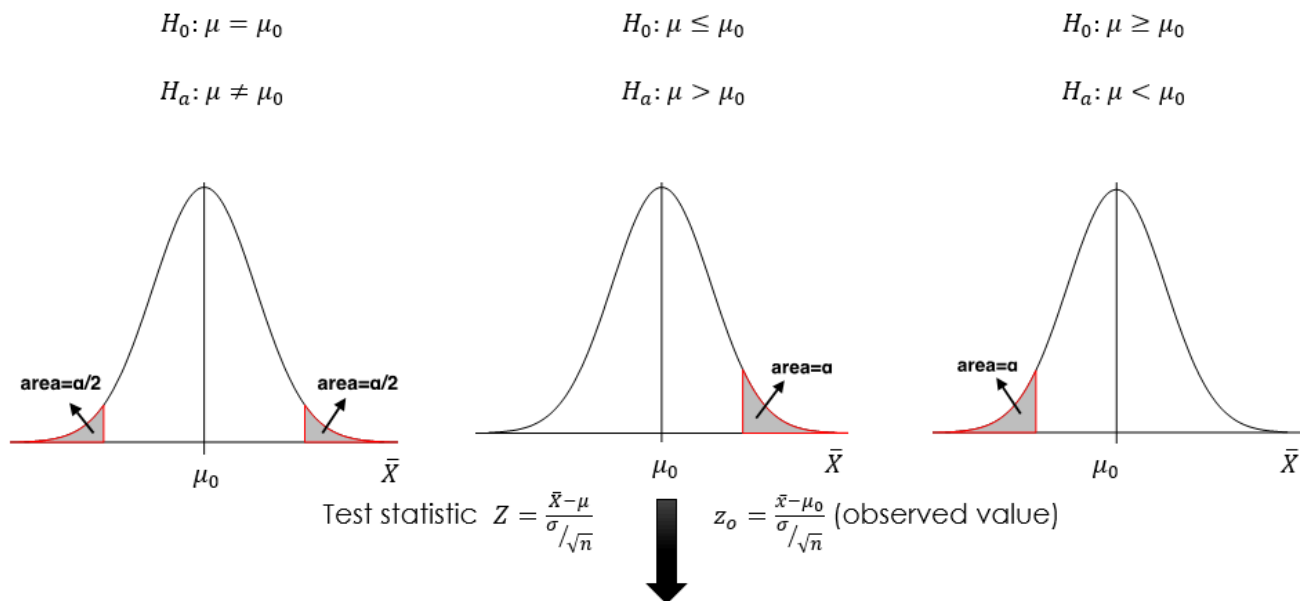
The shaded area is the rejection region.

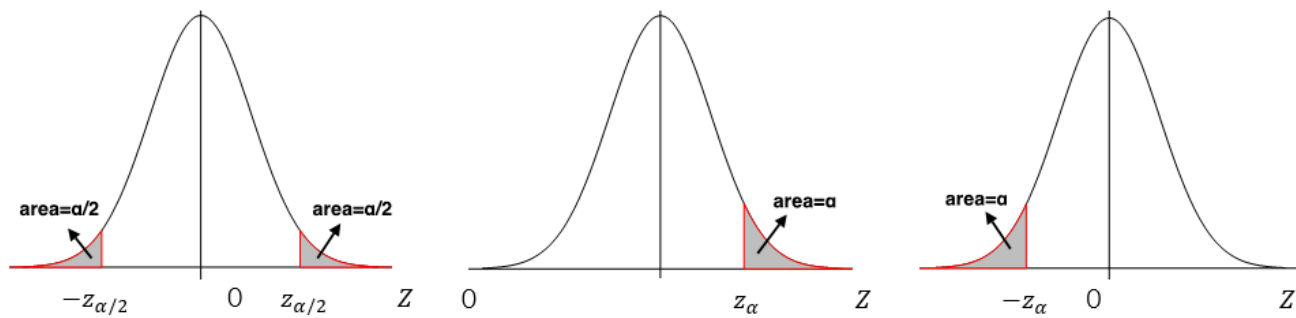
The shaded area is the P-value.

Figure 8.3: Rejection Region (left panel) and P-value (right panel). [[Image Description \(See Appendix D Figure 8.3\)](#)]

8.4.1 The Critical Value Approach

Recall that the main idea of hypothesis tests is to reject the null hypothesis H_0 if the sample mean \bar{x} is too extreme, i.e., we should reject H_0 if \bar{x} falls in the rejection region. If the population standard deviation σ is known, the observed test statistic is $z_o = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$. If \bar{x} is too extreme, the corresponding test statistic z_o will also be too extreme. Since the standardized variable follows a standard normal distribution, we can use the standard normal density curve to define the rejection region. **The key point for the critical value approach is that the total area of the rejection region equals the significance level of the test α .** The values dividing the density curve into rejection and non-rejection regions are called the **critical values**, such as z_{α} , $z_{\alpha/2}$, $-z_{\alpha}$, and $-z_{\alpha/2}$.





Reject H_0 if \bar{x} or z_o falls in the rejection region (the shaded area).

Reject H_0 if $z_o \geq z_{\alpha/2}$ or $z_o \leq -z_{\alpha/2}$

Reject H_0 if $z_o \geq z_{\alpha}$

Reject H_0 if $z_o \leq -z_{\alpha}$

Figure 8.4: Critical Values and Rejection Regions of One-Sample Z Test. [[Image Description \(See Appendix D Figure 8.4\)\]](#)

8.4.2 The P-value Approach

Another way to quantify the “extremeness” of the sample average is the P-value approach. We should reject the null H_0 if $\text{P-value} \leq \alpha$. The P-value is the probability that the test statistic is at least as extreme as the observed statistic, given that the null hypothesis is true. The P-value is a measure of evidence against H_0 in favour of H_a . **The smaller the P-value, the stronger the evidence. A small P-value indicates that the observed value of the test statistic is very unlikely if the null is true.** We should, therefore, reject the null hypothesis if the $\text{P-value} \leq \alpha$, where α is the significance level of the test. For example, if $\text{P-value} = 0.03$, we reject the null if the significance level is $\alpha = 0.1$, or 0.05 but not for $\alpha = 0.01$. Here are some important facts about the P-value:

- P-value is a probability; therefore, it must be between 0 and 1.
- P-value is a conditional probability, given that the null H_0 is true. **Note that the P-value is NOT the probability that the null is true, which is a common mistake.**
- The P-value is a measure of evidence against H_0 in favour of H_a .
- Therefore, when performing a one-sided test, the direction of the inequality in the P-value calculation should be in the same direction as the inequality in the alternative H_a .
- The P-value of a two-sided test is twice that of a one-sided test with the same value of test statistic.

Calculation of the P-value:

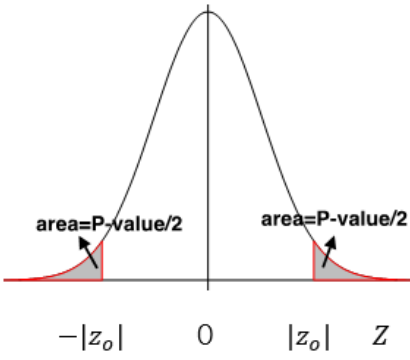
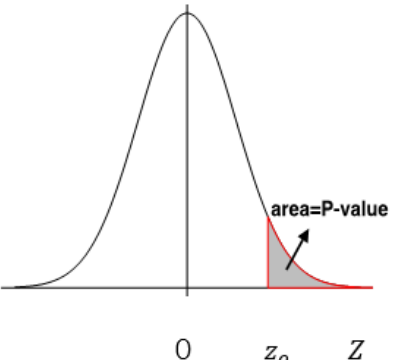
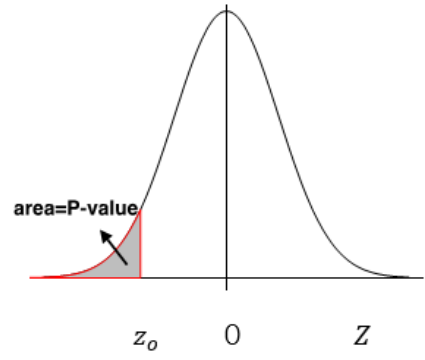
$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
		
$\text{P-value} = P(Z \geq z_o) + P(Z \leq - z_o) = 2P(Z \geq z_o)$	$\text{P-value} = P(Z \geq z_o)$	$\text{P-value} = P(Z \leq z_o)$
Reject H_0 if the P-value (the shaded area) $\leq \alpha$.		

Figure 8.5: P-Value of One-Sample Z Test. [[Image Description \(See Appendix D Figure 8.5\)\]](#)

8.5 Hypothesis Tests for One Population Mean μ

Recall that there are two different procedures used to construct confidence intervals for one population mean μ : the one-sample Z-interval (used when the population standard deviation σ is known) and the one-sample t-interval (used when σ is unknown). In a similar vein, there are two different procedures for hypothesis tests for one population mean: the one-sample Z-test is used when σ is known and the one-sample t-test is used when σ is unknown.

8.5.1 One-Sample Z-Test When σ is Known

Assumptions:

1. A simple random sample (SRS)
2. Normal population or large sample size ($n \geq 30$)
3. The population standard deviation σ is known

Steps:

1. Set up the hypotheses (one of the following three pairs):

Two tailed	Right tailed	Left tailed
$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$

2. State the significance level α .
3. Compute the value of the test statistic: $z_o = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.
4. Find the P-value or rejection region.

	Two tailed	Right tailed	Left tailed
H_0	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
H_a	$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$
P-value	$2P(Z \geq z_o)$	$P(Z \geq z_o)$	$P(Z \leq z_o)$
Rejection region	$Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$	$Z \geq z_{\alpha}$	$Z \leq -z_{\alpha}$

5. Reject the null H_0 if P-value $\leq \alpha$ or z_o falls in the rejection region.
6. Conclusion.

Example: One-Sample Z Test

One-Sample Z Test

A machine fills beer into bottles whose volume is supposed to be 341 ml, but the exact amount varies from bottle to bottle. We randomly picked 100 bottles and obtained the sample mean volume of 339 ml. Assume the population standard deviation $\sigma = 5$ ml. Test at the 5% significance level whether the machine is NOT working properly.

Check the assumptions:

1. We have a simple random sample (SRS).
2. We do not know whether the population is normal or not, but the sample size is large with $n = 100 \geq 30$.
3. $\sigma = 5$ ml is known.

Steps:

1. Set up the hypotheses: $H_0 : \mu = 341$ ml versus $H_a : \mu \neq 341$ ml.
This is a two-tailed test. If the machine works properly, the population mean volume $\mu = 341$ ml.
2. The significance level is $\alpha = 0.05$.
3. Compute the value of the test statistic:

$$z_o = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{339 - 341}{5 / \sqrt{100}} = \frac{-2}{0.5} = -4.$$

4. Find the P-value. For a two-tailed test with the observed test statistic $z_o = -4$.

$$P\text{-value} = 2P(Z \leq -4) \approx 2 \times 0 = 0.$$

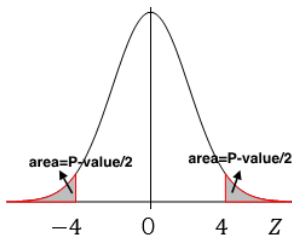



Table II: Area under the standard normal curve for negative z



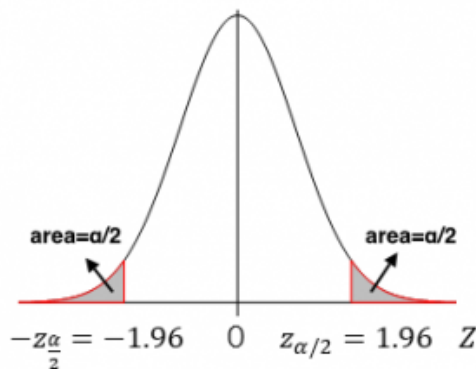
Second decimal place in z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	z
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-3.9
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	-3.8
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	-3.7
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	-3.6
0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	-3.5

[\[Image Description \(See Appendix D Example 8.1\)\]](#) Click on the image to enlarge itThe .

5. Decision: Since the P-value $\approx 0 \leq 0.05(\alpha)$, reject the null hypothesis H_0 .
6. Conclusion: At the 5% significance level, the data provide sufficient evidence that the machine is NOT working properly.

If using the critical value approach, steps 1-3 are the same, steps 4-6 become:

4. Rejection region:



[\[Image Description \(See Appendix D Example 8.2\)\]](#)

$\alpha = 0.05$, $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$.
For a two-tailed test, the critical values are -1.96 and 1.96.
The rejection region is either greater than 1.96 or smaller than -1.96.

5. Decision: Since the observed value $z_o = -4 < -1.96$ falls in the rejection region, we reject the null hypothesis H_0 .
6. Conclusion: At the 5% significance level, the data provide sufficient evidence that the machine is NOT working properly.



Instructor's Note

P-value approach is preferred for the following reasons:

1. It is more professional. P-value is required to be reported for all hypothesis tests in academia.
2. The P-value approach provides more information: it not only tells whether we should reject the null or not but also shows how strong the evidence is. However, the critical value approach only tells us whether we should reject the null or not.
3. The computer output only provides the P-value; no critical value is provided.

8.5.2 One-Sample t -Test When σ is Unknown

Assumptions:

1. A simple random sample (SRS)
2. Normal population or large sample size ($n \geq 30$)
3. The population standard deviation σ is unknown

Steps:

1. Set up the hypotheses:

Two tailed	Right tailed	Left tailed
$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$

2. State the significance level α .
3. Compute the value of the test statistic: $t_o = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ with a degree of freedom $df = n - 1$.
4. Use the t score table (Table IV) to find the P-value or rejection region.

	Two tailed	Right tailed	Left tailed
H_0	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
H_a	$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$
P-value	$2P(t \geq t_o)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_{\alpha}$	$t \leq -t_{\alpha}$

Note: Unlike the p-value of a one-sample Z test when σ is known, in general, only a range of values rather than an exact number will be obtained for a one-sample t test using Table IV.

An exact p-value will be obtained only when the observed test statistic is one of those twelve t-scores given in the table for a given degree of freedom.

5. Reject the null H_0 if P-value $\leq \alpha$ or t_o falls in the rejection region.
6. Conclusion.

Example: One-Sample t Test

A computer company claims that the average lifetime of its laptop is about 4 years. A simple random sample of 36 laptops yields an average lifetime of 3.5 years with a sample standard deviation of 4.2 years. Test at the 1% significance level whether the mean lifetime of this brand of laptops is less than 4 years.

Check the assumptions:

1. We have a simple random sample (SRS).
2. We do not know whether the population is normal or not, but the sample size is large with $n = 36 \geq 30$.
3. σ is unknown and estimated by $s = 4.2$.

Steps:

1. Set up the hypotheses: $H_0 : \mu \geq 4$ years versus $H_a : \mu < 4$ years.
2. The significance level is $\alpha = 0.01$.
3. Compute the value of the test statistic:
$$t_o = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.5 - 4}{4.2/\sqrt{36}} = \frac{-0.5}{0.7} = -0.714 \text{ with } df = n - 1 = 36 - 1 = 35.$$
4. Find the P-value. For a left-tailed test, the P-value is the area to the left of the observed test statistic t_o because the alternative H_a is "<".
P-value = $P(t \leq t_o) = P(t \leq -0.714) = P(t \geq 0.714) \approx 0.2$. To be more precisely,
 $0.2 < \text{p-value} < 0.3$.

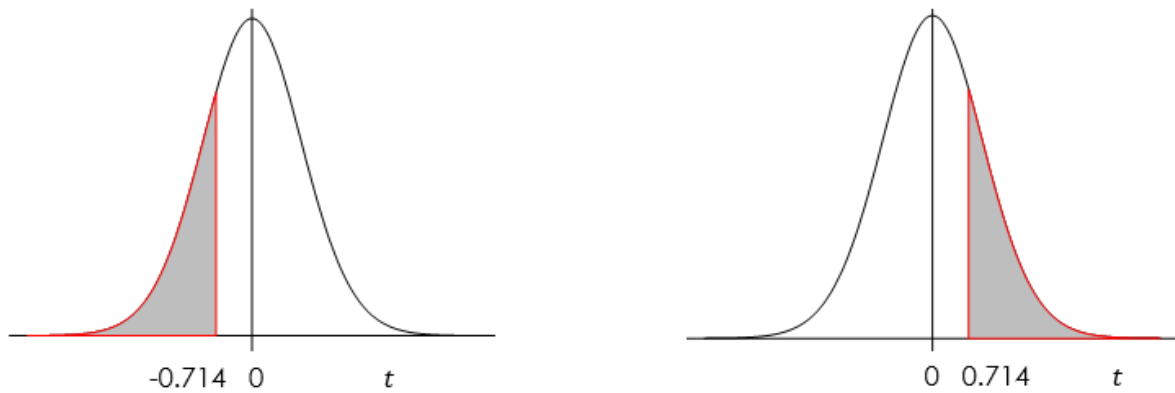
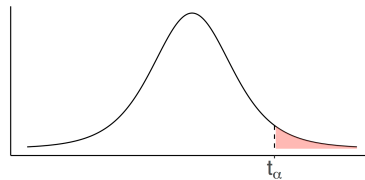


Table IV: Values of t_α of t -distribution



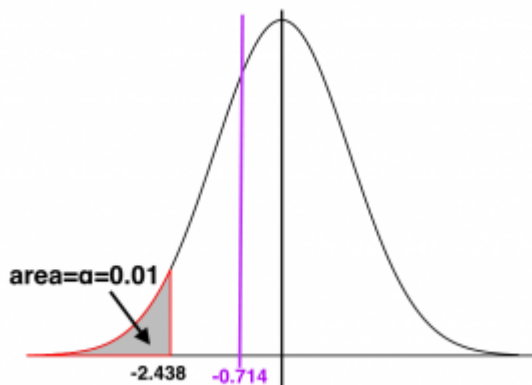
df	α : Area to the Right of t_α											
	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.010	0.0075	0.0050	0.0025	0.0005
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	31.821	42.433	63.657	127.321	636.619
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	6.965	8.073	9.925	14.089	31.599
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	3.365	3.634	4.032	4.773	6.869
31	0.256	0.530	0.853	1.054	1.309	1.696	2.040	2.453	2.576	2.744	3.022	3.633
32	0.255	0.530	0.853	1.054	1.309	1.694	2.037	2.449	2.571	2.738	3.015	3.622
33	0.255	0.530	0.853	1.053	1.308	1.692	2.035	2.445	2.566	2.733	3.008	3.611
34	0.255	0.529	0.852	1.052	1.307	1.691	2.032	2.441	2.562	2.728	3.002	3.601
35	0.255	0.529	0.852	1.052	1.306	1.690	2.030	2.438	2.558	2.724	2.996	3.591

[Image Description (See Appendix D Example 8.3)]

5. Decision: Since the P-value $0.2 > 0.01(\alpha)$, we can not reject the null H_0 .
6. Conclusion: At the 1% significance level, we do not have sufficient evidence that the mean lifetime of this brand of laptops is less than 4 years.

If we use the critical value approach, steps 1-3 are the same, and steps 4-6 become:

4. Rejection region



$$df = n - 1 = 36 - 1 = 35,$$

$$\alpha = 0.01, t_{\alpha} = t_{0.01} = 2.438$$

For a left-tailed test, the critical value is $-t_{\alpha} = -t_{0.01} = -2.438$.

[Image Description (See Appendix D Example 8.4)]

5. Decision: Since the observed value $t_o = -0.714 > -2.438$ falls in the non-rejection region, we can not reject the null hypothesis H_0 .
6. Conclusion: At the 1% significance level, the data do not provide sufficient evidence that the mean lifetime of this brand of laptops is less than 4 years.



Activity

Exercise: P-value for One sample t-Test

Use the same setting of the previous example (one-sample t-test with $df = 35$) to find the P-values of the following hypothesis tests.

- a. $H_0 : \mu = 4$ years versus $H_a : \mu \neq 4$ years, with the observed test statistic $t_o = 1.5$.
- b. $H_0 : \mu \geq 4$ years versus $H_a : \mu < 4$ years, with the observed test statistic $t_o = -2.5$.
- c. $H_0 : \mu \leq 4$ years versus $H_a : \mu > 4$ years, with the observed test statistic $t_o = 3.5$.

Show/Hide Answer

- a. For a **two-tailed** test, the P-value is **twice** the area to the right of the absolute value of the observed test statistic t_o . Note that the probability is the area under the density curve of the t-distribution with 35 degrees of freedom. $P\text{-value} = 2P(t \geq |t_o|) = 2P(t \geq 1.5)$. Since

$1.306(t_{0.1}) < 1.5 < 1.690(t_{0.05})$, we have

$0.05P(t \geq 1.5)0.1 \Rightarrow 2 \times 0.052P(t \geq 1.5)2 \times 0.1 \Rightarrow 0.1$ P-value 0.2. If use R commander, $2P(t \geq 1.5) = 2 \times 0.07129092 = 0.1425818$.

- b. For a **left-tailed** test, the P-value is the area to the **left** of the observed test statistic t_o . P -value $= P(t \leq t_o) = P(t \leq -2.5) = P(t \geq 2.5)$. Since

$2.438(t_{0.01}) < 2.5 < 2.558(t_{0.0075}) \Rightarrow 0.0075 < \text{P-value} < 0.01$. If use R commander, $P(t \geq 2.5) = 0.008627872$.

- c. For a **right-tailed** test, the P-value is the area to the right of the observed test statistic t_o . P -value $= P(t \geq t_o) = P(t \geq 3.5)$. Since

$(t_{0.0025})2.996 < 3.5 < 3.591(t_{0.0005}) \Rightarrow 0.0005 < \text{P-value} < 0.0025$. If use R commander, $P(t \geq 3.5) = 0.0006444197$.

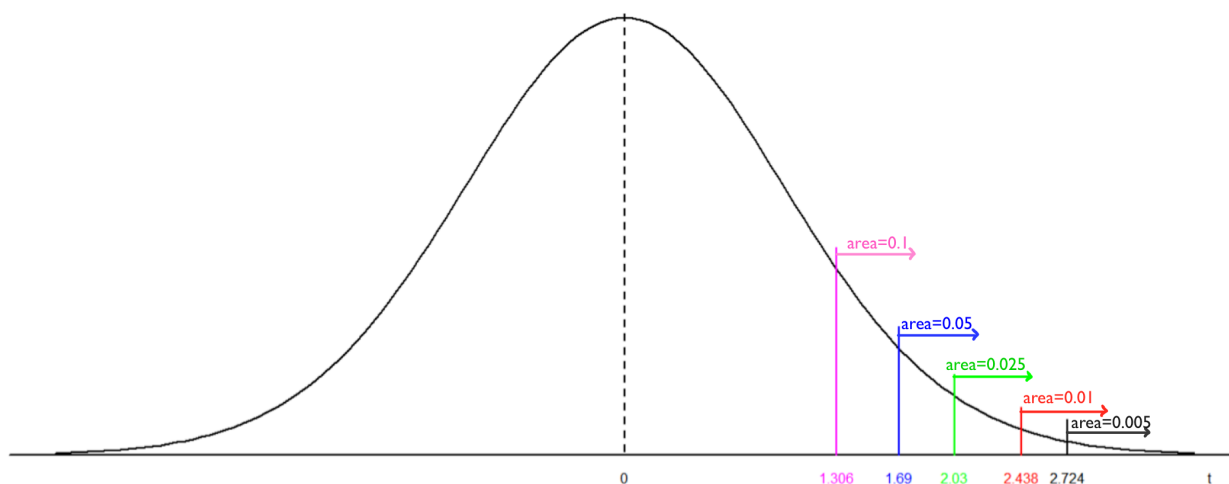


Figure 8.6: Selected Critical Values of t Distribution with $df=35$. [Image Description (See Appendix D Figure 8.6)]



Activity

Exercise: One-sample t -Test

The number of cell phone users has increased dramatically since 1997. Suppose the mean local monthly bill was \$50 for cell phone users in the United States in 2006. A simple random sample of 50

cell phone users was obtained in 2019, and the sample mean local monthly bill was $\bar{x} = 55$ with a sample standard deviation $s = 25$.

- At the 5% significance level, do the data provide sufficient evidence to conclude that the mean local monthly bill for cell phone users in 2019 has changed from the 2006 mean of \$50?
- Obtain a 95% confidence interval for the 2019 mean local monthly bill for all cell phone users. Interpret the confidence interval.
- Are the results in parts (a) and (b) consistent with each other? Explain why.

Show/Hide Answer

a. Check the assumptions:

- We have a simple random sample (SRS).
- We do not know whether the population is normal or not since we do not have the data, but the sample size is large with $n = 50 \geq 30$.
- σ is unknown and estimated by $s = 25$.

Steps:

- Set up the hypotheses: $H_0 : \mu = 50$ versus $H_a : \mu \neq 50$.
- The significance level is $\alpha = 0.05$.
- Compute the value of the test statistic: $t_o = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{55 - 50}{25/\sqrt{50}} = 1.414$ with $df = n - 1 = 50 - 1 = 49$.
- Find the P-value. For a two-tailed test, the P-value is twice the area to the right of the observed test statistic t_o .
P-value = $2P(t \geq t_o) = 2P(t \geq 1.414)$. Since $1.299(t_{0.1}) < 1.414 < 1.677(t_{0.05})$,
 $2 \times 0.05 < \text{P-value} < 2 \times 0.1 \implies 0.1 < \text{P-value} < 0.2$.
- Decision: Since the P-value $0.1 > 0.05(\alpha)$, we can not reject the null H_0 .
- Conclusion: At the 5% significance level, we do not have sufficient evidence that the 2019 mean local monthly bill for cell phone users has changed from the 2006 mean of \$50.

b. Steps:

- Find $t_{\alpha/2}$: $n = 50, df = n - 1 = 50 - 1 = 49$.
 $1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \alpha/2 = 0.025 \implies t_{\alpha/2} = t_{0.025} = 2.010$.
- Interval: $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 55 \pm 2.010 \times \frac{25}{\sqrt{50}} = (47.894, 62.106)$.

Interpretation: We can be 95% confident that the 2019 mean local monthly bill for cell phone users is somewhere between \$47.894 and \$62.106.

- Yes, they are consistent. We cannot reject $H_0 : \mu = 50$ and hence can not claim $\mu \neq 50$ in the hypothesis test in part (a). The interval in part (b) contains 50; there is no sufficient evidence that the population mean differs from 50. We cannot reject $H_0 : \mu = 50$ and claim $\mu \neq 50$. Therefore, they are consistent.

8.6 Relationship Between Confidence Intervals and Hypothesis Tests

Confidence intervals (CI) and hypothesis tests should give consistent results: we should not reject H_0 at the significance level α if the corresponding $(1 - \alpha) \times 100\%$ confidence interval contains the hypothesized value μ_0 . Two-sided confidence intervals correspond to two-tailed tests, upper-tailed confidence intervals correspond to right-tailed tests, and lower-tailed confidence intervals correspond to left-tailed tests.

A $(1 - \alpha) \times 100\%$ two-sided t confidence interval is given in the form $(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}})$. A $(1 - \alpha) \times 100\%$ upper-tailed t confidence interval is given by $(\bar{x} - t_{\alpha} \frac{s}{\sqrt{n}}, \infty)$ and the number $\bar{x} - t_{\alpha} \frac{s}{\sqrt{n}}$ is called the lower bound of the interval. A $(1 - \alpha) \times 100\%$ lower-tailed t confidence interval is given by $(-\infty, \bar{x} + t_{\alpha} \frac{s}{\sqrt{n}})$ and the number $\bar{x} + t_{\alpha} \frac{s}{\sqrt{n}}$ is called the upper bound of the interval. We can also use confidence intervals to make conclusions about hypothesis tests: reject the null hypothesis H_0 at the significance level α if the corresponding $(1 - \alpha) \times 100\%$ confidence interval does not contain the hypothesized value μ_0 . The relationship is summarized in the following table.

Table 8.3: Relationship Between Confidence Interval and Hypothesis Test

Null hypothesis	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
Alternative	$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$
$(1 - \alpha) \times 100\%$ CI	$(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}})$	$(\bar{x} - t_{\alpha} \frac{s}{\sqrt{n}}, \infty)$	$(-\infty, \bar{x} + t_{\alpha} \frac{s}{\sqrt{n}})$
Decision	Reject H_0 if μ_0 is outside the interval		



Instructor's Note

Here is the reason we should reject H_0 if μ_0 is outside the corresponding confidence interval. Take the right-tailed test for example, we should reject H_0 if the observed test statistic

t_o falls in the rejection region, that is if $t_o \geq t_\alpha$. This implies $t_o = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq t_\alpha \implies \mu_0 \leq \bar{x} - t_\alpha \frac{s}{\sqrt{n}}$. Given that the upper-tailed confidence interval for a right-tailed test is $(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \infty)$, $\mu_0 \leq \bar{x} - t_\alpha \frac{s}{\sqrt{n}}$ means the value of μ_0 is outside the confidence interval. The same rationale applies to two-tailed and left-tailed tests. Therefore, we can reject H_0 at the significance level α if μ_0 is outside the corresponding $(1-\alpha) \times 100\%$ confidence interval.

Example: Relationship Between Confidence Intervals and Hypothesis Tests

The ankle-brachial index (ABI) compares the blood pressure of a patient's arm to the blood pressure of the patient's leg. The ABI can be an indicator of different diseases, including arterial diseases. A healthy (or normal) ABI is 0.9 or greater. Researchers obtained the ABI of 100 women with peripheral arterial disease and obtained a mean ABI of 0.64 with a standard deviation of 0.15.

- a. At the 5% significance level, do the data provide sufficient evidence that, on average, women with peripheral arterial disease have an unhealthy ABI?

Steps:

1. Set up the hypotheses: $H_0 : \mu \geq 0.9$ versus $H_a : \mu < 0.9$.
2. The significance level is $\alpha = 0.05$.
3. Compute the value of the test statistic: $t_o = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.64 - 0.9}{0.15/\sqrt{100}} = \frac{-0.26}{0.015} = -17.333$ with $df = n - 1 = 100 - 1 = 99$ (not given in Table IV, use 95, the closest one smaller than 99).
4. Find the P-value. For a left-tailed test, the P-value is the area to the left of the observed test statistic t_o . $P\text{-value} = P(t \leq t_o) = P(t \leq -17.333) = P(t \geq 17.333) < 0.005$, since $17.333 > 2.629(t_{0.005})$.
5. Decision: Since the P-value $< 0.005 < 0.05(\alpha)$, we should reject the null hypothesis H_0 .
6. Conclusion: At the 5% significance level, the data provide sufficient evidence that, on average, women with peripheral arterial disease have an unhealthy ABI.

- b. Obtain a confidence interval corresponding to the test in part a).

For a **left-tailed** test at the significance level $\alpha = 0.05$, we should obtain a $(1 - \alpha) \times 100\% = 95\%$ **lower-tailed** interval. For $df = 99$, not given in Table IV, use $df = 95$, $t_\alpha = t_{0.05} = 1.661$

$$\left(-\infty, \bar{x} + t_\alpha \frac{s}{\sqrt{n}}\right) = \left(-\infty, 0.64 + 1.661 \times \frac{0.15}{\sqrt{100}}\right) = (-\infty, 0.665).$$

Interpretation: We are 95% confident that women with peripheral arterial disease have an average ABI below 0.665.

- c. Does the interval in part b) support the conclusion in part a)?

In part a), we reject H_0 and claim that the mean ABI is below 0.9 for women with peripheral arterial disease. In part b), we are 95% confident that the mean ABI is less than 0.9 since the entire confidence interval is below 0.9. In other words, the hypothesized value 0.9 is outside the

corresponding confidence interval, we should reject the null. Therefore, the results obtained in parts a) and b) are consistent.

8.7 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Write down the null and alternative hypotheses for a study (Section 8.1).
- Explain the difference between the type I and the type II errors and the relationship between these two types of errors (Section 8.2).
- Define the P-value, both mathematically and in plain English (Section 8.4).
- Distinguish between the one-mean Z test and the one-mean *t*-test, and identify when each test should be used (Section 8.5).
- Conduct a one-mean Z and a one-mean *t*-test using the P-value or critical value approaches (Section 8.5).
- Calculate the P-value or the range of the P-value for a hypothesis test (Section 8.5).
- Explain the relationship between confidence intervals and hypothesis tests. (Section 8.6).

8.8 Review Questions

- Determine whether the following interpretations of a 95% confidence interval (337, 343) ml for the population mean volume of beer μ are true or false. If false, correct it.
 - We can be 95% confident that μ is somewhere between 337 ml and 343 ml.
 - We can be 95% confident that the sample mean \bar{x} is somewhere between 337 and 343 ml.
 - The probability that the population mean μ is within the interval (337, 343) is 0.95.
 - The probability that the sample mean \bar{x} is within the interval (337, 343) is 0.95.
 - 95% of the \bar{x} values are within the interval (337, 343).
- Determine whether the following statements about the P -value are true or false. If false, correct it.
 - We should reject the null hypothesis H_0 if the P -value $\leq \alpha$.
 - We should accept the null H_0 if the P -value $> \alpha$.
 - P -value is the probability that the null H_0 is true.
 - P -value is the probability of rejecting H_0 .
- Suppose you perform a statistical test to decide whether a nuclear reactor should be approved. Further, suppose that failing to reject the null hypothesis (the reactor is safe to use) corresponds to approval.
 - Write down the null and alternative hypotheses.
 - What are the type I and type II errors in this example?
 - Which error has more serious consequence, type I or type II? Would you like to set α or β to be relatively small?
- The mean retail price of agriculture books in 2005 was \$57.61. This year's retail mean price for 28 randomly selected agriculture books was \$54.97. Assume that the population standard deviation of prices for this year's agriculture books is \$8.45.
 - At the 10% significance level, do the data provide sufficient evidence to conclude that this year's mean retail price of agriculture books has changed from the 2005 mean?
 - What is the P -value of the test in part (a)?
 - Obtain a confidence interval corresponding to the test in part (a).
 - Does the interval obtained in part (c) support the result in part (a)?
- The ankle brachial index (ABI) compares the blood pressure of a patient's arm to the blood pressure of the patient's leg. The ABI can be an indicator of different diseases, including arterial diseases. A healthy (or normal) ABI is 0.9 or greater. Researchers obtained the ABI of 100 women with peripheral arterial disease and obtained a mean

ABI of 0.64 with a standard deviation of 0.15.

- a. At the 5% significance level, do the data provide sufficient evidence that, on average, women with peripheral arterial disease have an unhealthy ABI?
- b. What is the P -value of the test in part (a)?
- c. Obtain a confidence interval corresponding to the test in part (a).
- d. Does the interval obtained in part (c) support the conclusion in part (a)?

Show/Hide Answer

1.

- a. True, a standard way to interpret the confidence interval.
- b. False. The sample mean \bar{x} is the center of the interval; we should be 100% confident that the sample mean x is in the interval.
- c. False. There is no randomness here and hence there is no probability, since the population mean μ is a constant and the interval (337, 343) is also fixed. μ is either within the interval or outside the interval.
- d. False. Similar arguments as the previous. The sample mean $\bar{\mu}$ is a fixed number, and the interval is also fixed; there is no randomness.
- e. False. 95% of the $\bar{\mu}$ values are within the interval $(\mu - 1.96 \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \frac{\sigma}{\sqrt{n}})$.

2.

- a. True.
- b. False. In general, never accept H_0 .
- c. False. P-value measures the strength of the evidence that the data contradicts H_0 and is in favour of H_a .
- d. False. P-value is the probability of observing $z_0(t_0)$ or more extreme values.

3.

- a. H_0 : the nuclear reactor is safe versus H_a : the nuclear reactor is not safe.
- b. Type I error: disapprove of the nuclear reactor of ruse given that the nuclear reactor is actually safe.
Type II error: approve the nuclear reactor for use given that the nuclear reactor is not safe.
- c. Type II error is more severe than type I. We probably need to set the type II error rate relatively small.

4.

- a. Assumptions:
We have a simple random sample.
We have a large sample with $n = 100 > 30$.
Population standard deviation σ is unknown.

We can use a one-sample t-test. Summarize the information:

$n = 100$, $\bar{x} = 0.64$, $s = 0.15$. The six steps to perform a one-sample t-test are:

1. Hypotheses: $H_0 : \mu \geq 0.9$ versus $H_a : \mu < 0.9$.

2. The significance level $\alpha = 0.05$.

3. Observed test statistic:

$$t_o = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.64 - 0.9}{0.15/\sqrt{100}} = -17.333$$

with $df = n - 1 = 99$.

4. A left-tailed test, P-value = $P(t \leq t_o) = P(t \leq -17.333) = P(t \geq 17.333) < 0.005$.

5. Since P-value $< 0.005 < 0.05(\alpha)$, we reject H_0 .

6. At the 5% significance level, we have sufficient evidence that, on average, women with peripheral arterial disease have an unhealthy ABI

b. P-value < 0.005

c. A left-tailed test at significance level $\alpha = 0.05$ corresponds to a $(1 - \alpha) \times 100$ lower-tailed confidence interval. With $df = 99$ not given in Table IV, use $df = 90$ the closed one but still no more than 99, $\alpha = 0.05 \implies t_\alpha = t_{0.05} = 1.662$

$$(-\infty, \bar{x} + t_\alpha \frac{s}{\sqrt{n}}) = (-\infty, 0.64 + 1.662 \times \frac{0.15}{\sqrt{100}}) = (-\infty, 0.665).$$

Interpretation: we can be 95% confident that the mean ABI of women with peripheral arterial disease is somewhere below 0.665.

Note: In this course, you are only required to know how to obtain a two-tailed interval. For $df=99$ (use 90), $t_{\alpha/2} = t_{0.025} = 1.987$. The 95% two-tailed interval is $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 0.64 \pm 1.987 \times \frac{0.15}{\sqrt{100}} = (0.610, 0.670)$.

Interpretation: we can be 95% confident that the mean ABI of women with peripheral arterial disease is somewhere between 0.610 and 0.670. The entire interval is below 0.9, so we can claim $\mu < 0.9$.

d. Yes, since a healthy (or normal) ABI is 0.9 or greater; however, 0.9 is outside the confidence interval (it is above the entire interval), so we can claim that the mean ABI of women with peripheral arterial disease is below 0.9, i.e., $\mu < 0.9$. This is consistent with the conclusion of the t-test in part a).

8.9 Assignment 8

Purposes

This assignment has two parts. The first part assesses your knowledge of conducting a one-sample z test and a one-sample t test, explaining and calculating the P-value of a hypothesis test, stating the Type I and Type II errors of a hypothesis test, and explaining the relationship between the results of a confidence interval and a hypothesis test. The second part assesses your skills in using R commander to conduct a one-sample t test for the population mean μ .

Resources

[M08_Age_Millionaire_Q4.xlsx](#)

[M08_BloodPressure_Diabete_Q6.xlsx](#)

[M08_Hour_Q7.xlsx](#)

Instructions

Part A

Complete the following:

1. The following statement appeared on a box of Tide laundry detergent: "Individual packages of Tide may weigh slightly more or less than the marked weight due to normal variations incurred with high-speed packaging machines, but each day's production of Tide will average slightly above the marked weight."
 - a. Explain in statistical terms what the statement means. (2 marks)
 - b. Suppose that the marked weight is 1 liter. State in words the null and alternative hypotheses for the hypothesis test. (2 marks)
 - c. State the Type I and Type II errors in the context of this application. (4 marks)
 - d. Propose procedures to collect data to test the statement. (4 marks)

2. For a fixed sample size, what happens to the probability of a Type II error if the significance level is decreased from 0.05 to 0.01? (2 marks)
3. For a one-sample t test with $df = 40$, find the P-values of the following hypothesis tests. (9 marks: 3+3+3)
 - a. $H_0 : \mu = 4$ years versus $H_a : \mu \neq 4$ years, with the observed test statistic $t_o = 1.5$.
 - b. $H_0 : \mu \geq 4$ years versus $H_a : \mu < 4$ years, with the observed test statistic $t_o = -2.5$.
 - c. $H_0 : \mu \leq 4$ years versus $H_a : \mu > 4$ years, with the observed test statistic $t_o = 3.5$.
4. The following table gives the age (in years) of 36 randomly selected U.S. millionaires. The sample mean $\bar{x} = 58.53$ years. Assume that the standard deviation of ages of all U.S. millionaires is 13.0 years. (See data on file: **M08_Age_Millionaire_Q4.xlsx**)

31	45	79	64	48	38	39	68	52
59	68	79	42	79	53	74	66	66
71	61	52	47	39	54	67	55	71
77	64	60	75	42	69	48	57	48

- a. Test at the 10% significance level whether the average year of all U.S. millionaires is above 55 years. (8 marks)
 - b. What is the P-value of the hypothesis test in part (a)? (2 marks)
 - c. Obtain a confidence interval corresponding to the hypothesis test in part (a). (4 marks)
 - d. Does the interval in part (c) support the result in part (a)? Explain your answer. (3 marks)
5. The mean retail price of agriculture books in 2005 was \$57.61. This year's retail mean price for 28 randomly selected agriculture books was \$54.97. Assume that the population standard deviation of prices for this year's agriculture books is \$8.45.
 - a. At the 5% significance level, do the data provide sufficient evidence to conclude that this year's mean retail price of agriculture books has changed from the 2005 mean? (8 marks)
 - b. What is the P-value of the test in part (a)? (2 marks)
 - c. Obtain a confidence interval corresponding to the test in part (a). (4 marks)
 - d. Does the interval in part (c) support the result in part (a)? Explain your answer. (3 marks)
6. Past studies showed that maternal diabetes results in obesity, blood pressure, and glucose tolerance complications in the offspring. Following are the arterial blood pressures, in millimetres of mercury (mm Hg), for a random sample of 16 children of diabetic mothers. The sample mean is $\bar{x} = 85.99$ mm Hg and the sample standard deviation is $s = 8.08$ mm Hg. (See data on file: **M08_BloodPressure_Diabete_Q6.xlsx**)

81.6	84.1	87.6	82.8	82.0	88.9	86.7	96.4
84.6	101.9	90.8	94.0	69.4	78.9	75.2	91.0

- Test at the 5% significance level whether the mean arterial blood pressure of all children of diabetic mothers is above 85 mm Hg. (8 marks)
 - What is the P-value of the test in part (a)? (2 marks)
 - Obtain a confidence interval corresponding to the test in part (a). (4 marks)
 - Does the interval in part (c) support the result in part (a)? Explain your answer. (3 marks)
7. Previous studies showed that the average person watched 4.55 hours of television daily in 2005. A random sample of 20 people gave the following number of hours of television watched per day for last year. The sample mean is $\bar{x} = 4.76$ hours and the sample standard deviation is $s = 2.30$ hours. (See data on file: **M08_Hour_Q7.xlsx**)

1.0	4.6	5.4	3.7	5.2
1.7	6.1	1.9	7.6	9.1
6.9	5.5	9.0	3.9	2.5
2.4	4.7	4.1	3.7	6.2

- Test at the 1% significance level, do the data provide sufficient evidence to conclude that the amount of television watched per day last year by the average person differed from that in 2005? (8 marks)
- What is the P-value of the test in part (a)? (2 marks)
- Obtain a confidence interval corresponding to the test in part (a). (4 marks)
- Does the interval in part (c) support the result in part (a)? Explain your answer. (3 marks)

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs as required and write down your answers in statements.

- Refer to the data in Question 6 in Part A.
 - Use the proper graphical tools in R and R commander to assess whether applying the one-sample t test procedure is reasonable. Make sure to write down the assumptions of the procedure and address whether every assumption is satisfied. (5 marks)
 - Conduct a one-sample t test at the 5% significance level using R commander to test whether the mean arterial blood pressure of all children of diabetic mothers is

above 85 mm Hg. Make sure to include all the six components of a hypothesis test. Keep the computer outputs in an appendix. Compare the answer with the one you obtained by hand in Question 6 parts (a) and (b). (5 marks)

- c. Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare the one you obtained by hand in Question 6 part (c). (2 marks)

2. Refer to the data in Question 7 in Part A.

- a. Use the proper graphical tools in R and R commander to assess whether applying the one-sample t test procedure is reasonable. Make sure to write down the assumptions of the procedure and address whether every assumption is satisfied. (5 marks)
- b. Conduct a one-sample t test at the 1% significance level using R commander to test whether the amount of television watched per day last year by the average person differed from that in 2005. Make sure to include all the six components of a hypothesis test. **Paste the computer output into the space below first**, then compare the answer with the one you obtained by hand in Question 7 parts (a) and (b). (5 marks)
- c. Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare the one you obtained by hand in Question 7 part (c). (2 marks)

Quiz 8



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2632#h5p-10>

CHAPTER 9: INFERENCES FOR TWO POPULATION MEANS

Chapters 7 and 8 introduced how to obtain a confidence interval and perform a hypothesis test for the population mean μ based on a simple random sample from a population with mean μ . This chapter covers how to obtain a confidence interval and conduct a hypothesis test for the difference between population means $\mu_1 - \mu_2$.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the sampling distribution of the difference between two sample means $\bar{X}_1 - \bar{X}_2$ for two independent samples.
- State the assumptions for inferences about the difference between two population means based on two independent samples.
- Obtain and interpret a $(1 - \alpha) \times 100\%$ two-mean confidence interval for $\mu_1 - \mu_2$.
- Conduct a two-mean (sample) *t*-test.
- Determine whether two samples are independent or paired.
- Obtain and interpret a paired *t*-confidence interval.
- Conduct a paired *t*-test.
- Explain the relationship between the results of a hypothesis test at significance level α and the corresponding $(1 - \alpha) \times 100\%$ confidence interval.

9.1 Distribution of the Difference between Two Sample Means for Two Independent Samples

Suppose two populations have means μ_1, μ_2 and standard deviations σ_1, σ_2 . Further, suppose that we obtain from each population simple random samples, from which we obtain sample means \bar{x}_1 and \bar{x}_2 . Our objective is to make inferences about $\mu_1 - \mu_2$ using the unbiased estimate $\bar{x}_1 - \bar{x}_2$ and as such, we need to know the distribution of $X_1 - X_2$.

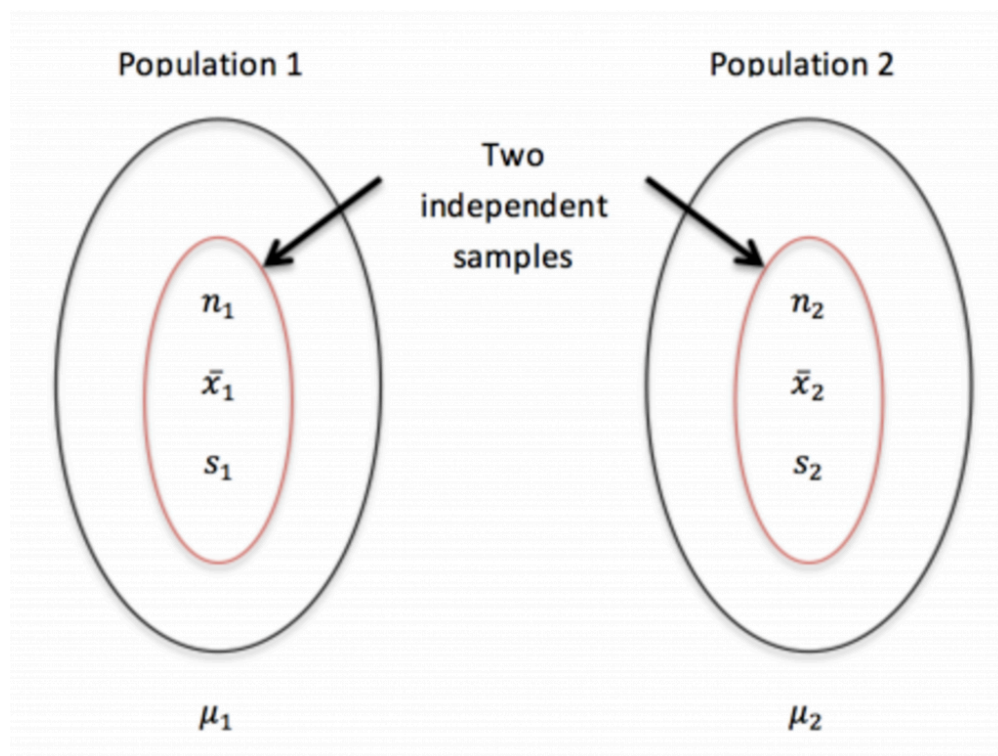


Figure 9.1: Two Independent Samples. [[Image Description \(See Appendix D Figure 9.1\)](#)]

Recall the conclusions about the sampling distribution of the sample mean \bar{X} based on samples of size n taken from a population with mean μ and standard deviation σ :

1. The mean of \bar{X} equals the population mean μ , i.e., $\mu_{\bar{X}} = \mu$.
2. The standard deviation of \bar{X} equals the population standard deviation divided by the square root of the sample size n , i.e., $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
These two conclusions are always true regardless of the population distribution and the sample size n .
3. The shape of the distribution of \bar{X} :
 - a. If the population is normally distributed, so is \bar{X} regardless of the sample size n .
 - b. If the population is not normally distributed, but the sample size n is relatively large, say $n \geq 30$, then the sample mean \bar{X} is approximately normally distributed.

A similar idea applies to the distribution of $\bar{X}_1 - \bar{X}_2$.

Key Facts: Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

1. The mean of $\bar{X}_1 - \bar{X}_2$ equals the difference of the population means: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$.
2. The standard deviation of $\bar{X}_1 - \bar{X}_2$ is: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
These two conclusions are always true regardless of the population distributions and the sample sizes n_1 and n_2 .
3. The shape of the distribution of $\bar{X}_1 - \bar{X}_2$:
 - a. If the populations are normally distributed, $\bar{X}_1 - \bar{X}_2$ is exactly normally distributed regardless of the sample sizes n_1 and n_2 .
 - b. If the populations are not normally distributed, but sample sizes n_1 and n_2 are relatively large, say $n_1 \geq 30$ and $n_2 \geq 30$, then by the central limit theorem both \bar{X}_1 and \bar{X}_2 are approximately normally distributed. The difference of two normal distributions is still normal; therefore, for $n_1 \geq 30$ and $n_2 \geq 30$, $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed.

To summarize, for normal populations **OR** large sample sizes

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

We can also standardize $\bar{X}_1 - \bar{X}_2$ to convert it into a standard normal random variable:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

If the population standard deviations σ_1 and σ_2 are unknown and estimated by sample standard deviations s_1 and s_2 , the studentized version of $\bar{X}_1 - \bar{X}_2$ is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t \text{ distribution}$$

with degrees of freedom

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2} \text{ rounded down to the nearest integer.}$$



Instructor's Note

The degrees of freedom calculation given in the above equation is very complicated, so for exams, you can use the conservative lower bound, which is defined as the smaller value of $n_1 - 1$ and $n_2 - 1$. That is, you may use $df = \min\{n_1 - 1, n_2 - 1\}$.

For example, if $n_1 = 40, n_2 = 50$, then
 $df = \min\{n_1 - 1, n_2 - 1\} = \min\{40 - 1, 50 - 1\} = \min\{39, 49\} = 39$.

9.2 Two-Sample t Test and t Interval Based on Two Independent Samples

Two-sample *t*-tests are used to test hypotheses regarding the difference between two population means. Depending on whether the two population standard deviations (σ_1 and σ_2) are equal or not, we have the non-pooled and pooled two-sample *t*-tests and *t* interval. Minor advantages of the pooled *t*-test are a slightly narrower confidence interval, a slightly more powerful test, and a simpler formula for the degrees of freedom. However, the pooled *t*-test is valid only when the two population standard deviations are close; otherwise, it gives poor results. Therefore, we recommend using the non-pooled *t*-test unless we are quite confident that $\sigma_1 = \sigma_2$, which is very difficult to verify.

9.2.1 Non-Pooled Two-Sample *t* Test and *t* Interval

Assumptions:

1. Simple random samples
2. Two samples are independent
3. Normal populations or large sample sizes ($n_1 \geq 30, n_2 \geq 30$)

Steps:

1. Set up the hypotheses:

Two-tailed test	Right (upper)-tailed test	Left (lower)-tailed test
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 \leq \Delta_0$	$H_0 : \mu_1 - \mu_2 \geq \Delta_0$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$H_a : \mu_1 - \mu_2 > \Delta_0$	$H_a : \mu_1 - \mu_2 < \Delta_0$

Note that Δ_0 can be zero or any value you want to test. In most cases, however, $\Delta_0 = 0$.

2. State the significance level α .

3. Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1 - \bar{x}_2) - (\Delta_0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ with $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$, rounded **down** to the nearest integer or $\min\{n_1 - 1, n_2 - 1\}$.

4. Use the t-score table (Table IV) to find the P-value or rejection region.

	Two-tailed	Right-tailed	Left-tailed
Null	$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 \leq \Delta_0$	$H_0 : \mu_1 - \mu_2 \geq \Delta_0$
Alternative	$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$H_a : \mu_1 - \mu_2 > \Delta_0$	$H_a : \mu_1 - \mu_2 < \Delta_0$
P-value	$2P(t \geq t_o)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_{\alpha}$	$t \leq -t_{\alpha}$

5. Decision: Reject the null H_0 if P-value $\leq \alpha$ or t_o falls in the rejection region.

6. Conclusion.

A $(1 - \alpha) \times 100\%$ two-sample t confidence interval for $\mu_1 - \mu_2$ is

Two-tailed	Right-tailed	Left-tailed
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 \leq \Delta_0$	$H_0 : \mu_1 - \mu_2 \geq \Delta_0$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$H_a : \mu_1 - \mu_2 > \Delta_0$	$H_a : \mu_1 - \mu_2 < \Delta_0$
$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) - t_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \infty$	$-\infty, (\bar{x}_1 - \bar{x}_2) + t_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Example: Two-Sample Non-Pooled t -Test and t Interval

Some students attend class regularly, but some do not. An instructor wants to compare the class averages for those who attend lectures regularly (μ_1) with those who do not (μ_2). A simple random sample of size $n_1 = 135$ is selected from the attendees and a simple random sample of size $n_2 = 35$ is taken from the non-attendees. The sample mean and sample standard deviation for attendees are $\bar{x}_1 = 67, s_1 = 17$; and for non-attendees are $\bar{x}_2 = 49, s_2 = 18$.

- a. Test at the 1% significance level whether those who attend lectures have a **higher average**, i.e., $\mu_1 > \mu_2$ or $\mu_1 - \mu_2 > 0$.

Check the assumptions:

1. We have simple random samples from attendees and non-attendees.
2. The two samples are independent.
3. We do not have the data, so we cannot check whether two populations are normally distributed using normal probability plot (Q-Q plot); however, we have large sample sizes with $n_1 = 135$, $n_2 = 35$.

Therefore, the assumptions are met.

Steps:

1. Set up the hypotheses: $H_0 : \mu_1 - \mu_2 \leq 0$ versus $H_a : \mu_1 - \mu_2 > 0$.
This is a right-tailed test.
2. The significance level is $\alpha = 0.01$.
3. Compute the value of the test statistic:

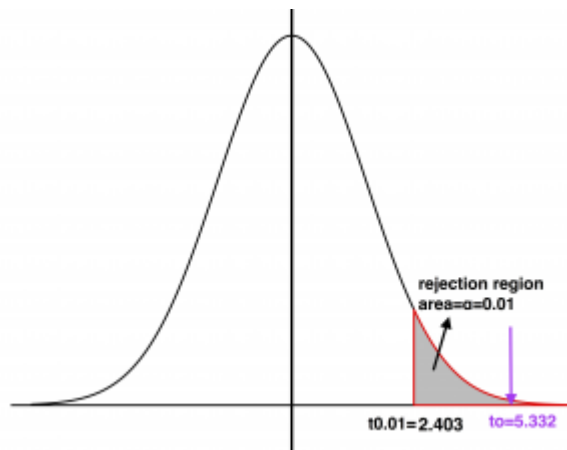
$$t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(67 - 49) - 0}{\sqrt{\frac{17^2}{135} + \frac{18^2}{35}}} = 5.332 \text{ with}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{17^2}{135} + \frac{18^2}{35}\right)^2}{\frac{1}{135 - 1} \left(\frac{17^2}{135}\right)^2 + \frac{1}{35 - 1} \left(\frac{18^2}{35}\right)^2} = 50.85, \text{ rounded down to } df = 50.$$

4. Find the P-value. For a right-tailed test with the observed test statistics $t_o = 5.332$, the P-value is the area to the right of t_o , i.e.,
P-value = $P(t \geq t_o) = P(t \geq 5.332) < 0.0005$, since $t_o = 5.332 > t_{0.0005}$
5. Decision: Since the P-value $< 0.0005 < 0.01(\alpha)$, reject the null hypothesis H_0 .
6. Conclusion: At the 1% significance level, the data provide sufficient evidence that those who attend lectures have a **higher average**.

If using the critical value approach, steps 1-3 are the same, steps 4-6 become:

4. Rejection region:



$\alpha = 0.01, t_{\alpha} = t_{0.01} = 2.403$
 For a right-tailed test, the critical value is 2.403. The rejection region is to the right of 2.403.

Figure 9.2: Rejection Region and Observed Value. [\[Image Description \(See Appendix D Figure 9.2\)\]](#)

5. Decision: Since the observed value $t_o = 5.332$ falls in the rejection region, we reject the null hypothesis H_0 .
 6. Conclusion: At the 1% significance level, the data provide sufficient evidence that those who attend lectures have a **higher average**.
- b. Obtain a confidence interval for the difference between the class average for attendees and non-attendees $\mu_1 - \mu_2$ corresponding to the test in part a).
- Part a) contains a right-tailed test at the 1% significance level. Therefore, we should obtain a 99% upper-tailed interval: $\left((\bar{x}_1 - \bar{x}_2) - t_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \infty \right)$.
- $\alpha = 0.01, df = 50, t_{\alpha} = t_{0.01} = 2.403$.
- The lower bound for the upper-tailed interval is:
- $$(\bar{x}_1 - \bar{x}_2) - t_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (67 - 49) - 2.403 \times \sqrt{\frac{17^2}{135} + \frac{18^2}{35}} = 9.887.$$
- Thus, the corresponding 99% confidence interval for $\mu_1 - \mu_2$ is $(9.887, \infty)$.
- Interpretation: we are 99% confident that the difference in average grades is at least 9.887 between attendees and non-attendees.
- c. Does the interval in part (b) support the conclusion in part a)?
 In part a), we reject H_0 at the 1% significance level and claim that $\mu_1 - \mu_2 > 0$.
 In part b), since the entire interval is above 0, we can claim that $\mu_1 - \mu_2 > 0$ with 99% confidence, which supports the results obtained in part a).
 - d. Based on the interval obtained in part b), can we claim that the class average of attendees is at least 5 marks higher than that of the non-attendees? How about 10 marks higher?
 We can claim that the class average of attendees is at least 5 marks higher than that of the non-attendees since the entire interval is above 5. However, we cannot claim that the class average of attendees is at least 10 marks higher than that of the non-attendees since the interval contains 10.

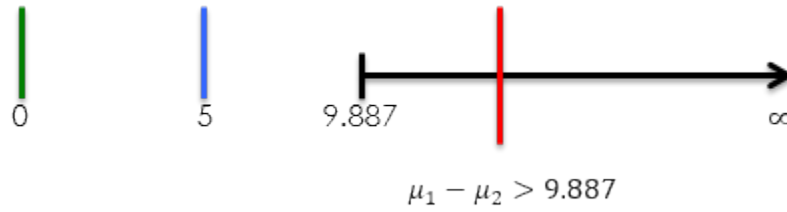


Figure 9.3: Confidence Interval of difference in Class Average. [\[Image Description \(See Appendix D Figure 9.3\)\]](#)

9.2.2 Pooled Two-Sample t Test and t Interval

If the two population standard deviations are equal, i.e., $\sigma_1 = \sigma_2 = \sigma$, we can pool the two samples together to get a better estimate of the common standard deviation σ

$$\hat{\sigma} = s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1) + (n_2-1)}}$$

where the term $(n_1 - 1)s_1^2 = \sum_{\text{sample 1}} (x - \bar{x}_1)^2$ is the variation of the data within sample 1, and $(n_2 - 1)s_2^2 = \sum_{\text{sample 2}} (x - \bar{x}_2)^2$ is the variation of the data within sample 2.

Recall that the standard deviation of $\bar{X}_1 - \bar{X}_2$ is $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Thus, if $\sigma_1 = \sigma_2 = \sigma$, then $\sigma_{\bar{X}_1 - \bar{X}_2}$ reduces to $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Estimating σ with s_p leads to the pooled test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t \text{ distribution}$$

with $df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

The assumption $\sigma_1 = \sigma_2$ is very difficult to verify. Some textbooks suggest a rule of thumb:

If the ratio of the larger to the smaller sample standard deviation is less than 2, then the assumption is considered to be reasonable, i.e., $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} < 2$.

Assumptions:

1. Simple random samples.
2. Independent samples.
3. Normal populations or large sample sizes ($n_1 \geq 30, n_2 \geq 30$).
4. Equal population standard deviations. This assumption is reasonable if $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} < 2$.

Steps:

1. Set up the hypotheses:

Two-tailed test	Right (upper)-tailed test	Left (lower)-tailed test
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 \leq \Delta_0$	$H_0 : \mu_1 - \mu_2 \geq \Delta_0$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$H_a : \mu_1 - \mu_2 > \Delta_0$	$H_a : \mu_1 - \mu_2 < \Delta_0$

Note that Δ_0 can be zero or any value you want to test.

2. State the significance level α .
3. Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, with $df = n_1 + n_2 - 2$ and $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$.
4. Use the t-score table (Table IV) to find the P-value or rejection region.

	Two-tailed	Right-tailed	Left-tailed
Null	$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 \leq \Delta_0$	$H_0 : \mu_1 - \mu_2 \geq \Delta_0$
Alternative	$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$H_a : \mu_1 - \mu_2 > \Delta_0$	$H_a : \mu_1 - \mu_2 < \Delta_0$
P-value	$2P(t \geq t_o)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_{\alpha}$	$t \leq -t_{\alpha}$

5. Decision: Reject the null H_0 if P-value $\leq \alpha$ or t_o falls in the rejection region.
6. Conclusion.

A $(1 - \alpha) \times 100$ two-sample t confidence interval for $\mu_1 - \mu_2$ is

Two-tailed	Right-tailed	Left-tailed
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 \leq \Delta_0$	$H_0 : \mu_1 - \mu_2 \geq \Delta_0$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$H_a : \mu_1 - \mu_2 > \Delta_0$	$H_a : \mu_1 - \mu_2 < \Delta_0$
$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) - t_{\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty$	$-\infty, (\bar{x}_1 - \bar{x}_2) + t_{\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Some students attend class regularly, but some do not. An instructor wants to compare the class averages for those who attend lectures regularly (μ_1) with those who do not (μ_2). A simple random sample of size $n_1 = 135$ is selected from the attendees, and a simple random sample of size $n_2 = 35$ is taken from the non-attendees. The sample mean and sample standard deviation for attendees are $\bar{x}_1 = 67, s_1 = 17$; and for non-attendees are $\bar{x}_2 = 49, s_2 = 18$.

- a. Is it reasonable to conduct a pooled two-sample t-test to test whether those who attend lectures have a **higher average**? If yes, run the test at the 1% significance level.

Check the assumptions:

1. We have simple random samples.
2. The two samples are independent.
3. We have large sample sizes ($n_1 = 135 > 30, n_2 = 35 > 30$).
4. Equal standard deviation $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} = \frac{\max\{17, 18\}}{\min\{17, 18\}} = \frac{18}{17} < 2$.

It is reasonable to conduct a pooled two-sample t-test since all the assumptions for pooled two-sample t-test are met.

Steps:

1. Set up the hypotheses: $H_0 : \mu_1 - \mu_2 \leq 0$ versus $H_a : \mu_1 - \mu_2 > 0$. This is a right-tailed test.
2. The significance level is $\alpha = 0.01$.
3. Compute the value of the test statistic:

$$t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(67 - 49) - 0}{17.207 \sqrt{\frac{1}{135} + \frac{1}{35}}} = 5.515 \text{ with } df = n_1 + n_2 - 2 = 135 + 35 - 2 = 168$$

(not given in Table IV, use $df=100$), and with

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(135 - 1)17^2 + (35 - 1)18^2}{135 + 35 - 2}} = 17.207.$$

4. Find the P-value. For a right-tailed test with the observed test statistics $t_o = 5.515$, the P-value is the area to the right of t_o i.e., $p\text{-value} = P(t \geq t_o) = P(t \geq 5.515) < 0.0005$, since $t_o = 5.515 > 3.390(t_{0.0005})$ with $df = 100$.
5. Decision: Since the P-value $< 0.0005 < 0.01(\alpha)$ reject the null hypothesis H_0 .
6. Conclusion: At the 1% significance level, the data provide sufficient evidence that those who attend lectures have a **higher average**.

- b. Obtain a confidence interval for the difference between the class average for attendees and non-attendees, $\mu_1 - \mu_2$, corresponding to the test in part a).

Part a) contains a right-tailed test at the 1% significance level. Therefore, we should obtain a 99% upper-tailed interval $((\bar{x}_1 - \bar{x}_2) - t_{\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty)$, with $\alpha = 0.01, df = 100$, and $t_{0.01} = 2.364$. The lower bound for the upper-tailed interval is

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (67 - 49) - 2.364 \times 17.207 \times \sqrt{\frac{1}{135} + \frac{1}{35}} = 10.284.$$

Thus, the corresponding 99% confidence interval for $\mu_1 - \mu_2$ is $(10.284, \infty)$.

Interpretation: we are 99% confident that the difference in average grades is at least 10.284 between attendees and non-attendees.

- c. Based on the confidence interval in part b), can we claim that the class average of attendees is at least 10 marks higher than that of the non-attendees?

Yes, since the entire interval is above 10, we can claim that $\mu_1 - \mu_2 \geq 10$.



Activity

Exercise: Two-Sample Test

The following table summarizes the operative times of neurosurgeries conducted by a dynamic system (Z-plate) and a static system (ALPS plate).

Table 9.1: Operating Time of Dynamic and Static System

Dynamic	Static
$\bar{x}_1 = 400$	$\bar{x}_2 = 480$
$s_1 = 85$	$s_2 = 40$
$n_1 = 60$	$n_2 = 30$

- Test at the 5% significance level whether the dynamic system (Z-plate) has a lower mean operative time than the static system (ALPS plate).
- Obtain a confidence interval for the difference in mean operative time between the dynamic and the static systems, $\mu_1 - \mu_2$, corresponding to the test in part a).

Show/Hide Answer

- a. **Check the assumptions:**

- We have simple random samples.
- The two samples are independent.
- We have large sample sizes $n_1 = 60 > 30, n_2 = 30 \geq 30$.
- Equal standard deviations $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} = \frac{\max\{85, 40\}}{\min\{85, 40\}} = \frac{85}{40} \approx 2.125$.

Since the equal standard deviation assumption is violated, we should use the non-pooled two-sample t -test.

Steps:

1. Set up the hypotheses: $H_0 : \mu_1 - \mu_2 \geq 0$ versus $H_a : \mu_1 - \mu_2 < 0$.

This is a left-tailed test.

2. The significance level is $\alpha = 0.05$.

3. Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(400 - 480) - 0}{\sqrt{\frac{85^2}{60} + \frac{40^2}{30}}} = -6.069$ with

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{85^2}{60} + \frac{40^2}{30}\right)^2}{\frac{1}{60 - 1} \left(\frac{85^2}{60}\right)^2 + \frac{1}{30 - 1} \left(\frac{40^2}{30}\right)^2} = 87.797, \text{ rounded down to } df = 87$$

4. Find the P-value. For a left-tailed test with the observed test statistics $t_o = -6.069$, the P-value is the area to the left of t_o , i.e.,

$$\text{P-value} = P(t \leq t_o) = P(t \leq -6.069) = P(t \geq 6.069) < 0.0005, \text{ since } 6.069 > 3.406(t_{0.0005}).$$

5. Decision: Since the P-value $< 0.0005 < 0.05(\alpha)$, reject the null hypothesis H_0 .
6. Conclusion: At the 5% significance level, the data provide sufficient evidence that the dynamic system (Z-plate) has a lower mean operative time than the static system (ALPS plate).

- b. For a left-tailed test at the 5% significance level, the corresponding confidence interval is a 95% lower-tailed interval $(-\infty, (\bar{x}_1 - \bar{x}_2) + t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$ with the upper confidence bound

$$(\bar{x}_1 - \bar{x}_2) + t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (400 - 480) + 1.663 \times \sqrt{\frac{85^2}{60} + \frac{40^2}{30}} = -58.079. \text{ Note that for } df = 87, t_{0.05} = 1.663. \text{ Therefore, the 95\% lower-tailed interval is}$$

$$(-\infty, (\bar{x}_1 - \bar{x}_2) + t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}) = (-\infty, -58.079).$$

Interpretation: we are 95% confident that the difference in mean operative time between the dynamic and the static systems is below -58.097. Since the entire interval is below 0, we can claim that $\mu_1 - \mu_2 < 0$, which supports the conclusion of the hypothesis test in part a).



Instructor's Note

It is safer to use the non-pooled two-sample t test if we are not sure whether the two population standard deviations are equal. Use the pooled two-sample t test only if we have evidence that the population standard deviations are equal. For example, we can use the pooled two-sample t test when we compare two independent groups in a one-way ANOVA (analysis of variance) analysis since equal standard deviation is one of the assumptions of the one-way ANOVA F test which will be covered in Chapter 13.

9.3 Paired t Test and Interval Based on Paired Sample

Two samples are considered **paired** if each observation in the first sample is related to exactly one observation in the second sample and each observation in the second sample is related to exactly one observation in the first sample. Some examples of paired observations include:

- Reaction times of an individual before and after consuming caffeine.
- The weight of a patient before and after a medical treatment.
- The fuel consumption of the same vehicle when it is driven at two different speeds.
- The ages of a husband and wife in the same marriage.

Example: Independent Sample or Paired Sample?

1. We randomly selected 40 males and 40 females and compared the average time they spent watching TV. Is this an independent sample or a paired sample?
An independent sample since there is no relationship between those 40 males and 40 females.
2. We randomly selected 40 couples and compared the time the husbands and wives spent watching TV. Is this an independent sample or a paired sample?
Paired sample, since those 40 males and 40 females are husbands and wives from the same households.
3. This table shows men's and women's winning times (in minutes) in the New York City Marathon between 1978 and 2006 (www.nycmarathon.org). Is this an independent sample or a paired sample?

Table 9.2: Winning Times for Men and Women of New York City Marathon 1978-2006

Year	Men	Women	Difference	Year	Men	Women	Difference
1978	132.2	152.5	20.3	1993	130.1	146.4	16.3
1979	131.7	147.6	15.9	1994	131.4	147.6	16.2
1980	129.7	145.7	16.0	1995	131	148.1	17.1
1981	128.2	145.5	17.3	1996	129.9	148.3	18.4
1982	129.5	147.2	17.7	1997	128.2	148.7	20.5
1983	129	147	18.0	1998	128.8	145.3	16.5
1984	134.9	149.5	14.6	1999	129.2	145.1	15.9
1985	131.6	148.6	17.0	2000	130.2	145.8	15.6
1986	131.1	148.1	17.0	2001	127.7	144.4	16.7
1987	131	150.3	19.3	2002	128.1	145.9	17.8
1988	128.3	148.1	19.8	2003	130.5	142.5	12.0
1989	128	145.5	17.5	2004	129.5	143.2	13.7
1990	132.7	150.8	18.1	2005	129.5	144.7	15.2
1991	129.5	147.5	18.0	2006	130	145.1	15.1
1992	129.5	144.7	15.2				

Paired sample since the winning times for men and women in the same year were compared. We should not compare the winning time for men in 2006 with the winning time for women in 2000 since the weather conditions vary from year to year, affecting the winning time.

A paired t-test and a paired t-interval are exactly a one-sample t-test and a one-sample t-interval on the **paired differences**. Therefore, the assumptions and the procedures for a paired t-test are the same as those for a one-sample t-test.

Assumptions:

1. The sample of paired differences $d_i, i = 1, \dots, n$ is a simple random sample (SRS) from the population of all possible paired differences.
2. The paired differences follow a normal distribution or a large number of paired differences ($n \geq 30$).

Steps:

1. Set up the hypotheses:

Two-tailed	Right-tailed	Left-tailed
$H_0 : \mu_1 - \mu_2 = \delta_0$	$H_0 : \mu_1 - \mu_2 \leq \delta_0$	$H_0 : \mu_1 - \mu_2 \geq \delta_0$
$H_a : \mu_1 - \mu_2 \neq \delta_0$	$H_a : \mu_1 - \mu_2 > \delta_0$	$H_a : \mu_1 - \mu_2 < \delta_0$

Note: δ_0 can be any value tested, but in most cases $\delta_0 = 0$. Some textbooks state the hypotheses using $\mu_d = \mu_1 - \mu_2$.

- State the significance level α .
- Compute the value of the test statistic: $t_o = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}}$, with degrees of freedom $df = n - 1$, where n is the number of paired differences and the mean and standard deviation of the paired differences are given by

$$\bar{d} = \frac{\sum d_i}{n}, s_d = \sqrt{\frac{(\sum d_i^2) - \frac{(\sum d_i)^2}{n}}{n-1}}.$$

- Use the t-score table (Table IV) to find the P-value or rejection region.

	Two-tailed	Right-tailed	Left-tailed
Null	$H_0 : \mu_1 - \mu_2 = \delta_0$	$H_0 : \mu_1 - \mu_2 \leq \delta_0$	$H_0 : \mu_1 - \mu_2 \geq \delta_0$
Alternative	$H_a : \mu_1 - \mu_2 \neq \delta_0$	$H_a : \mu_1 - \mu_2 > \delta_0$	$H_a : \mu_1 - \mu_2 < \delta_0$
P-value	$2P(t \geq t_o)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_{\alpha}$	$t \leq -t_{\alpha}$

- Reject the null H_0 if P-value $\leq \alpha$ or t_o falls in the rejection region.
- Conclusion.

A $(1 - \alpha) \times 100\%$ confidence interval for $\mu_d = \mu_1 - \mu_2$ corresponding to a hypothesis test at the significance level α is

	Two-tailed	Right-tailed	Left-tailed
Null	$H_0 : \mu_1 - \mu_2 = \delta_0$	$H_0 : \mu_1 - \mu_2 \leq \delta_0$	$H_0 : \mu_1 - \mu_2 \geq \delta_0$
Alternative	$H_a : \mu_1 - \mu_2 \neq \delta_0$	$H_a : \mu_1 - \mu_2 > \delta_0$	$H_a : \mu_1 - \mu_2 < \delta_0$
$(1 - \alpha) \times 100\%$ CI	$(\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}})$	$(\bar{d} - t_{\alpha} \frac{s_d}{\sqrt{n}}, \infty)$	$(-\infty, \bar{d} + t_{\alpha} \frac{s_d}{\sqrt{n}})$
Decision	Reject H_0 if δ_0 is outside the interval		

Example: Paired t-test and Paired t-interval

This table shows men and women's winning times (in minutes) in the New York City Marathon between 1978 and 2006.

Year	Men	Women	Difference	Year	Men	Women	Difference
1978	132.2	152.5	20.3	1993	130.1	146.4	16.3
1979	131.7	147.6	15.9	1994	131.4	147.6	16.2
1980	129.7	145.7	16.0	1995	131	148.1	17.1
1981	128.2	145.5	17.3	1996	129.9	148.3	18.4
1982	129.5	147.2	17.7	1997	128.2	148.7	20.5
1983	129	147	18.0	1998	128.8	145.3	16.5
1984	134.9	149.5	14.6	1999	129.2	145.1	15.9
1985	131.6	148.6	17.0	2000	130.2	145.8	15.6
1986	131.1	148.1	17.0	2001	127.7	144.4	16.7
1987	131	150.3	19.3	2002	128.1	145.9	17.8
1988	128.3	148.1	19.8	2003	130.5	142.5	12.0
1989	128	145.5	17.5	2004	129.5	143.2	13.7
1990	132.7	150.8	18.1	2005	129.5	144.7	15.2
1991	129.5	147.5	18.0	2006	130	145.1	15.1
1992	129.5	144.7	15.2				

- a. At the 1% significance level, do the data provide sufficient evidence that, there is a difference in mean winning times between males and females? Note that the sample mean and standard deviation of the paired differences are $\bar{d} = 16.85$ and $s_d = 1.98$ respectively.

Steps:

1. Set up the hypotheses: $H_0 : \mu_F - \mu_M = 0$ versus $H_a : \mu_F - \mu_M \neq 0$.
2. The significance level is $\alpha = 0.01$.
3. Compute the value of the test statistic: $t_o = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}} = \frac{16.85 - 0}{1.98 / \sqrt{29}} = 45.828$ with $df = n - 1 = 29 - 1 = 28$.
4. Find the P-value. For a two-tailed test, the P-value is twice the area to the right of the absolute value of the observed test statistic t_o .
P-value = $2P(t \geq |t_o|) = 2P(t \geq 45.828) < 2 \times 0.0005 = 0.001$, since $45.828 > 3.674(t_{0.0005})$
5. Decision: Since the P-value $< 0.001 < 0.01(\alpha)$, we reject the null hypothesis H_0 .
6. Conclusion: At the 1% significance level, the data provide sufficient evidence that there is a difference in mean winning times between males and females.

- b. Obtain a 99% two-tailed interval for the difference in mean winning times between males and females, i.e., $\mu_F - \mu_M$.

Since $1 - \alpha = 0.99 \implies \alpha = 0.01$, use Table IV with $df = 28$, $t_{\alpha/2} = t_{0.005} = 2.763$. Therefore, a 99% two-tailed interval for $\mu_F - \mu_M$ is given by

$$\begin{aligned}
& (\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}}) \\
& = (16.85 - 2.763 \times \frac{1.98}{\sqrt{29}}, 16.85 + 2.763 \times \frac{1.98}{\sqrt{29}}) \\
& = (15.834, 17.866).
\end{aligned}$$

Interpretation: we are 99% confident that the mean difference in winning time between females and males is somewhere between 15.834 and 17.866 minutes, i.e., the mean winning time of females is 15.834 to 17.866 minutes longer than the mean winning time of males.

- c. Does the interval in part (b) support the conclusion in part (a)?
In part (a), we reject H_0 and claim that $\mu_F - \mu_M \neq 0$, with 1% significance. In part (b), the 99% confidence interval does not contain $\delta_0 = 0$, and so we can claim that $\mu_F - \mu_M \neq 0$ with 99% confidence. Therefore, the results from part b) support the results obtained in part (a).
- d. Based on the confidence interval in part (b), what is the conclusion of testing $H_0 : \mu_F - \mu_M = 16$ versus $H_a : \mu_F - \mu_M \neq 16$ at the 1% significance level?
Since the hypothesized value $\delta_0 = 16$ is inside the 99% confidence interval (15.834, 17.866), we cannot reject the null hypothesis $H_0 : \mu_F - \mu_M = 16$ at the 1% significance level.



Activity

Exercise: Paired t-Test and Paired t Interval

Eleven people participate in a diet program; their weights in pounds before and after taking the program are listed below.

Table 9.3: Working Table for Weight Lose

Before	After	Paired Differences $d_i = \text{Before} - \text{After}$	d_i^2
130	100	30	900
140	115	25	625
160	140	20	400
110	115	-5	25
120	120	0	0
150	130	20	400
160	130	30	900
100	110	-10	100
180	140	40	1600
200	150	50	2500
130	120	10	100
Sum		$\sum d_i = 210$	$\sum d_i^2 = 7550$

Normal Probability Plot on Paired Differences

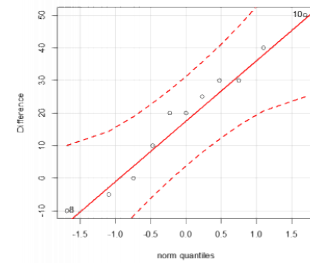


Figure 9.4: Normal Probability Plot for Paired Difference.
[\[Image Description \(See Appendix D Figure 9.4\)\]](#) Click on the image to enlarge it.

- Test at the 1% significance level whether the diet program is effective in reducing weight.
- Obtain a confidence interval corresponding to the test in part a).
- Does the interval in part b) support the conclusion in part a)?
- Is it possible to claim that the diet program can reduce weight by more than 5 pounds on average? Explain why.

Show/Hide Answer

Answer

a. **Check the assumptions:**

- We have a simple random sample of paired differences.
- We have only $n = 11$ pairs, which is too small for the CLT to apply. Therefore, we should draw a Q-Q plot of the **paired differences** to see whether they are from a normal population. Since all the points are roughly on a straight line, there is no strong evidence against the normality assumption.

Let μ_B and μ_A be the mean weight before and after the diet program, respectively. If the diet program is effective in reducing weight, the average weight before the program should be larger than the average weight after the program.

Steps:

1. Set up the hypotheses: $H_0 : \mu_B - \mu_A \leq 0$ versus $H_a : \mu_B - \mu_A > 0$.
2. The significance level is $\alpha = 0.01$.
3. Compute the value of the test statistic:

$$t_o = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}} = \frac{19.091 - 0}{18.817 / \sqrt{11}} = 3.365 \text{ with } df = n - 1 = 11 - 1 = 10 \text{ and } \bar{d} = \frac{\sum d_i}{n} = \frac{210}{11} = 19.091,$$

$$s_d = \sqrt{\frac{(\sum d_i^2) - \frac{(\sum d_i)^2}{n}}{n-1}} = \sqrt{\frac{7550 - \frac{(210)^2}{11}}{11-1}} = 18.817.$$
4. Find the P-value. For a right-tailed test, the P-value is the area to the right of the observed test statistic t_o , i.e.,

$$\text{P-value} = P(t \geq t_o) = P(t \geq 3.365) \implies 0.0025 < \text{P-value} < 0.005. \text{ Note that with } df = 10, 3.169(t_{0.005}) < 3.365 < 3.581(t_{0.0025}).$$
5. Decision: Since the P-value $< 0.005 < 0.01(\alpha)$, we reject the null hypothesis H_0 .
6. Conclusion: At the 1% significance level, the data provide sufficient evidence that the diet program is effective in reducing average weight.

- For a right-tailed test at the 1% significance level, the corresponding confidence interval is a 99% upper-tailed interval $(\bar{d} - t_{\alpha} \frac{s_d}{\sqrt{n}}, \infty)$ with $df = n - 1 = 10$,

$$t_{0.01} = 2.764, \bar{d} - t_{\alpha} \frac{s_d}{\sqrt{n}} = 19.091 - 2.764 \times \frac{18.817}{\sqrt{11}} = 3.409. \text{ The 99\% upper-tailed interval is } (3.409, \infty).$$

Interpretation: We are 99% confident that the diet program reduces weight by at least 3.409 pounds on average.

- Does the interval in part b) support the conclusion in part a)?
 Yes. In part a), we reject H_0 and claim that $\mu_B - \mu_A > 0$. In part b), since the interval does not contain $\delta_0 = 0$ and the entire interval is above 0, we are 99% confident that $\mu_B - \mu_A > 0$. Thus, the results from part b) support the results obtained in part a).
- Is it possible to claim that, on average, the diet program reduces weight by more than 5 pounds? Explain why.
 This question asks us to test $H_0 : \mu_B - \mu_A \leq 5$ versus $H_a : \mu_B - \mu_A > 5$. Since the hypothesized difference $\delta_0 = 5$ is within the interval $(3.409, \infty)$, we cannot reject 5 (or any value as low as 3.409) as a possible value for $\mu_B - \mu_A$. Therefore, we cannot reject $H_0 : \mu_B - \mu_A \leq 5$.

9.4 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the sampling distribution of the difference between two sample means $\bar{X}_1 - \bar{X}_2$ for two independent samples (Section 9.1).
- State the assumptions for inferences about the difference between two population means based on two independent samples (Section 9.2).
- Obtain and interpret a $(1 - \alpha) \times 100\%$ two-mean confidence interval for $\mu_1 - \mu_2$ (Section 9.2).
- Conduct a two-mean (sample) *t*-test (Section 9.2).
- Determine whether two samples are independent or paired (Section 9.3).
- Obtain and interpret a paired *t*-confidence interval (Section 9.3).
- Conduct a paired *t*-test (Section 9.3).
- Explain the relationship between the results of a hypothesis test at significance level α and the corresponding $(1 - \alpha) \times 100\%$ confidence interval (Section 9.3).

9.5 Review Questions

1. Determine whether we should use a two-sample t test or a paired t test in the following applications.
 - a. The data in the following table give the number of young per litter for 24 female cottonmouths in Florida and 44 female cottonmouths in Virginia.

Florida			Virginia					
8	6	7	5	12	7	7	6	8
7	4	3	12	9	7	4	9	6
⋮		⋮	⋮		⋮			⋮
5	5	4	5	4				

At the 1% significance level, do the data provide sufficient evidence to conclude that, on average, the number of young per litter of cottonmouths in Florida is less than that in Virginia?

- b. Independent random samples of 10 homes each in Atlantic City and Las Vegas yielded the following data on home prices in thousands of dollars. At the 5% significance level, can you conclude that the mean costs for existing single-family homes differ in Atlantic City and Las Vegas?

Atlantic City		Las Vegas	
234.0	192.8	226.4	231.5
213.0	256.4	214.7	210.9
⋮	⋮	⋮	⋮
236.1	301.9	349.4	178.5

- c. The following table gives data, for a sample of eight years, on the number of days that ice stayed on two lakes in Madison, Wisconsin-Lake Mendota and Lake Monona. At the 10% significance level, do the data provide sufficient evidence to conclude that a difference exists in the mean length of time that ice stays on these two lakes?

Year	Mendota	Monona
1	119	107
2	115	108
⋮	⋮	⋮
8	87	91

- d. The fiber density of 10 samples with varying fiber density was obtained using both an eye-piece method and a TV-screen method. The results, in fibers per square millimeter, are presented in the following table. Test at the 5% significance level whether, on average, the eyepiece method gives a greater fiber-density reading than the TV-screen method.

Sample ID	Eyepiece	TV Screen
1	182.2	177.8
2	118.5	116.6
⋮	⋮	⋮
10	85.4	86.6

- e. Compare the average IQ score of students from U of A and MacEwan. Randomly pick 30 students from each University and obtain their IQ scores.
- f. Compare weekly earnings of male and female workers. Randomly pick 40 male and 40 female workers and compare their average weekly earnings.
- g. Compare weekly earnings of male and female workers. Randomly pick 40 households and compare the husbands' and wives' average weekly earnings.
2. The following table summarizes the operative times of neurosurgeries conducted by a dynamic system (Z-plate) and a static system (ALPS plate), respectively.

Dynamic	Static
$\bar{x}_1 = 400$	$\bar{x}_2 = 480$
$s_1 = 85$	$s_2 = 40$
$n_1 = 60$	$n_2 = 30$

- a. Test at a 5% significance level whether the dynamic system (Z-plate) reduced the mean operative time relative to the static system (ALPS plate).
- b. Obtain a confidence interval for the difference in mean operative time between the dynamic and the static systems, $\mu_1 - \mu_2$, corresponding to the test in part (a).
3. This table shows men and women's winning times (in minutes) in the New York City

Marathon between 1978 and 2006.

Men(y_i)	Women(x_i)	Difference(d_i)	Men(y_i)	Women(x_i)	Difference(d_i)
132.2	152.5	20.3	130.1	146.4	16.3
131.7	147.6	15.9	131.4	147.6	16.2
129.7	145.7	16.0	131.0	148.1	17.1
128.2	145.5	17.3	129.9	148.3	18.4
129.5	147.2	17.7	128.2	148.7	20.5
129.0	147.0	18.0	128.8	145.3	16.5
134.9	149.5	14.6	129.2	145.1	15.9
131.6	148.6	17.0	130.2	145.8	15.6
131.1	148.1	17.0	127.7	144.4	16.7
131.0	150.3	19.3	128.1	145.9	17.8
128.3	148.1	19.8	130.5	142.5	12.0
128.0	145.5	17.5	129.5	143.2	13.7
132.7	150.8	18.1	129.5	144.7	15.2
129.5	147.5	18.0	130.0	145.1	15.1
129.5	144.7	15.2	$\bar{d} = 16.85, s_d = 1.98$		

- At the 1% significance level, do the data provide sufficient evidence that there is a difference in winning times between males and females?
- Obtain a confidence interval corresponding to the test in part (a).
- Does the interval in part (b) support the conclusion in part (a)?
- Based on the interval obtained in part (b), can we claim that the winning time of males is at least 15 minutes fast than that of females? How about 20 minutes fast?

Show/Hide Answer

1.

- The data are two independent samples, use two-sample t test
- The data are two independent samples, use two-sample t test
- The data is a paired sample, use paired t test.
- The data is a paired sample, use paired t test.
- Two independent samples, use two-sample t test.
- Two independent samples, use two-sample t test.
- Paired sample, use paired

2.

a. Check the assumptions:

- We have simple random samples.
- The two samples are independent.
- We have large samples $n_1 = 60 > 30$ and $n_2 = 30 > 30$.

Steps:

1. Hypotheses. $H_0 : \mu_1 - \mu_2 \geq 0$ versus $H_a : \mu_1 - \mu_2 < 0$.

2. Significance level $\alpha = 0.05$.

$$t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(400 - 480) - 0}{\sqrt{\frac{85^2}{60} + \frac{40^2}{30}}} = -6.069, df = \min\{n_1 - 1, n_2 - 1\} =$$

3. Test statistic. $\min\{60 - 1, 30 - 1\} = 29$.

4. P -value: for a left-tailed test, P -value =

$$P(t \leq t_o) = P(t \leq -6.069) = P(t \geq 6.069) < 0.0005 < 0.05(\alpha).$$

5. Decision: Reject H_0 , since P -value $< \alpha$.

6. Conclusion: At the 5% significance level, we have sufficient evidence that the dynamic system (Z-plate) reduced the mean operative time relative to the static system (ALPS plate).

b. We should construct a lower-tailed 95% confidence interval for a left-tailed test at the 5% significance level. The upper confidence bound for a 95% lower-tailed interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) + t_\alpha \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (400 - 480) + 1.699 \times \sqrt{\frac{85^2}{60} + \frac{40^2}{30}} = -57.605.$$

The 95% lower-tailed confidence interval is $(-\infty, -57.605)$.

Interpretation: we can be 95% confident that the mean operative time of the dynamic system is at least 57.605 minutes shorter than the mean operative time of the static system.

3.

a. Check the assumptions:

- We have simple random samples.
- We have large number of pairs $n = 29 \approx 30$.

Steps:

1. Hypotheses. $H_0 : \mu_M - \mu_F = 0$ versus $H_a : \mu_M - \mu_F \neq 0$.

2. Significance level $\alpha = 0.01$.

3. Test statistic. $t_o = \frac{\bar{d} - \delta_0}{\frac{s_d}{\sqrt{n}}} = \frac{16.85 - 0}{\frac{1.98}{\sqrt{29}}} = 45.828$, with $df = n - 1 = 29 - 1 = 28$.

4. P -value: for a two-tailed test, P -value =

$$2P(t \leq |t_o|) = 2P(t \geq 45.828) < 2 \times 0.0005 = 0.001 < 0.01(\alpha).$$

5. Decision: Reject H_0 , since $P\text{-value} < \alpha$.

6. Conclusion: At the 1% significance level, we have sufficient evidence that there is a difference in winning times between males and females.

- b. A two-tailed test at the 1% significance level corresponds to a 99% two-tailed confidence interval. For $df = 28$,

$$1 - \alpha = 0.99 \implies \alpha = 0.01 \implies \frac{\alpha}{2} = \frac{0.01}{2} = 0.005 \implies t_{\alpha/2} = t_{0.005} = 2.763. \quad \text{A 99\% two-tailed}$$

$$\text{confidence interval for } \mu_d \text{ is } \bar{d} \pm t_{\alpha/2} \times \frac{s_d}{\sqrt{n}} = 16.85 \pm 2.763 \times \frac{1.98}{\sqrt{29}} = (15.834, 17.866).$$

Interpretation: we can be 99% confident that the difference in the mean winning times between male and female is somewhere between 15.834 minutes and 17.866 minutes.

- c. Yes, because the hypothesized value $\delta_0 = 0$ is outside the entire interval; therefore, we can claim that $\mu_M - \mu_F \neq 0$ which is the conclusion of the paired t test. Since the entire interval is above 0, we can further claim that $\mu_M - \mu_F > 0$.
- d. We can claim that the mean winning time of males is at least 15 minutes faster than that of females, since the entire interval is above 15, we can claim $\mu_M - \mu_F > 15$. However, we cannot claim that the mean winning time of males is at least 20 minutes faster than that of females since the 20 is inside the interval.

Note: for a right-tailed test at the 1% significance level, it is more precise to construct a 99% upper-tailed interval with a lower bound:

$$\bar{d} \pm t_{\alpha} \times \frac{s_d}{\sqrt{n}} = 16.85 - 2.467 \times \frac{1.98}{\sqrt{29}} = 15.943. \text{ The 99\% upper-tailed interval is } (-\infty, 15.943).$$

Interpretation: we can be 99% confident that the winning time of males is at least 15.943 minutes faster than that of females. Since $\delta_0 = 15$ is outside the confidence interval, we can claim that $\mu_M - \mu_F > 15$; however, $\delta_0 = 20$ is inside the confidence interval, we cannot claim that $\mu_M - \mu_F > 20$. There is insufficient evidence that the mean winning time of males is at least 20 minutes faster than that of females.

9.6 Assignment 9

Purposes

This assignment has two parts. The first part assesses your knowledge of properties of the distribution of the difference between two sample means $\bar{X}_1 - \bar{X}_2$, conducting a two-sample t test, and performing a paired t test. The second part assesses your skills in using R commander to conduct a two-sample and a paired t test in comparing two population means μ_1 and μ_2 .

Resources

[M09_Gas_Q7.xlsx](#)

[M09_Direction_Q4.xlsx](#)

[M09_Driver_Q5_TwoColumn.xlsx](#)

[M09_Driver_Q5.xlsx](#)

[M09_Treadwear_Q6.xlsx](#)

Instructions

Part A

Complete the following:

1. Consider the quantities μ_1 , \bar{X}_1 , \bar{x}_1 , s_1 , μ_2 , X_2 , \bar{x}_2 , and s_2 .
 - a. Which quantities represent parameters and which represent statistics? (4 marks)
 - b. Which quantities are fixed numbers and which are variables? (4 marks)
2. A variable of two populations has a mean of 8 and a standard deviation of 6 for one of the populations and a mean of 7 and a standard deviation of 5 for the other population.
 - a. For independent samples of sizes 3 and 6, respectively, find the mean and

- standard deviation of $X_1 - X_2$. (3 marks)
- Must the variable under consideration be normally distributed on each of the two populations for you to answer part (a)? Explain your answer. (2 marks)
 - Can you conclude that the variable $X_1 - X_2$ is normally distributed? Explain your answer. (3 marks)
- A variable of two populations has a mean of 40 and a standard deviation of 12 for one of the populations and a mean of 40 and a standard deviation of 6 for the other population. Moreover, the variable is normally distributed on each of the two populations.
 - For independent samples of sizes 9 and 4, respectively, determine the mean and standard deviation of $X_1 - X_2$. (3 marks)
 - Can you conclude that the variable $\bar{X}_1 - \bar{X}_2$ is normally distributed? Explain your answer. (3 marks)
 - Determine the percentage of all pairs of independent samples of sizes 9 and 4, respectively, from the two populations with the property that the difference $\bar{X}_1 - \bar{X}_2$ between the sample means is between -10 and 10. (5 marks)
 - A study examined the sense of direction of 30 male and 30 female students. After being taken to an unfamiliar wooded park, the students were given some spatial orientation tests, including pointing to the south, which tested their absolute frame of reference. The students pointed by moving a pointer attached to a 360° protractor. Following are the absolute pointing errors, in degrees, of the participants.

Male					Female				
13	130	39	33	10	14	8	20	3	138
13	68	18	3	11	122	78	69	111	3
38	23	60	5	9	128	31	18	35	111
59	5	86	22	70	109	36	27	32	35
58	3	167	15	30	12	27	8	3	80
8	20	67	26	19	91	68	66	176	15

	Male	Female
Mean	$\bar{x}_1 = 37.6$	$\bar{x}_2 = 55.8$
SD	$s_1 = 38.49$	$s_2 = 48.26$

- Given the statistical summaries, test at the 1% significance level whether males have a better sense of direction than females on average. (8 marks)
- What is the P-value of the hypothesis test in part (a)? (2 marks)
- Obtain a confidence interval for the difference between the mean absolute pointing errors for males and females corresponding to the hypothesis test in part (a). (4 marks)
- Interpret the confidence interval obtained in part (c). Does this interval support

the conclusion of the hypothesis test in part (a)? Justify your answer. (4 marks: 2+2)

5. Frustrated passengers, congested streets, time schedules, and air and noise pollution are just some of the physical and social pressures that lead many urban bus drivers to retire prematurely with disabilities such as coronary heart disease and stomach disorders. An intervention program was implemented to improve the work conditions of the city's bus drivers. The following table reported the heart rates, in beats per minute, of the drivers who drove on the improved routes (intervention) and those who drove on the regular routes (control).

Intervention (μ_1)	Control (μ_2)
68 66 72 62 69 63 68 71 64 76	74 52 67 63 77 57 80 77 58 72 54 73 54 55 82 63 60 68 64 66 75 72 55 71 84 63 79 59 74 58 82
$\bar{x}_1 = 67.90, s_1 = 4.36, n_1 = 10$	$\bar{x}_2 = 67.35, s_2 = 9.66, n_2 = 31$

- Is applying the pooled two-sample t test reasonable? Justify your answer. (2 marks)
 - At the 5% significance level, do the data provide sufficient evidence that the intervention program reduces the mean heart rate of urban bus drivers? (8 marks)
 - Obtain a confidence interval for the difference between the mean heart rates of urban bus drivers in the two environments corresponding to the hypothesis test in part (a). (4 marks)
 - Interpret the confidence interval obtained in part (c). Does this interval support the conclusion of the hypothesis test in part (a)? Justify your answer. (4 marks: 2+2)
6. Eleven tires were each measured for treadwear by two methods, one based on weight and the other on groove wear. The following are the data, in thousands of miles.

Weight method (μ_1)	Groove method (μ_2)	Difference ($\mu_1 - \mu_2$)
30.5	28.7	1.8
30.9	25.9	5
31.9	23.3	8.6
30.4	23.1	7.3
27.3	23.7	3.6
20.4	20.9	-0.5
24.5	16.1	8.4
20.9	19.9	1
18.9	15.2	3.7
13.7	11.5	2.2
11.4	11.2	0.2
$\bar{x}_1 = 23.71, s_1 = 7.19$	$\bar{x}_2 = 19.95, s_2 = 5.77$	$d = 3.75, s_d = 3.22$

- Are the data two independent samples or a simple paired sample? (2 marks)
 - At the 5% significance level, do the data provide sufficient evidence to conclude that, on average, the two measurement methods give different results? (8 marks)
 - What is the P-value of the hypothesis test in part (b)? (2 marks)
 - Obtain a confidence interval for the mean difference in measurement by the weight and groove methods corresponding to the hypothesis test in part (b). (4 marks)
 - Interpret the confidence interval obtained in part (d). Does this interval support the conclusion of the hypothesis test in part (b)? Justify your answer. (4 marks: 2+2)
7. The manufacturer of an Engine Energizer System (EES) claims that it improves gas mileage and reduces emissions in automobiles by using magnetic-free energy to increase the amount of oxygen in the fuel for greater combustion efficiency. Following are test results, performed under international and U.S. government agency standards, on a random sample of 14 vehicles. The data (also see file **M09_Gas_Paired.txt** or **M09_Gas_Q7.xlsx**) give the carbon monoxide (CO) levels, in parts per million, of each vehicle tested, both before installation of EES and after installation.

Before	After
1.60	0.15
0.30	0.20
3.80	2.80
6.20	3.60
3.60	1.00
1.50	0.50
2.00	1.60
2.60	1.60
0.15	0.06
0.06	0.16
0.60	0.35
0.03	0.01
0.10	0.00
0.19	0.00

- Test at the 1% significance level whether, on average, EES reduces CO emissions. (11 marks)
- Obtain a 99% confidence interval for the difference between the mean CO emissions before and after installation of EES corresponding to the test in part (a). (4 marks)
- Interpret the confidence interval obtained in part (b). Does this interval support the conclusion of the hypothesis test in part (a)? Justify your answer. (4 marks: 2+2)

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs and write down your answers in statements.

- Refer to Question 4 in Part A. The data are provided in the data file **M09_Direction_Q4.xlsx**. Import the data into R commander.
 - Use the proper graphical tools in R and R commander to assess whether it is reasonable to apply the procedure you chose in Question 4 of Part A. Make sure to write down the assumptions of the procedure and address whether every assumption is satisfied. (5 marks)
 - Re-conduct the test in Question 4 (a) using R commander. Make sure to include all the six components of a hypothesis test. **Copy and paste the computer output first** and then compare the answer on the output with the one you obtained by hand in Question 4 (a) and (b). (5 marks)
Note: R commander uses Female-Male by default; please pay attention to this when you set up the hypotheses.
 - Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare it to the one you obtained by hand in Question 4 (c). (2 marks)
- Refer to Question 5 in Part A. The data are provided in the data files **M09_Driver_Q5_twocolumn.xlsx** and **M09_Driver_Q5.xlsx**.

- a. Use the proper graphical tools in R and R commander to assess whether applying the procedure you chose in Question 5 (b) is reasonable. Make sure to write down the assumptions of the procedure and address whether every assumption is satisfied. (Hint: use the data file **M09_Driver_Q5_twocolumn.xlsx** to draw the normal probability plots). (5 marks)
 - b. Import the data file **M09_Driver_Q5.xlsx** into R commander. Re-conduct the test in Question 5 (b) using R commander. Make sure to include all the six components of a hypothesis test. **Copy and paste the computer output first**, then compare the answer with the one you obtained by hand in Question 5 (b). (5 marks)
 - c. Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare it to the one you obtained by hand in Question 5 (c). (2 marks)
3. Refer to Question 6 in Part A. The data are provided in the data file **M09_Treadwear_Q6.xlsx**.
- a. Use the proper graphical tools in R and R commander to assess whether applying the procedure you chose in Question 6 (b) is reasonable. Make sure to write down the assumptions of the procedure and address whether every assumption is satisfied. (5 marks)
 - b. Re-conduct the test in Question 6 (b) using R commander. Make sure to include all the six components of a hypothesis test. **Copy and paste the computer output first**, then compare the answer with the one you obtained by hand in Question 6 (b) and (c). (5 marks)
 - c. Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare the one you obtained by hand in Question 6 (d). (2 marks)
4. Refer to Question 6. The data are provided in the data file **M09_Gas_Q7.xlsx**.
- a. Use the proper graphical tools in R and R commander to assess whether applying the procedure you chose in Question 7 (a) is reasonable. Make sure to write down the assumptions of the procedure and address whether every assumption is satisfied. (5 marks)
 - b. Re-conduct the test in Question 7 (a) using R commander. Make sure to include all the six components of a hypothesis test. **Copy and paste the computer output first** and then Compare the answer with the one you obtained by hand in Question 7 (a). (5 marks)
 - c. Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare it to the one you obtained by hand in Question 7 (b). (2 marks)

Quiz 9



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2639#h5p-11>

CHAPTER 10: INFERENCES FOR POPULATION PROPORTIONS

Overview

Chapters 8 and 9 introduced inferences for population means. This chapter focuses on inferences for another population parameter: **the population proportion p** , defined as the proportion (or percentage) of a population with a specified attribute. For example, the proportion of times that athletes wearing blue uniforms win a judo match, the proportion of customers who respond to an advertisement, and the proportion of women who have arthritis.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain why the sample proportion $\hat{p} = \frac{x}{n}$ is a special type of sample mean $\bar{x} = \frac{\sum x_i}{n}$.
- Describe the sampling distribution of the sample proportion \hat{p} .
- Conduct a one-proportion z-test.
- Obtain a $(1 - \alpha) \times 100$ confidence interval for the population proportion p .
- Describe the sampling distribution of the difference between two sample proportions $(\hat{p}_1 - \hat{p}_2)$.
- Conduct a two-proportion z-test.
- Obtain a $(1 - \alpha) \times 100$ confidence interval for the difference between two population proportions $(p_1 - p_2)$.

10.1 Population Proportion and the Sample Proportion

Recall that the population mean $\mu = \frac{\sum x_i}{N}$ is a population parameter used to describe the population, where N is the population size (number of individuals in the population). The population proportion

$$p = \frac{\text{of individuals having a certain attribute}}{\text{population size}} = \frac{\text{of successes}}{N}$$

is another parameter used to describe the population. For example, the proportion of female students at MacEwan is defined as

$$p = \frac{\text{of female students at MacEwan}}{\text{total number of students at MacEwan}} = \frac{\text{of successes}}{N}.$$

In this instance, picking a female student is regarded as a success.

Just as the sample mean $\bar{x} = \frac{\sum x_i}{n}$ is used to estimate the population mean μ , the sample proportion \hat{p} is used to estimate the population proportion p , where

$$\hat{p} = \frac{\text{of individuals having a certain attribute in the sample}}{\text{sample size}} = \frac{\text{of successes in the sample}}{n}.$$

Here are several examples:

Examples

- A random sample of $n = 100$ students is obtained from MacEwan University. Of the 100 students in the sample, 65 are female. The sample proportion $\hat{p} = \frac{x}{n} = \frac{65}{100}$ provides a point estimate of p , the proportion of female students at MacEwan.
- A random sample of $n = 1000$ judo matches is obtained, and it is determined that 510 of the matches are won by the athletes wearing a blue uniform. The sample proportion $\hat{p} = \frac{x}{n} = \frac{510}{1000}$ is a point estimate of p , the proportion of winners in blue.
- A credit card company sends an advertisement to $n = 500$ randomly chosen customers and only 10 customers respond. The sample proportion $\hat{p} = \frac{x}{n} = \frac{10}{500}$ is a point estimate of p , the proportion of respondents.

10.2 Distribution of the Sample Proportion

Inferences about the population mean μ are based on the distribution of the sample mean \bar{X} . Similarly, inferences about the population proportion p are based on the distribution of the sample proportion \hat{p} .

The **population proportion** is defined as

$$p = \frac{\text{of individuals having a certain attribute}}{\text{of individuals in the population}} = \frac{\text{of successes}}{N}.$$

The population proportion can be regarded as a special type of population mean if we let the variable of interest be an indicator variable as follows:

$$x_i = \begin{cases} 1 & \text{if the } i\text{th individual has the attribute (a success),} \\ 0 & \text{if the } i\text{th individual does not have the attribute (a failure).} \end{cases}$$

Then, the population proportion can be rewritten as

$$p = \frac{\text{of individuals having a certain attribute}}{\text{of individuals in the population}} = \frac{\text{of successes}}{N} = \frac{\sum x_i}{N}.$$

The variable of interest X has only two possible values: 1 if the individual has the attribute and 0 if not. Randomly select one individual and define p as the probability that this individual has the attribute. As a result, the probability distribution of X is

Table 10.1: Probability Distribution of an Indicator Variable

x	1	0
$P(X = x)$	p	$1 - p$

with a population mean and population standard deviation:

$$\mu = \sum xP(X = x) = 1 \times p + 0 \times (1 - p) = p,$$

$$\sigma = \sqrt{\sum x^2P(X = x) - \mu^2} = \sqrt{1^2 \times p + 0^2 \times (1 - p) - p^2} = \sqrt{p - p^2} = \sqrt{p(1 - p)}.$$

The sample proportion can be viewed as a special type of sample mean (in the same way that the population proportion can be viewed as a special type of population mean). That is,

in a simple random sample of size n , the proportion of individuals with the specific attribute is the sample proportion:

$$\begin{aligned}\hat{p} &= \frac{\text{of individuals having a certain attribute in the sample}}{\text{sample size}} \\ &= \frac{\text{of successes in the sample}}{n} = \frac{\sum x_i}{n} = \bar{x}\end{aligned}$$

with $x_i = 1$ if the individual has the attribute and $x_i = 0$ if not.

Recall from Chapter 6, the sampling distribution of the sample mean \bar{X} :

- Centre: the mean of the sample mean \bar{X} equals the population mean μ . That is,

$$\mu_{\bar{X}} = \mu.$$

- Spread: the standard deviation of the sample mean equals the population standard deviation divided by the square root of the sample size. That is,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

These two arguments are true for any population distribution and sample size n .

- Shape:
 - When the population distribution is normal, \bar{X} is also normal regardless of n .
 - When the population distribution is non-normal but the sample size n is large, \bar{X} is approximately normally distributed. This is guaranteed by the central limit theorem (CLT).

The same conclusions can be applied to the sampling distribution of the sample proportion \hat{p} , where the variable of interest is

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

with the population mean $\mu = p$ and standard deviation $\sigma = \sqrt{p(1-p)}$. Therefore, the sampling distribution of the sample proportion \hat{p} is summarized as follows.

Key Facts: Sampling Distribution of the Sample Proportion

- **Centre:** the mean of the sample proportion \hat{p} equals the population mean μ . That is,

$$\mu_{\hat{p}} = \mu = p.$$

- **Spread:** the standard deviation of the sample proportion \hat{p} equals the population standard deviation σ divided by the square root of the sample size. That is,

$$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

These two arguments are true for any population proportion p and any sample size n .

- **Shape:** The population distribution is non-normal. By the central limit theorem (CLT), however, \hat{p} is approximately normal if n is large enough. The rule of thumb is to guarantee both $np \geq 5$ and $n(1-p) \geq 5$, i.e., $n \geq \max\left\{\frac{5}{p}, \frac{5}{1-p}\right\}$. Some textbooks require both $np \geq 10$ and $n(1-p) \geq 10$.

Central limit theorem for the sample proportion:

If the sample size n is large enough ($np \geq 5$ and $n(1-p) \geq 5$), the sampling distribution of the sample proportion \hat{p} is approximately normally distributed.

For example, suppose the population proportion is $p = 0.05$. Then the sampling distribution of the sample proportion \hat{p} is approximately normally distributed if the sample size is at least

$$n = \max\left\{\frac{5}{p}, \frac{5}{1-p}\right\} = \max\left\{\frac{5}{0.05}, \frac{5}{1-0.05}\right\} = \max\{100, 5.26\} = 100.$$

The following figures show the sampling distribution of the sample proportion with $p = 0.05$ and sample sizes $n = 50, 100, 200$, and 1000.

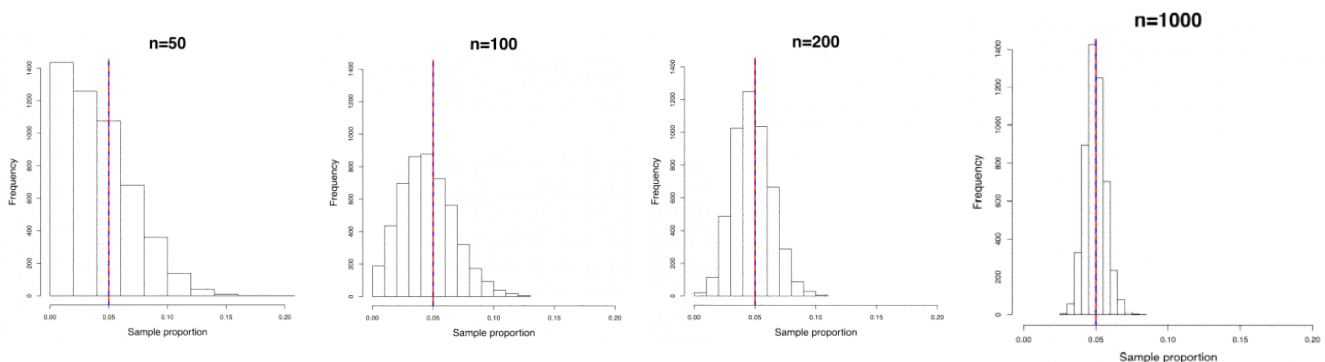


Figure 10.1: Histograms of Sample Proportions with Different Sample Size. [\[Image Description \(See Appendix D Figure 10.1\)\]](#) Click on the image to enlarge it.

There are several findings:

- The sampling distribution of the sample proportion becomes increasingly normal as the sample size n increases. When $n = 50$, the sampling distribution of sample proportion is skewed. When $n = 100$, the distribution is still slightly right skewed. For $n = 200$ and $n = 1000$, the sampling distribution appears bell-shaped and symmetric (indicative of a normal distribution).
- The mean of the sample proportion (blue dashed line) is always identical to the population proportion $p = 0.05$ (red solid line) regardless of the sample size n .
- The standard deviation of the sample proportion decreases as n increases.

To summarize, for $np \geq 5$ and $n(1 - p) \geq 5$, $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. The standardized version of \hat{p} is $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$. As a result, inferences about the population proportions are based on the standard normal distribution.

10.3 One-Proportion z Interval

Assumptions:

1. A simple random sample.
2. Large sample size: both the number of successes x and the number of failures $n - x$ are at least 5.

Note: Recall that one proportion inferences require $np \geq 5$ and $n(1 - p) \geq 5$. However, p is generally unknown, and estimated with $\hat{p} = \frac{x}{n}$. Thus, since $n\hat{p} = n\frac{x}{n} = x$ and $n(1 - \hat{p}) = n(1 - \frac{x}{n}) = n(\frac{n-x}{n}) = n - x$, the sample is deemed sufficiently large if $n\hat{p} = x \geq 5$ and $n(1 - \hat{p}) = n - x \geq 5$. We require at least 5 successes and at least 5 failures in the sample.

A point estimate for the population proportion p is the sample proportion $\hat{p} = \frac{x}{n}$. Therefore, a $(1 - \alpha) \times 100\%$ confidence interval for the population proportion p is

Two-Tailed	Upper-Tailed	Lower-Tailed
$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$	$\left(\hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1 \right)$	$\left(0, \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$

Note: Since the range of proportion is between 0 and 1, the right-end point of the upper-tailed interval is bounded by 1 and the left-end point of the lower-tailed interval is bounded by 0.

Example: One-Proportion Z Interval

A credit card company sent out $n = 400$ advertisements, and $x = 30$ customers responded. Obtain a 95% confidence interval for the proportion of respondents.

Check the assumptions:

1. We have a simple random sample (SRS).
2. Both the number of successes $x = 30$ and number of failures $n - x = 400 - 30 = 370$ are greater than 5.

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{30}{400} = 0.075.$$

$$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies z_{\alpha/2} = z_{0.025} = 1.96.$$

A 95% confidence interval for the proportion of respondents is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.075 \pm 1.96 \times \sqrt{\frac{0.075(1-0.075)}{400}} = (0.049, 0.101).$$

Interpretation: We are 95% confident that the proportion of respondents is somewhere between 0.049 and 0.101, i.e., we are 95% confident that the percentage of respondents is somewhere between 4.9% and 10.1%.

10.4 Margin of Error and Sample Size Calculation for Proportion

A $(1 - \alpha) \times 100\%$ confidence interval for the population proportion p is $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The margin of error is $E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is half of the length of the interval; solving for n yields $n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$. Consequently, we are $(1 - \alpha) \times 100\%$ confident that the margin of error is at most E if the sample size $n \geq \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$. However, this formula cannot be used because the sample proportion $\hat{p} = \frac{x}{n}$ is unknown until the sample is obtained. One solution to this problem is to use the maximum value of $\hat{p}(1 - \hat{p})$, which is 0.25 when $\hat{p} = 0.5$. This leads to the conservative bound on the sample size.

$$n = 0.5(1 - 0.5) \left(\frac{z_{\alpha/2}}{E} \right)^2 = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2$$

rounded up to the nearest integer. However, if we have some extra information about the value of \hat{p} , we can use that information to obtain the guess $\hat{p} = p_g$. This alternative approach leads to the sample size

$$n = p_g(1 - p_g) \left(\frac{z_{\alpha/2}}{E} \right)^2$$

rounded up to the nearest integer.

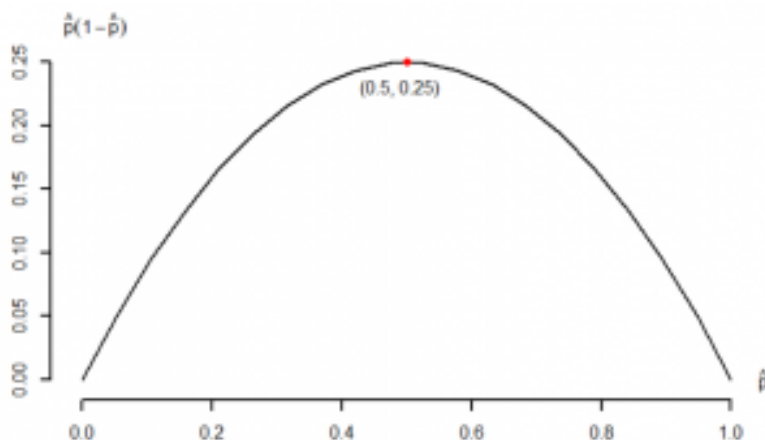


Figure 10.2: Graph of $\hat{p}(1 - \hat{p})$ versus \hat{p} . [Image Description (See Appendix D Figure 10.2)]

\hat{p}	$\hat{p}(1 - \hat{p})$
0	$0 \times (1 - 0) = 0$
0.2	$0.2 \times (1 - 0.2) = 0.16$
0.5	$0.5 \times (1 - 0.5) = 0.25$
0.8	$0.8 \times (1 - 0.8) = 0.16$
1	$1 \times (1 - 1) = 0$

Table 10.2: Relationship Between \hat{p} and $\hat{p}(1 - \hat{p})$.

Example: Sample Size Calculation for Proportion

1. Determine the sample size n such that we are 95% confident that the error is at most 0.05 when \hat{p} is used to estimate p . Use the conservative estimate $\hat{p} = 0.5$.

Since we do not have any extra information about \hat{p} , we will use $n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2$.

$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies z_{\alpha/2} = z_{0.025} = 1.96$, and $E = 0.05$.

$$n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2 = 0.25 \times \left(\frac{1.96}{0.05} \right)^2 = 384.16 \text{ rounded up to } n = 385.$$

2. Suppose p is known to be in between 0.6 and 0.8. Equipped with this new information, obtain a sample size to ensure the margin of error is at most 0.05 with 95% confidence.

We should take this information into account and use $p_g = 0.6$, the value closest to 0.5 within the range $[0.6, 0.8]$. Therefore, the required sample size is

$$n = p_g(1 - p_g) \left(\frac{z_{\alpha/2}}{E} \right)^2 = 0.6(1 - 0.6) \times \left(\frac{1.96}{0.05} \right)^2 = 368.79,$$

rounded up to $n = 369$.

10.5 One-Proportion z Test for p

The assumptions and steps of a one-proportion z test are as follows.

Assumptions:

1. A simple random sample.
2. Both np_0 and $n(1 - p_0)$ are at least 5, where p_0 is the hypothesized value of p under the null H_0 .

Steps to perform a one-proportion z test:

1. Set up the hypotheses:

Two-tailed	Right-tailed	Left-tailed
$H_0 : p = p_0$	$H_0 : p \leq p_0$	$H_0 : p \geq p_0$
$H_a : p \neq p_0$	$H_a : p > p_0$	$H_a : p < p_0$

2. State the significance level α .
3. Compute the value of the test statistic: $z_o = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, with $\hat{p} = \frac{x}{n}$.
4. Find the P-value **or** rejection region.

	Two-tailed	Right-tailed	Left-tailed
Null	$H_0 : p = p_0$	$H_0 : p \leq p_0$	$H_0 : p \geq p_0$
Alternative	$H_a : p \neq p_0$	$H_a : p > p_0$	$H_a : p < p_0$
P-value	$2P(Z \geq z_o)$	$P(Z \geq z_o)$	$P(Z \leq z_o)$
Rejection region	$Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$	$Z \geq z_{\alpha}$	$Z \leq -z_{\alpha}$

5. Reject the null H_0 if the P-value $\leq \alpha$ or z_o falls in the rejection region.
6. Conclusion.

Example: One-Proportion z Test

Let p be the proportion of athletes wearing blue suits who win a judo match. Randomly select $n = 100$ Olympic judo matches and suppose 55 winners wore a blue suit. The other 45 wore a white suit.

a. **Test at the 5% significance level whether a color bias exists.**

If there is no colour bias, the proportions of blue and white winners should be 0.5 and 0.5.

Therefore, letting p be the proportion of winners in blue, the hypotheses are $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$.

Check the assumptions:

1. We have a simple random sample (SRS).
2. Both $np_0 = 100 \times 0.5 = 50$ and $n(1 - p_0) = 100 \times (1 - 0.5) = 50$ are at least 5.

Steps:

1. Set up the hypotheses. $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$
2. State the significance level is $\alpha = 0.05$.
3. Compute the test statistic:

$$\hat{p} = \frac{x}{n} = \frac{55}{100} = 0.55, z_o = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.55 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 1.$$

4. Find the P-value. For a two-tailed test, the P-value is twice the area to the right of the absolute value of the observed test statistic z_o . That is:
P-value = $2P(Z \geq |z_o|) = 2P(Z \geq 1) = 2P(Z \leq -1) = 2 \times 0.1587 = 0.3174$.
5. Decision: Since the P-value = 0.3174 $>$ 0.05(α), we cannot reject the null H_0 .
6. Conclusion: At the 5% significance level, we do not have sufficient evidence that color bias exists in judging Olympic Judo matches.

b. **Obtain a confidence interval corresponding to the test in part (a).**

For a two-tailed test at the 5% significance level, we obtain a 95% two-tailed interval.

The sample proportion is $\hat{p} = \frac{x}{n} = \frac{55}{100} = 0.55$.

$$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies z_{\alpha/2} = z_{0.025} = 1.96.$$

A 95% confidence interval for the proportion of winners in blue is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.55 \pm 1.96 \times \sqrt{\frac{0.55(1-0.55)}{100}} = (0.4525, 0.6475).$$

Interpretation: We are 95% confident that the proportion of winners in blue is somewhere between 0.4525 and 0.6475, i.e., we are 95% confident that the percentage of winners in blue is somewhere between 45.25% and 64.75%.

c. **Does this interval support the conclusion of the one-proportion z test?**

Yes. In part a), we failed to reject $H_0 : p = 0.5$ at the 5% significance level. Similarly, in part b), the 95% confidence interval (0.4525, 0.6475) contains the hypothesized value $p_0 = 0.5$, meaning we don't have sufficient evidence to claim that p is significantly different from 0.5.

10.6 Inferences for Two Population Proportions

Previous studies suggest that more women than men have arthritis. The Centers for Disease Control and Prevention reported a survey of randomly selected Americans aged 65 and older. They found 411 of 1,012 men and 535 of 1,062 women had arthritis. Is there any evidence that women are more likely to suffer from arthritis than men? Let p_1 be the proportion of male arthritis sufferers and p_2 be the proportion of female sufferers. We want to test $H_0 : p_1 \geq p_2$ versus $H_a : p_1 < p_2$ or $H_0 : p_1 - p_2 \geq 0$ versus $H_a : p_1 - p_2 < 0$.

Inference on the population mean μ is based on the distribution of the sample mean \bar{X} ; inference on the difference of two population means $\mu_1 - \mu_2$ is based on the distribution of the difference between the sample means $X_1 - X_2$; and inference on the population proportion p is based on the distribution of the sample proportion \hat{p} . Similarly, inference on the difference of two population proportions $p_1 - p_2$ is based on the distribution of the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$.

10.6.1 Sampling Distribution of Difference Between Two Sample Proportions $\hat{p}_1 - \hat{p}_2$

Key Facts: Sampling Distribution of Difference Between Two Sample Proportions

For independent samples of size n_1 and n_2 from the two populations:

- The mean of $\hat{p}_1 - \hat{p}_2$ equals the difference of the population proportions, i.e.,

$$\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2.$$

- The standard deviation of $\hat{p}_1 - \hat{p}_2$: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$

These two conclusions are always true regardless of the sample sizes n_1 and n_2 .

- The shape of the distribution of $\hat{p}_1 - \hat{p}_2$: by the central limit theorem, when the sample sizes n_1 and n_2 are large enough, $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed. The rule of thumb is $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$ and $n_2 p_2 \geq 5$, $n_2(1 - p_2) \geq 5$.

To summarize, when $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$ and $n_2 p_2 \geq 5$, $n_2(1 - p_2) \geq 5$,

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

The standardized version is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1).$$

10.6.3 Two-Proportion z Interval for the Difference Between Two Proportions $p_1 - p_2$

A point estimate for the difference between two population proportions ($p_1 - p_2$) is the difference between the sample proportions ($\hat{p}_1 - \hat{p}_2$).

Assumptions:

1. Both samples are simple random samples from their respective populations.
2. The two samples are independent.
3. Large samples, all the number of successes, and the number of failures $x_1, n_1 - x_1, x_2$, and $n_2 - x_2$ are at least 5.

Note: As was the case with one-proportion inferences, p_1 and p_2 are generally unknown and estimated with $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$. Thus, since $n_i \hat{p}_i = n_i \frac{x_i}{n_i} = x_i$ and $n_i(1 - \hat{p}_i) = n_i \left(1 - \frac{x_i}{n_i}\right) = n_i \left(\frac{n_i - x_i}{n_i}\right) = n_i - x_i$, the sample is deemed sufficiently large if $n_i \hat{p}_i = x_i \geq 5$ and $n_i(1 - \hat{p}_i) = n_i - x_i \geq 5$ for $i = 1, 2$.

A $(1 - \alpha) \times 100\%$ confidence interval for the difference between the population proportions ($p_1 - p_2$) is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{\alpha/2}$ is the z score such that the area under the standard normal curve to its right is $\frac{\alpha}{2}$. This is a two-tailed interval.

A $(1 - \alpha) \times 100\%$ upper-tail confidence interval is

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, 1 \right),$$

and a $(1 - \alpha) \times 100\%$ lower-tailed confidence interval is

$$\left(-1, (\hat{p}_1 - \hat{p}_2) + z_{\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right).$$

Note that the largest possible value of $p_1 - p_2$ is 1 when $p_1 = 1, p_2 = 0$, and the smallest possible value of $p_1 - p_2$ is -1 when $p_1 = 0, p_2 = 1$.

10.6.2 Two-Proportion z Test for the Difference Between Two Proportions $p_1 - p_2$

Recall that the population proportion can be viewed as the average of the indicator random variable $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$ with a mean $\mu = p$ and standard deviation $\sigma = \sqrt{p(1-p)}$. Note that the standard deviation is a function in p . For a two-tailed test, the null hypothesis is that two population proportions are equal, that is, $H_0 : p_1 = p_2$; consequently, if the null hypothesis is true, it follows that the populations have the same standard deviation. Therefore, similar to a pooled two-sample t-test, we can pool the two samples together to obtain a better estimate of the common standard deviation. If $H_0 : p_1 = p_2$ is true, let $p_1 = p_2 = p_p$, where p_p is the common standard deviation. Then, the test statistic becomes

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{p_p(1-p_p)}{n_1} + \frac{p_p(1-p_p)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_p(1-p_p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

The common proportion p_p is estimated by

$$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}.$$

Assumptions:

1. Both samples are simple random samples from their respective populations.
2. The two samples are independent.
3. Large samples: all the number of successes and failures $x_1, n_1 - x_1, x_2$, and $n_2 - x_2$ and are at least 5.

Steps to perform a two-proportion z test:

1. Set up the hypotheses:

Two-tailed	Right-tailed	Left-tailed
$H_0 : p_1 = p_2$	$H_0 : p_1 \leq p_2$	$H_0 : p_1 \geq p_2$
$H_a : p_1 \neq p_2$	$H_a : p_1 > p_2$	$H_a : p_1 < p_2$

2. State the significance level α .
3. Compute the value of the test statistic:

$$z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ with } \hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}, \hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}.$$

4. Find the P-value **or** rejection region.

	Two-tailed	Right-tailed	Left-tailed
Null	$H_0 : p_1 = p_2$	$H_0 : p_1 \leq p_2$	$H_0 : p_1 \geq p_2$
Alternative	$H_a : p_1 \neq p_2$	$H_a : p_1 > p_2$	$H_a : p_1 < p_2$
P-value	$2P(Z \geq z_o)$	$P(Z \geq z_o)$	$P(Z \leq z_o)$
Rejection region	$Z \geq z_{\alpha/2} \text{ or } Z \leq -z_{\alpha/2}$	$Z \geq z_{\alpha}$	$Z \leq -z_{\alpha}$

5. Reject the null H_0 if the P-value $\leq \alpha$ or z_o falls in the rejection region.
 6. Conclusion.

Example: Two-Proportion z Test and z Interval

The Centers for Disease Control and Prevention reported a survey of randomly selected Americans aged 65 and older. They found 411 of 1,012 men and 535 of 1,062 women had arthritis.

- a. Is there any evidence that women are more likely to suffer from arthritis than men? Test at the 1% significance level.

Let p_1 be the proportion of men who have arthritis and p_2 be the proportion of women who have arthritis.

Check the assumptions:

1. We have simple random samples.
2. The two samples are independent.
3. All the number of successes and failures $x_1 = 411, n_1 - x_1 = 601, x_2 = 535$, and $n_2 - x_2 = 527$ are at least 5.

Steps:

1. Set up the hypotheses: $H_0 : p_1 \geq p_2$ versus $H_a : p_1 < p_2$.
2. State the significance level $\alpha = 0.01$.
3. The test statistic:

$$z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.406 - 0.504}{\sqrt{0.456(1-0.456)}\sqrt{\frac{1}{1012} + \frac{1}{1062}}} = -4.479$$

where

$$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{411 + 535}{1012 + 1062} = 0.456, \hat{p}_1 = \frac{x_1}{n_1} = \frac{411}{1012} = 0.406, \hat{p}_2 = \frac{x_2}{n_2} = \frac{535}{1062} = 0.504.$$

4. Find the P-value. For a left-tailed test, the P-value is the area to the left of the observed test statistic z_o :

$$P\text{-value} = P(Z \leq z_o) = P(Z \leq -4.479) \approx 0.$$

5. Decision: Since the P-value $\approx 0 < 0.01(\alpha)$, we should reject the null H_0 .

6. Conclusion: At the 1% significance level, we have sufficient evidence that women are **more likely** to suffer from arthritis than men.

- b. Obtain a confidence interval for $p_1 - p_2$, corresponding to the test in part a).

For a left-tailed test at the 1% significance level, we should obtain a 99% lower-tailed interval.

$$1 - \alpha = 0.99 \implies \alpha = 0.01 \implies z_\alpha = z_{0.01} = 2.33.$$

A 99% lower-tail confidence interval for $p_1 - p_2$ is

$$\begin{aligned} & \left(-1, (\hat{p}_1 - \hat{p}_2) + z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right) \\ &= \left(-1, (0.406 - 0.504) + 2.33 \sqrt{\frac{0.406(1-0.406)}{1012} + \frac{0.504(1-0.504)}{1062}} \right) = (-1, -0.047). \end{aligned}$$

Interpretation: We are 99% confident that $(p_1 - p_2)$ is below -0.047. That is, we are 99% confident that the proportion of women who have arthritis is at least 0.047 higher than the proportion of men.

- c. Does the interval in part b) support the conclusion of the test in part a)?

Yes. In part a), we reject H_0 and claim $H_a : p_1 < p_2$ (suggesting men have a smaller proportion than women). In part b), the entire interval is below 0, so we are 99% confident that $p_1 - p_2 < 0$.



Activity

Exercises: Inference on Proportions

It is believed that there is an association between breast cancer and smoking. The following table summarizes the results of an observational study of 200 females classified by their disease and smoking status.

	Smoker	Non-smoker	Total
Breast Cancer	10	30	40
Cancer Free	20	140	160
Total	30	170	200

- a. Obtain a 99% confidence interval for the proportion of females with breast cancer.

- Obtain the minimum sample size n needed so that we are 95% confident that the error is at most 0.02 when \hat{p} is used to estimate p . Use the conservative estimate $\hat{p} = 0.5$.
- Test at the 5% significance level whether the proportion of females with breast cancer is higher among smokers than non-smokers.
- Obtain a confidence interval corresponding to the test in part c).

Show/Hide Answer

- Obtain a 99% confidence interval for the proportion of females with breast cancer.
The point estimate for the proportion of females with breast cancer is $\hat{p} = \frac{x}{n} = \frac{40}{200} = 0.2$.
 $1 - \alpha = 0.99 \implies \alpha = 0.01 \implies z_{\alpha/2} = z_{0.005} = 2.575$.

The 99% confidence interval for the proportion of breast cancer is

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.2 \pm 2.575 \times \sqrt{\frac{0.2(1-0.2)}{200}} = (0.127, 0.273).$$

Interpretation: We are 99% confident that the proportion of females with breast cancer is somewhere between 0.127 and 0.273.

- Obtain the minimum sample size n needed so that we are 95% confident that the error is at most 0.02 when \hat{p} is used to estimate p . Use the conservative estimate $\hat{p} = 0.5$.

$$n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2 = 0.25 \left(\frac{2.575}{0.02} \right)^2 = 4144.14, \text{ rounded up to } n = 4145.$$

- Test at the 5% significance level whether the proportion of females with breast cancer is higher among smokers than non-smokers.

Let p_1 be the proportion of females with breast cancer among smokers and p_2 be the proportion of females with breast cancer among non-smokers.

Check the assumptions:

- We have simple random samples.
- The two samples are independent.
- All the number of successes and failures $x_1 = 10, n_1 - x_1 = 20, x_2 = 30$ and $n_2 - x_2 = 140$ are at least 5.

Steps:

- Set up the hypotheses: $H_0 : p_1 \leq p_2$ versus $H_a : p_1 > p_2$.
- The significance level $\alpha = 0.05$.
- Compute the test statistic:

$$z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.333 - 0.176}{\sqrt{0.2(1-0.2)}\sqrt{\frac{1}{30} + \frac{1}{170}}} = 1.982,$$

where

$$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{10 + 30}{30 + 170} = 0.2, \hat{p}_1 = \frac{x_1}{n_1} = \frac{10}{30} = 0.333, \hat{p}_2 = \frac{x_2}{n_2} = \frac{30}{170} = 0.176.$$

- Find the P-value. For a right-tailed test, the P-value is the area to the right of the observed test statistic z_o .

$$\text{P-value} = P(Z \geq z_o) = P(Z \geq 1.982) = P(Z \leq -1.982) = 0.0239.$$

- Decision: Since the P-value = 0.0239 < 0.05(α), we should reject the null H_0 .

6. Conclusion: At the 5% significance level, we have sufficient evidence that the proportion of females with breast cancer is higher among smokers than non-smokers.
- d. Obtain a confidence interval corresponding to the test in part c).
For a right-tailed test at the 5% significance level, we should obtain a 95% upper-tailed confidence interval.

$$\begin{aligned} & \left((\hat{p}_1 - \hat{p}_2) - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, 1 \right) \\ &= \left((0.333 - 0.176) - 1.645 \sqrt{\frac{0.333(1-0.333)}{30} + \frac{0.176(1-0.176)}{170}}, 1 \right) = (0.0075, 1). \end{aligned}$$

Interpretation: We are 95% confident that the proportion of females with breast cancer is at least 0.0075 higher for smokers than non-smokers.

10.7 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain why the sample proportion $\hat{p} = \frac{x}{n}$ is a special type of sample mean $\bar{x} = \frac{\sum x_i}{n}$ (Section 10.1).
- Describe the sampling distribution of the sample proportion \hat{p} (Section 10.2).
- Conduct a one-proportion z-test (Section 10.5).
- Obtain a $(1 - \alpha) \times 100$ confidence interval for the population proportion p (Section 10.3).
- Describe the sampling distribution of the difference between two sample proportions $(\hat{p}_1 - \hat{p}_2)$ (Section 10.6).
- Conduct a two-proportion z-test (Section 10.6).
- Obtain a confidence interval for the difference between two population proportions $(p_1 - p_2)$ (Section 10.6).

10.8 Review Questions

1. In a poll of 1961 randomly selected U.S. adults, 1137 said that they do not believe that abstinence programs are effective in reducing or preventing AIDS.
 - a. At the 2% significance level, do the data provide sufficient evidence to conclude that a majority of all U.S. adults feel that way?
 - b. Obtain a 98% confidence interval for the percentage of U.S. adults who believe that abstinence programs are effective in reducing or preventing AIDS.
 - c. Interpret the interval obtained in part (b). Does it support the result of the test in part (a)?
2. In a clinical trial, 56 patients were randomly assigned to use the Bug Buster kit and 70 were assigned to use the standard treatment. Thirty-two patients in the Bug Buster kit group were cured, whereas nine of those in the standard treatment group were cured.
 - a. At the 5% significance level, do these data provide sufficient evidence to conclude that a difference exists in the cure rates of the two types of treatment?
 - b. Determine a 95% confidence interval for the difference in cure rates for the two types of treatment.
 - c. Interpret the interval obtained in part (b). Does it support the result of the test in part (a)?

Show/Hide Answer

1.

a. Check the assumptions:

■ we have a simple random sample

■ Sample size assumption:

$$np_0 = 1961 \times 0.5 = 980.5 > 5, n(1 - p_0) = 1961 \times (1 - 0.5) = 980.5 > 5.$$

Steps: we have $n = 1961$, $x = 1137$, $\hat{p} = \frac{x}{n} = \frac{1137}{1961} = 0.5798$.

1. Hypotheses: $H_0 : p \leq 0.5$ versus $H_a : p > 0.5$.

2. Significance level $\alpha = 0.02$.

3. Test statistics.

$$z_o = \frac{\hat{p} - p_0}{\sqrt{p_0 \times (1 - p_0) / n}} = \frac{\frac{1137}{1961} - 0.5}{\sqrt{0.5 \times (1 - 0.5) / 1961}} = 7.068.$$

4. P -value: For a right-tailed test, $P\text{-value} = P(Z \geq 7.068) = P(Z \leq -7.068) \approx 0$.

5. Decision: We reject H_0 since $P\text{-value} \approx 0 < 0.02(\alpha)$.

6. Conclusion: At the 2% significance level, we have sufficient evidence that a majority of all U.S. adults do not believe that abstinence programs are effective in reducing or preventing AIDS.

b. $n = 1961, x = 1137, \hat{p} = \frac{x}{n} = \frac{1137}{1961} = 0.5798, 1 - \alpha = 0.98, \alpha = 0.02, z_{\alpha/2} = z_{0.01} = 2.33.$

A 98% confidence interval for the proportion of U.S. adults who do not believe that abstinence programs are effective in reducing or preventing AIDS is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.5798 \pm 2.33 \times \sqrt{\frac{0.5798(1-0.5798)}{1961}} = (0.5538, 0.6057)$$

which gives the percentage is somewhere between 55.38% and 60.57%.

c. Interpretation: we can be 98% confident that the percentage of U.S. adults who do not believe that abstinence programs are effective in reducing or preventing AIDS is somewhere between 55.38% and 60.57%.

Yes, we reject H_0 and claim that $p > 0.5$ in the hypothesis test. The entire 98% confidence interval is above 0.5, so we can claim that $p > 0.5$.

2.

a. Check the assumptions:

- we have simple random samples from the two treatment groups.
- the two samples are independent.
- $x_1 = 32, n_1 - x_1 = 56 - 32 = 24, x_2 = 9$, and $n_2 - x_2 = 70 - 9 = 61$ are all greater than 5.

Steps:

1. Hypotheses: $H_0 : p_1 - p_2 = 0$ versus $H_a : p_1 - p_2 \neq 0$.

2. Significance level $\alpha = 0.05$.

3. Test statistics.

$$z_o = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_p(1-\hat{p}_p)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.5714 - 0.1286}{\sqrt{0.3254(1-0.3254)} \sqrt{\frac{1}{56} + \frac{1}{70}}} = 5.271$$

where the pooled proportion is given by

$$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{32 + 9}{56 + 70} = 0.3254; \hat{p}_1 = \frac{x_1}{n_1} = \frac{32}{56} = 0.5714; \hat{p}_2 = \frac{x_2}{n_2} = \frac{9}{70} = 0.1286.$$

4. P -value: For a two-tailed test, $P\text{-value} = 2P(Z \geq 5.271) = 2P(Z \leq -5.271) \approx 0$.

5. Decision: We reject H_0 since $P\text{-value} \approx 0 < 0.05(\alpha)$.

6. Conclusion: At the 5% significance level, we have sufficient evidence that a difference exists in the cure rates of the two types of treatment.

b.

$$\begin{aligned} & (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= (0.5714 - 0.1286) \pm 1.96 \times \sqrt{\frac{0.5714(1-0.5714)}{56} + \frac{0.1286(1-0.1286)}{70}} \\ &= (0.2913, 0.5943). \end{aligned}$$

- c. Interpretation: We can be 95% confident that the cure rate of the Bug Buster kits group is 0.2913 to 0.5943 higher than the cure rate of the standard treatment. Yes, since the entire interval is above 0, we can claim that $p_1 - p_2 > 0$, which is the conclusion of the hypothesis in part (a).

10.9 Assignment 10

Purposes

The following questions assess your knowledge of properties of the distribution of the sample proportion \hat{p} and the distribution of difference between two sample proportions $\hat{p}_1 - \hat{p}_2$, conducting a one-proportion z test and obtaining a one-proportion z interval, performing a two-proportion z test and determining a two-proportion z interval, and using R commander to conduct a one-proportion and a two-proportion z test.

Resources

[M10_BloodPressure_Proportion_Q5.xlsx](#)

Instructions

Part A

Complete the following:

1. Think of a scenario (you can make up one) you can observe in your daily life and answer the following questions about the basic notation and terminology for proportions.
 - a. What is a population proportion? (2 marks)
 - b. What symbol is used for a population proportion? (1 mark)
 - c. What is a sample proportion? (2 marks)
 - d. What symbol is used for a sample proportion? (1 mark)
 - e. Explain why the sample proportion is a special case of the sample mean. (4 marks)
2. This exercise involves using an unrealistically small population to provide a concrete illustration for the exact distribution of a sample proportion. A population consists of three men and two women. The men's first names are Jose, Pete, and Carlo; the women's first names are Gail and Frances. Suppose that the specified attribute is "female."

- a. Determine the population proportion, p . (2 marks)
- b. The first column of the following table provides the possible samples of size 2, where each person is represented by the first letter of their first name; the second column gives the number of successes—the number of females obtained—for each sample; and the third column shows the sample proportion. Complete the table. (4 marks)

Sample	Number of females x	Sample proportion \hat{p}
J, G	1	0.5
J, P	0	0.0
J, C	0	0.0
J, F	1	0.5
G, P		
G, C		
G, F		
P, C		
P, F		
C, F		

- c. Use the third column of the table to obtain the mean of the variable. (2 marks)
 - d. Compare your answers from parts (a) and (c). What property about the distribution of the sample mean does it verify? (3 marks)
3. In a poll of 1,961 randomly selected U.S. adults, 1,137 said that they do not believe that abstinence programs are effective in reducing or preventing AIDS.
- a. At the 2% significance level, do the data provide sufficient evidence to conclude that a majority of all U.S. adults feel that way? (8 marks)
 - b. Obtain a confidence interval for the percentage of U.S. adults who believe that abstinence programs effectively reduce or prevent AIDS, corresponding to the hypothesis test in part (a). (4 marks)
 - c. Interpret the confidence interval obtained in part (b). Does this interval support the conclusion of the hypothesis test in part (a)? Justify your answer. (4 marks: 2+2)
4. In a clinical trial, 56 patients were randomly assigned to use the Bug Buster kit and 70 were assigned to use the standard treatment. Thirty-two patients in the Bug Buster kit group were cured, whereas nine of those in the standard treatment group were cured.
- a. At the 5% significance level, do these data provide sufficient evidence to conclude that a difference exists in the cure rates of the two types of treatment? (8 marks)

- b. Determine a confidence interval for the difference in cure rates for the two types of treatment corresponding to the hypothesis test in part (a). (4 marks)
 - c. Interpret the confidence interval obtained in part (b). Does this interval support the conclusion of the hypothesis test in part (a)? Justify your answer. (4 marks: 2+2)
5. The following table summarizes the age (≤ 50 or above 50) and blood pressure (BP; normal or high blood pressure) status of 380 Canadian adults in 2017.

	Age 50 or Below	Age Above 50	Total
High BP	74	74	148
Normal BP	139	93	232
Total	213	167	380

- a. It was reported that the percentage of Canadian adults with high blood pressure was 30% in 2007. Test at the 1% significance level whether the percentage of Canadian adults with high blood pressure in 2017 differs from that in 2007. Report the P-value of the test. (8 marks)
- b. Obtain a 99% confidence interval for the proportion of Canadian adults having high blood pressure. (4 marks)
- c. Interpret the confidence interval obtained in part (b). Does this interval support the conclusion of the hypothesis test in part (a)? Justify your answer. (4 marks: 2+2)
- d. Test at the 5% significance level whether the proportion of high blood pressure is higher among Canadian adults above age 50 than among those age 50 or below. (8 marks)
- e. Obtain a confidence interval for the difference between the proportions of high blood pressure among Canadian adults above age 50 and among those age 50 or below corresponding to the hypothesis test in part (d). (5 marks)
- f. Interpret the confidence interval obtained in part (e). Does this interval support the conclusion of the hypothesis test in part (d)? Justify your answer. (4 marks: 2+2)

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs as required and write down your answers in statements.

Refer to Question 5 in Part A. The data are provided in the data file **M10_Bloodpressure_Proportion_Q5.xlsx**. Import the data into R commander.

- a. Re-conduct the test in Question 5 (a) using R commander. Make sure to include all the six components of a hypothesis test. Copy and paste the computer output first and

then compare the answer with the one you obtained by hand in Question 5 (a). (5 marks)

- b. Obtain a confidence interval corresponding to the hypothesis test in part (b). Compare the one you obtained by hand in Question 5 (b). (2 marks)
- c. Re-conduct the test in Question 5 (d) using R commander. Compare the answer with the one you obtained by hand in Question 5 part (d). (5 marks)
- d. Obtain a confidence interval corresponding to the hypothesis test in part (c). Compare the one you obtained by hand in Question 5 part (e). (2 marks)

Quiz 10



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2641#h5p-13>

CHAPTER 11: CHI-SQUARE PROCEDURES

Overview

Chapter 10 covers the z test and z interval for one and two proportions. Chi-square tests are used when at least two proportions are compared.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the main idea behind chi-square tests.
- Describe the differences between the chi-square goodness-of-fit and chi-square independence (homogeneity) tests.
- Identify situations where the chi-square goodness-of-fit or the chi-square independence (homogeneity) test should be used.
- Perform a chi-square goodness-of-fit test and a chi-square independence test.

11.1 Introduction

Could you design an experiment to check whether a coin is unbalanced? A coin is said to be balanced if each of its two faces is equally likely to occur when the coin is tossed. Hence, if we define p as the proportion of times that a head occurs among an infinite number of coin tosses, it follows that $p = 0.5$ if the coin is balanced, while $p \neq 0.5$ if the coin is unbalanced. Therefore, in order to test whether the coin is unbalanced, we may toss the coin n times, record the number of heads observed, and then perform a one-proportion z test to test $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$.

How about an experiment to test whether a die is unbalanced? A 6-sided die is considered balanced if each of its six faces is equally likely to occur when the die is rolled. Hence, define p_i as the proportion of times that face i occurs among an infinite number of rolls. If the die is balanced, then $p_i = \frac{1}{6}$, for $i = 1, 2, 3, 4, 5, 6$; if the die is unbalanced, then $p_i \neq \frac{1}{6}$ for at least one of the faces. Therefore, in order to test whether the die is unbalanced, we may roll the die n times and compute the sample proportions \hat{p}_1 through \hat{p}_6 ; there is evidence that the die is unbalanced, if any \hat{p}_i is significantly different from $\frac{1}{6}$.

The question arises: how do we conduct a hypothesis test when there are six proportions of interest? The naïve approach is to perform six one-proportion z tests. However, this approach is problematic for two main reasons. First, it is time-consuming to conduct six consecutive hypothesis tests; a single test would be more efficient. Second, when several hypothesis tests are performed in succession, the overall type I error rate increases (this is called the multiple comparisons problem). The solution to these problems is to perform a single hypothesis test, with hypotheses

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6} \text{ versus } H_a : p_i \neq \frac{1}{6} \text{ for at least one } i = 1, 2, 3, 4, 5, 6.$$

To test such hypotheses, we rely on new types of tests based on the chi-square distribution. The tests are referred to as chi-square tests.

11.2 Chi-Square Distribution

Just as z-tests are based on the normal distribution and t-tests are based on the t-distribution, chi-square tests are based on the chi-square distribution.

Key Facts: Chi-Square Distribution

1. If $Z \sim N(0, 1)$, then $Z^2 \sim \chi^2$ distribution with degrees of freedom $df = 1$.
2. If Z_1, Z_2, \dots, Z_n are independent and follow a standard normal distribution $N(0, 1)$, then $Z_1^2 + Z_2^2 + \dots + Z_n^2 \sim \chi^2$ distribution with $df = 1 + 1 + \dots + 1 = n$.
3. If $W \sim \chi^2$ with $df = p$, $V \sim \chi^2$ with $df = q$ and they are independent, then $W + V \sim \chi^2$ with $df = p + q$.

Like the t distribution, the chi-square distribution is determined by one parameter, the degrees of freedom. The figure below shows the density curves of chi-square distributions with $df = 1, 3, 5, 9, 15$.

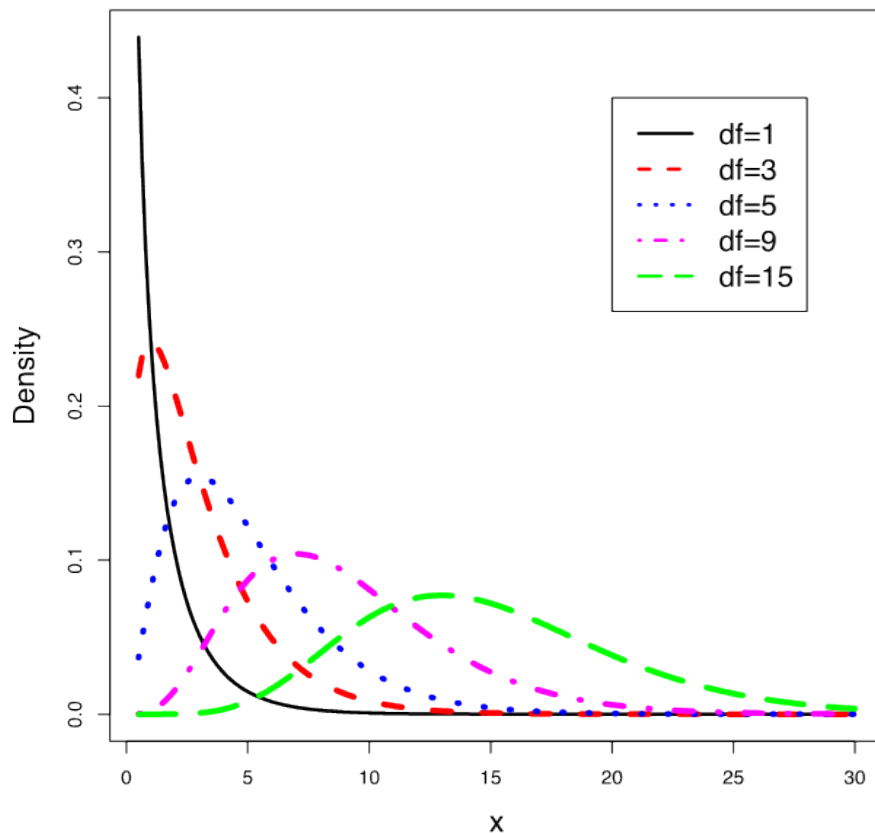


Figure 11.1: Chi-Square Density Curves. [[Image Description \(See Appendix D Figure 11.1\)](#)]

The properties of the chi-square density curve are as follows:

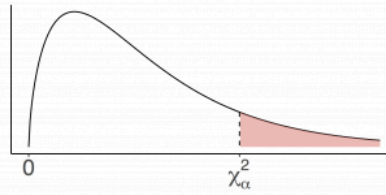
Key Facts: Properties of Chi-Square Density Curve

- The total area under the curve is 1.
- It is right skewed.
- As the degrees of freedom increase, the chi-square curves appear more symmetric.
- Every chi-square random variable is non-negative with possible values between 0 and ∞ .
- The mean of a chi-square is equal to its degrees of freedom and the standard deviation is the square root of twice the degrees of freedom. Suppose the degrees of freedom of a chi-square distribution is γ , then $\mu = \gamma, \sigma = \sqrt{2\gamma}$.

Like the [t score table](#), the χ^2 table (Table V) gives critical values χ_α^2 . Each critical value χ_α^2 has an area of α to its right under the curve of the chi-square distribution with a certain degrees of freedom df .

Table 11.1: Part of Chi-Square Table (Table V)

Table V: Values of χ^2_α of χ^2 -distribution



df	α : Area to the Right of χ^2_α									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.0 ⁴ 393	0.0 ³ 157	0.0 ³ 982	0.0 ² 393	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

[\[Image Description \(See Appendix D Table 11.1\)\]](#)

11.3 Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit test can be applied to either a categorical or discrete quantitative variable with a finite number of values. The objective of the chi-square goodness-of-fit test is to test whether the variable does not follow the probability distribution specified in the null hypothesis H_0 .

The main idea behind the chi-square goodness-of-fit test is to compare the observed frequencies (O) to the expected frequencies (E), which are based on the probability distribution specified in H_0 . If H_0 is true, the observed and expected frequencies should be reasonably similar. Therefore, we reject H_0 if the observed and expected frequencies are very different. The discrepancy between the observed and expected frequencies can be quantified by chi-square statistic

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

which follows a chi-square distribution with $df = k - 1$, where k is the number of possible values for the variable under consideration. The chi-square statistic will be large when the observed and expected frequencies are very different. Thus, we reject the null hypothesis when the chi-square statistic is sufficiently large. More specifically, at the significance level of α , we reject H_0 if the chi-square statistic is larger than the critical value χ^2_{α} . Since we only reject H_0 if the chi-square statistic is sufficiently large, chi-square tests are always right-tailed. That is, both the rejection region and the p-value are upper-tailed probabilities.

Chi-Square Goodness-of-Fit Test

Assumptions:

1. All expected frequencies are at least 1.
2. At most 20% of the expected frequencies are less than 5.
3. Simple random sample (if you need to generalize the conclusion to a larger population).

Note: If assumptions 1 or 2 are violated, one can consider combining the cells to increase the counts in those cells.

Steps to perform a chi-square goodness-of-fit test:

First, check the assumptions. Calculate the expected frequency for each possible value of the variable using $E = np$, where n is the total number of observations and p is the relative frequency (or probability) specified in the null hypothesis. Check whether the expected frequencies satisfy assumptions 1 and 2. If not, consider combining some cells.

1. Set up the hypotheses:

H_0 : The variable has the specified distribution

H_a : The variable does not have the specified distribution.

2. State the significance level α .
3. Compute the value of the test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$ with $df = k - 1$.
4. Find the P-value **or** rejection region based on the χ^2 curve with $df = k - 1$.

Rejection region	$\chi^2 \geq \chi_\alpha^2$ the region to the right of χ_α^2 , the area is α
P-value	$P(\chi^2 \geq \chi_o^2)$ the area to the right of χ_o^2 under the curve

5. Reject the null H_0 if P-value $\leq \alpha$ or χ_o^2 falls in the rejection region.
6. Conclusion.

Example: Chi-Square Goodness-of-Fit Test

According to the results of the federal election in 2015, 31.9% of votes supported the Conservative Party, 39.5% supported the Liberal Party, 19.7% supported the New Democratic Party (NDP), 4.7% supported Bloc Québécois, and 3.4% supported the Green Party (data from Wikipedia). Thirty-seven students in my Stat151 class responded to an online survey and their preferences are summarized in the following table:

Table 11.2: Voting Preference of the Class

Conservative	Green	Liberal	NDP	Not Voting	Others
9	2	17	6	3	0

Test at the 5% significance level whether the class had different voting preferences than all Canadians in the 2015 election.

Check the assumptions: since $n = 37$, each expected frequency is computed as $E = np = 37 \times p$. For example, the expected count of conservative voters is $E = 37 \times 0.319 = 11.803$. The following table gives all expected counts:

Table 11.3: Expected Frequency of Voting Preference

	Conservative	Green	Liberal	NDP	Bloc Québécois	Others
Proportion (p)	0.319	0.034	0.395	0.197	0.047	0.008
Counts	11.803	1.258	14.615	7.289	1.739	0.296

There are $k = 6$ cells and at most $6 \times 0.2 = 1.2$ cells are expected to have expected counts less than 5; however, there are actually three cells less than 5. We could combine the cells “Green”, “Bloc Québécois” and “Others”, and name it as “Others”. Therefore, we have the working table as follows.

Table 11.4: Working Table for a Chi-Square Goodness of Fit Test (Example)

Parties	Proportion p	Observed O	Expected $E = np = 37 \times p$	$\frac{(O-E)^2}{E}$
Conservative	0.319	9	$37 \times 0.319 = 11.803$	$\frac{(9-11.803)^2}{11.803} = 0.6657$
Liberal	0.395	17	$37 \times 0.395 = 14.615$	$\frac{(17-14.615)^2}{14.615} = 0.3892$
NDP	0.197	6	$37 \times 0.197 = 7.289$	$\frac{(6-7.289)^2}{7.289} = 0.2279$
Others	0.089	$2 + 3 + 0 = 5$	$37 \times 0.089 = 3.293$	$\frac{(5-3.293)^2}{3.293} = 0.8849$
	Sum = 1	Sum = 37	Sum = 37	Sum = $\chi_o^2 = 2.1667$

Note: After combining the cells, all the expected counts are greater than 1, while 25% of the expected counts are below 5 (the expected count for Others is below 5). Since more than 20% of the expected counts are below 5, there is still a violation in the assumptions. However, the expected frequency for “Others” is 3.293 which is not very far away from 5. To maintain a meaningful number of parties, we proceed to conduct the chi-square goodness-of-fit test.

Steps to perform a chi-square goodness-of-fit test:

1. Set up the hypotheses:

$$H_0 : p_C = 0.319, p_L = 0.395, p_{NDP} = 0.197, p_{Others} = 0.089$$

$$H_a : \text{At least one proportion is different from those specified in } H_0.$$

2. The significance level is $\alpha = 0.05$.
3. The test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = 2.1677$, with $df = k - 1 = 4 - 1 = 3$.
4. Find the P-value. Since chi-square tests are always right-tailed, the p-value is
P-value = $P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 2.1677) = 0.1$.
5. Decision: We do not reject the null H_0 , since P-value 0.1 $>$ 0.05(α).

- Conclusion: At the 5% significance level, we do not have sufficient evidence that the class had different voting preferences than all Canadians in the 2015 election.

If using the critical value approach, steps 4–6 are as follows:

- Find the rejection region. For a right-tailed test with $df = 3$, the rejection region is to the right of the critical value $\chi^2 \geq \chi^2_{\alpha} = \chi^2_{0.05} = 7.815$.
- Decision: We do not reject the null H_0 since $\chi^2_o = 2.1667 < 7.815$ falls in the non-rejection region.
- Conclusion: At the 5% significance level, we do not have sufficient evidence that the class had different voting preferences than all Canadians in the 2015 election.



Activity

Exercise: Chi-square goodness-of-fit test

A company claims their deluxe mixed nuts consist of 20% peanuts, 60% cashews, and 20% almonds. An inspector obtains a random sample of $n = 100$ nuts and observes 30 peanuts, 55 cashews, and 15 almonds. Test at the 5% significance level whether the percentages differ from what the company claims.

Show/Hide Answer

Answers:

Check the assumptions: $n = 100$ and the expected counts are

$E_{\text{peanut}} = 100 \times 0.2 = 20$, $E_{\text{cashew}} = 100 \times 0.6 = 60$, $E_{\text{almond}} = 100 \times 0.2 = 20$ and all greater than 5.

Steps to perform a chi-square goodness-of-fit test:

- Set up the hypotheses:

$$H_0 : p_{\text{peanut}} = 0.2, p_{\text{cashew}} = 0.6, p_{\text{almond}} = 0.2$$

$$H_a : \text{at least one proportion is different from those specified in } H_0.$$

- The significance level is $\alpha = 0.05$.
- The test statistic with the working table:

Table 11.5: Working Table for Chi-Square Goodness-of-Fit Test (Exercise)

Nuts	Proportion p	Observed (O)	Expected $E = np = 100 \times p$	$\frac{(O-E)^2}{E}$
Peanut	0.2	30	$100 \times 0.2 = 20$	$\frac{(30-20)^2}{20} = 5.000$
Cashew	0.6	55	$100 \times 0.6 = 60$	$\frac{(55-60)^2}{60} = 0.417$
Almond	0.2	15	$100 \times 0.2 = 20$	$\frac{(15-20)^2}{20} = 1.250$
Sum = 1		Sum = 100	Sum = 100	Sum = $\chi_o^2 = 6.667$

$$\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = 6.667 \text{ with } df = k - 1 = 3 - 1 = 2.$$

4. Find the P-value: P-value $P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 6.667)$.
Since $5.991(\chi_{0.05}^2) < \chi_o^2 = 6.667 < 7.378(\chi_{0.025}^2)$, $0.025 < \text{P-value} < 0.05$.
5. Decision: We should reject the null H_0 since P-value $< 0.05(\alpha)$.
6. Conclusion: At the 5% significance level, we have sufficient evidence that the percentages of nuts are different from what the company claims.

11.4 Chi-Square Independence Test

The chi-square independence test is used to test for an association between two categorical variables of a population.

11.4.1 Terminologies Used for a Contingency Table

Recall that a contingency table summarizes the counts of two categorical variables. For example, the following contingency table groups 200 females according to their breast cancer status and smoking status:

Table 11.6: Contingency Table of Cancer Status (row) and Smoking Status (column)

	Smoker (S_1)	Non-smoker (S_2)	Total
Breast Cancer (C_1)	10 ($C_1 \& S_1$)	30 ($C_1 \& S_2$)	40
Cancer-free (C_2)	20 ($C_2 \& S_1$)	140 ($C_2 \& S_2$)	160
Total	30	170	200

Suppose we randomly select an individual from this sample. Define the events:

- S_1 = the subject is a smoker;
- S_2 = the subject is a non-smoker;
- C_1 = the subject has breast cancer;
- C_2 = the subject does not have breast cancer.

The joint events are:

- $C_1 \& S_1$ = the subject has cancer and is a smoker;
- $C_1 \& S_2$ = the subject has cancer and is a non-smoker;
- $C_2 \& S_1$ = the subject does not have cancer and is a smoker;
- $C_2 \& S_2$ = the subject does not have cancer and is a non-smoker.

The variable “Cancer Status” is called the **row variable**, and it has two possible values—**cancer** or **cancer-free**. The variable “Smoking Status” is the **column variable**, and it

has two values—smoker and non-smoker. The two numbers in the last column (40 and 160) are the **row totals** and the two in the last row (30, 170) are the **column totals**. The sample size is also called the *grand total*. The four numbers in bold are the joint frequencies. The boxes that contain the joint frequencies are referred to as cells.

Based on the $\frac{f}{N}$ rule, the **marginal distribution** of the **row** (**column**) variable equals the **row** (**column**) totals divided by n . The **joint distribution** is given by the joint frequencies divided by n . The following table shows the marginal distribution of “Cancer Status” in the last column, the marginal distribution of “Smoking Status” in the last row, and the joint distribution of the four cells.

Table 11.7: Marginal and Joint Probability Distributions of Cancer Status and Smoking Status

	Smoker (S_1)	Non-smoker (S_2)	Total
Breast Cancer (C_1)	$P(C_1 \& S_1) = \frac{10}{200} = 0.05$	$P(C_1 \& S_2) = \frac{30}{200} = 0.15$	$P(C_1) = \frac{40}{200} = 0.2$
Cancer-free (C_2)	$P(C_2 \& S_1) = \frac{20}{200} = 0.1$	$P(C_2 \& S_2) = \frac{140}{200} = 0.7$	$P(C_1) = \frac{160}{200} = 0.8$
Total	$P(S_1) = \frac{30}{200} = 0.15$	$P(S_2) = \frac{170}{200} = 0.85$	1

We want to test for an association between the two variables in a contingency table. Two variables are said to be associated if they are NOT independent. If two variables are associated, then differences exist among the conditional distributions of one variable, given different values of the other variable. For example, the conditional distributions of “Cancer Status” given “Smoking Status” are given in the following table. Notice that the conditional distributions are simply the relative frequencies of “Cancer” within smoker and non-smoker groups.

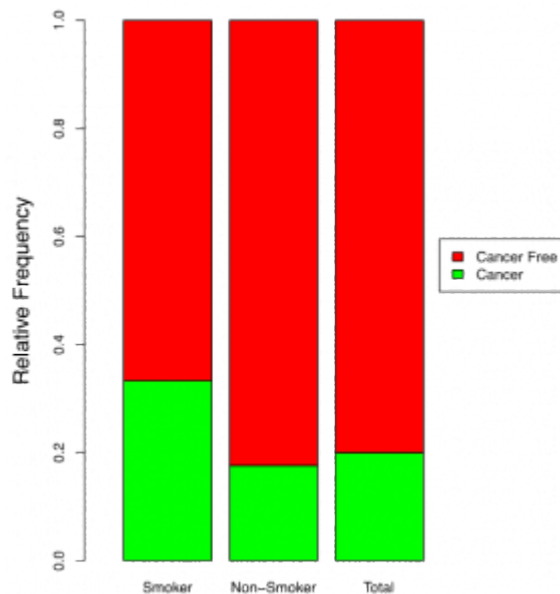
Table 11.8: Conditional Probability Distribution of Cancer Status Given Smoking Status

Breast Cancer (C_1)	$P(C_1 S_1) = \frac{10}{30} = 0.333$	$P(C_1 S_2) = \frac{30}{170} = 0.176$	$P(C_1) = \frac{40}{200} = 0.2$
Cancer-free (C_2)	$P(C_2 S_1) = \frac{20}{30} = 0.677$	$P(C_2 S_2) = \frac{140}{170} = 0.824$	$P(C_2) = \frac{160}{200} = 0.8$
Total	1	1	1

A segmented bar graph helps us visualize conditional distributions and the concept of association. The figure below is the segmented bar graph that displays the conditional distributions of “Cancer Status” for smokers and non-smokers and the marginal distribution of “Cancer Status”. The three bars should be identical if “Cancer Status” and

“Smoking Status” are independent. That is, the conditional probabilities should equal the unconditional probabilities:

$$P(C_1|S_1) = P(C_1|S_2) = P(C_1); P(C_2|S_1) = P(C_2|S_2) = P(C_2).$$



Interpretation:

The proportion or percentage of females with breast cancer (the green bar) is higher among the smokers than the non-smokers. Therefore, “Cancer Status” and “Smoking Status” might be associated; we can test this by a chi-square independence test.

Figure 11.2: Segment Bar Chart. [\[Image Description \(See Appendix D Figure 11.2\)\]](#) Click on the image to enlarge it.

11.4.2 Main Idea Behind Chi-Square Independence Test

The null hypothesis is that the two variables are independent; the alternative is that they are associated. The test statistic is the same as that from the chi-square goodness-of-fit test; for each cell, compute the difference between the observed frequency (O) and the expected frequency (E), square it, and divide by the expected frequency. The expected frequency is the number we expect to observe if the null is true. A large chi-square statistic means the observed and the expected frequencies are significantly different, which provides evidence against the null hypothesis. Therefore, we should reject the null if the observed chi-square statistic is sufficiently large. More specifically, given the significance level α , reject H_0 if the $P\text{-value} \leq \alpha$, where the $P\text{-value}$ is the area to the **right** of the observed test statistic under the chi-square curve.

The test procedure is straightforward—the key is calculating each cell’s expected frequency. Recall that two events, A and B , are independent if $P(A \& B) = P(A) \times P(B)$. For example, if the events “Breast Cancer” and “Smoker” are independent, then

$P(\text{Breast Cancer} \mid \text{Smoker}) = P(\text{Breast Cancer}) \times P(\text{Smoker})$ where $P(\text{Breast Cancer})$ and $P(\text{Smoker})$ are given by the marginal distribution of “Cancer Status” and “Smoking Status” respectively. That is,

$$P(\text{Breast Cancer}) = \frac{40}{200} = 0.2; P(\text{Smoker}) = \frac{30}{200} = 0.15.$$

If H_0 (the two variables are independent) is true, the expected frequency for the cell “Cancer and Smoker” is

$$E = nP(\text{Cancer and Smoker}) = nP(\text{Cancer})P(\text{Smoker}) = 200 \times \frac{40}{200} \times \frac{30}{200} = \frac{40 \times 30}{200} = 6.$$

In general,

$$\text{Expected frequency of the cell in } r\text{th row and } c\text{th column} = \frac{r\text{th row total} \times c\text{th column total}}{n}.$$

Applying the above formula to each cell yields the following expected frequencies:

- “Cancer” & “Smoker”: $E = \frac{40 \times 30}{200} = 6$.
- “Cancer” & “Non-smoker”: $E = \frac{40 \times 170}{200} = 34$.
- “Cancer free” & “Smoker”: $E = \frac{160 \times 30}{200} = 24$.
- “Cancer free” & “Non-smoker”: $E = \frac{160 \times 170}{200} = 136$.

To compute the test statistic, it is helpful to write each expected frequency in the same cell as the corresponding observed frequency. The following table gives both the observed and expected frequencies for each cell (the expected frequencies are displayed in brackets):

Table 11.9: Observed and Expected Frequency (in Brackets) of Chi-Square Independent Test

	Smoker (S_1)	Non-smoker (S_2)	Total
Breast Cancer (C_1)	10 (6)	30 (34)	40
Cancer-free (C_2)	20 (24)	140 (136)	160
Total	30	170	200

Chi-Square Independence Test

The assumptions and steps of conducting a chi-square independence test are as follows.

Assumptions:

1. All expected frequencies are at least 1.
2. At most 20% of the expected frequencies are less than 5.
3. Simple random sample (required only if you need to generalize the conclusion to a larger population).

Note: If either assumption 1 or 2 is violated, one can consider combining the cells to make the counts in those cells larger.

Steps to perform a chi-square independence test:

First, check the assumptions. Calculate the expected frequency for each possible value of the variable using $E = \frac{\text{rth row total} \times \text{cth column total}}{n}$, where n is the total number of observations. Check whether the expected frequencies satisfy assumptions 1 and 2. If not, consider combining some cells.

1. Set up the hypotheses:

H_0 : The two variables are independent

H_a : The two variables are associated.

2. State the significance level α .
3. Compute the value of the test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$ with, $df = (r - 1) \times (c - 1)$ where $E = \frac{\text{rth row total} \times \text{cth column total}}{n}$, r is the number of rows and c is number of columns of the cells.
4. Find the P-value **or** rejection region based on the χ^2 curve with $df = (r - 1) \times (c - 1)$

P-value	$P(\chi^2 \geq \chi_o^2)$ the area to the right of χ_o^2 under the curve
Rejection region	$\chi^2 \geq \chi_\alpha^2$ the region to the right of χ_α^2

5. Reject the null H_0 if the P-value $\leq \alpha$ or χ_o^2 falls in the rejection region.
6. Conclusion.

Example: Chi-Square Independence Test

Test at the 10% significance level whether the variables “Cancer Status” and “Smoking Status” are associated.

	Smoker (S1)	Non-smoker(S2)	Total
Breast Cancer (C_1)	10 (6)	30 (34)	40
Cancer-free (C_2)	20 (24)	140 (136)	160
Total	30	170	200

Check the assumptions: The expected frequencies are the values given in brackets, all greater than 5. We must assume this is a simple random sample of females.

Steps:

- Set up the hypotheses:
 H_0 : The variables "Cancer Status" and "Smoking Status" are independent
 H_a : The variables "Cancer Status" and "Smoking Status" are associated.
- The significance level is $\alpha = 0.1$.
- Compute the value of the test statistic:

$$\begin{aligned}\chi_o^2 &= \sum_{\text{all cells}} \frac{(O - E)^2}{E} \\ &= \frac{(10 - 6)^2}{6} + \frac{(30 - 34)^2}{34} + \frac{(20 - 24)^2}{24} + \frac{(140 - 136)^2}{136} \\ &= 3.922.\end{aligned}$$

with $df = (r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1$.

- Find the P-value:
P-value = $P(\chi^2 \geq X_0^2) = P(\chi^2 \geq 3.992) \implies 0.025 < \text{P-value} < 0.05$ since $3.841(\chi_{0.05}^2) < \chi_o^2 = 3.922 < 5.024(\chi_{0.025}^2)$.
- Decision: Reject the null H_0 since $\text{P-value} \leq 0.05 < 0.1(\alpha)$.
- Conclusion: At the 10% significance level, we have sufficient evidence of an association between the variables "Cancer Status" and "Smoking Status".



Activity

Exercise: Chi-Square Independence Test

A random sample of 230 adults yields the following data regarding age and Internet usage. At the 1%

significance level, do the data provide sufficient evidence of an association between age and Internet usage?

Table 11.10: Contingency Table of Internet Usage (row) and Age (column)

	18-24	25-64	65+	Total
Never	6	38	31	75
Sometimes	14	31	5	50
Every day	50	50	5	105
Total	70	119	41	230

Show/Hide Answer

Answers:

Check the assumptions.

Applying the formula Expected frequency of the cell in rth row and cth column = $\frac{\text{rth row total} \times \text{cth column total}}{n}$ to each cell, the expected frequencies are given by:

- “Never” & “18-24”: $E = \frac{75 \times 70}{230} = 22.826$.
- “Never” & “25-64”: $E = \frac{75 \times 119}{230} = 38.804$.
- “Never” & “65+”: $E = \frac{75 \times 41}{230} = 13.370$.
- “Sometimes” & “18-24”: $E = \frac{50 \times 70}{230} = 15.217$.
- “Sometimes” & “25-64”: $E = \frac{50 \times 119}{230} = 25.870$.
- “Sometimes” & “65+”: $E = \frac{50 \times 41}{230} = 8.913$.
- “Every day” & “18-24”: $E = \frac{105 \times 70}{230} = 31.957$.
- “Every day” & “25-64”: $E = \frac{105 \times 119}{230} = 54.326$.
- “Every day” & “65+”: $E = \frac{105 \times 41}{230} = 18.717$.

The expected frequencies are given in brackets; they are all greater than 5. We are told this is a random sample. Therefore, assumptions for the chi-square independence test are satisfied.

Table 11.11: Observed and Expected Frequency of Internet Usage (row) and Age (column)

	18-24	25-64	65+	Total
Never	6 (22.826)	38 (38.804)	31 (13.370)	75
Sometimes	14 (15.217)	31 (25.870)	5 (8.913)	50
Every day	50 (31.957)	50 (54.326)	5 (18.717)	105
Total	70	119	41	230

Steps:

1. Set up the hypotheses:

H_0 : The variables “Age” and “Internet usage” are independent

H_a : The variables “Age” and “Internet usage” are associated.

2. The significance level is $\alpha = 0.01$.

3. Compute the value of the test statistic:

$$\begin{aligned}\chi_o^2 &= \sum_{\text{all cells}} \frac{(O - E)^2}{E} \\ &= \frac{(6 - 22.826)^2}{22.826} + \frac{(38 - 38.804)^2}{38.804} + \frac{(31 - 13.370)^2}{13.370} + \frac{(14 - 15.217)^2}{15.217} \\ &\quad + \frac{(31 - 25.870)^2}{25.870} + \frac{(5 - 8.913)^2}{8.913} + \frac{(50 - 31.957)^2}{31.957} + \frac{(50 - 54.326)^2}{54.326} \\ &\quad + \frac{(5 - 18.717)^2}{18.717} = 59.084.\end{aligned}$$

with

$$df = (r - 1) \times (c - 1) = (3 - 1) \times (3 - 1) = 4.$$

4. Find the P-value: $P\text{-value} = P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 59.084) < 0.005$.
5. Decision: Reject the null H_0 since $P\text{-value} \leq 0.005 < 0.01(\alpha)$.
6. Conclusion: At the 1% significance level, we have sufficient evidence that there is an association between age and Internet usage.

11.5 Chi-Square Homogeneity Test

The chi-square homogeneity test is the same as the chi-square independence test, except for the wording of the hypotheses. The hypotheses for a chi-square homogeneity test are:

H_0 : The population proportions are homogeneous.

H_a : The population proportions are nonhomogeneous.

For example, recall the chi-square independence test for “cancer status” and “smoking status”. The hypotheses for the corresponding chi-square homogeneity test are:

H_0 : The proportions of females with without cancer status are homogeneous between smokers and nonsmokers.

H_a : The proportions of females with without cancer status are nonhomogeneous between smokers and nonsmokers.

That is,

$H_0 : p_{\text{cancer—smoker}} = p_{\text{cancer—nonsmoker}}; p_{\text{cancer free—smoker}} = p_{\text{cancer free—non smoker}}.$

$H_a : p_{\text{cancer—smoker}} \neq p_{\text{cancer—nonsmoker}}; p_{\text{cancer free—smoker}} \neq p_{\text{cancer free—non smoker}}.$

Note: In the case where the variable has more than 2 levels, lack of homogeneity does not imply that all pairs of proportions are different between the populations but that at least one pair of proportions is different.

The remaining steps for the chi-square homogeneity test are identical to the chi-square independence test.

11.6 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Explain the main idea behind chi-square tests (Section 11.1).
- Describe the differences between the chi-square goodness-of-fit and chi-square independence (homogeneity) tests (Sections 11.3, 11.4 and 11.5).
- Identify situations where the chi-square goodness-of-fit or the chi-square independence (homogeneity) test should be used (Sections 11.3, 11.4 and 11.5).
- Perform a chi-square goodness-of-fit test and chi-square independence test (Sections 11.3 and 11.4).

11.7 Review Questions

1. An American roulette wheel contains 18 red numbers, 18 black numbers, and 2 green numbers. The following table shows the frequency with which the ball landed on each color in 300 trials.

Color	Red	Black	Green
Frequency	140	120	40

- At the 5% significance level, do the data suggest that the wheel is out of balance?
2. A gambler thinks a die may be loaded and that the six numbers are not equally likely. To test his suspicion, he rolled the die 150 times and obtained the data in the following table.

Number	1	2	3	4	5	6
Frequency	23	26	23	21	31	26

- Do the data provide sufficient evidence to conclude that the die is loaded? Perform the hypothesis test at the 5% significance level.
3. The following table reported the survey results on how members would prefer to receive ballots in annual elections. At the 5% significance level, do the data provide sufficient evidence to conclude that gender (column) and preference (row) are associated?

	Male	Female	Total
Mail	60	30	90
Email	150	90	240
Both	70	40	110
N/A	80	50	130
Total	360	210	570

Show/Hide Answer

1. If the wheel is balanced, the chance of landing on a red number is $p_1 = \frac{18}{18+18+2}$, on a black number is $p_2 = \frac{18}{18+18+2}$, and on a green number is $p_3 = \frac{2}{18+18+2}$. If the wheel is out of balance, at least one proportion differs from the specified value. We can use the chi-

square goodness of fit test to test this. The assumptions of a goodness-of-fit test are:

1. All expected frequencies are at least 1.
2. At most 20% of the expected frequencies are less than 5.
3. Simple random sample (if you need to generalize the conclusion to a larger population)

The expected frequencies are $E_1 = np_1 = 300 \times \frac{18}{38} = 142.1053$, $E_2 = np_2 = 300 \times \frac{18}{38} = 142.1053$, $E_3 = np_3 = 300 \times \frac{2}{38} = 15.78947$. All the expected frequencies are greater than 5. We assume the trials were conducted randomly.

Steps:

1. Hypotheses. $H_0 : p_1 = \frac{18}{38}, p_2 = \frac{18}{38}, p_3 = \frac{2}{38}$ versus H_a : at least one proportion is different from the specified value.
2. Significance level $\alpha = 0.05$
3. Test statistic.

Color	Observed	Expected	Contribution
red	140	142.10526	0.0311891
black	120	142.10526	3.4385965
green	40	15.78947	37.1228070

$$\begin{aligned}\chi_o^2 &= \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} = \frac{(140 - 142.10526)^2}{142.10526} + \frac{(120 - 142.10526)^2}{142.10526} + \frac{(40 - 15.78947)^2}{15.78947} \\ &= 0.03118908 + 3.43859649 + 37.12280702 \\ &= 40.59259.\end{aligned}$$

with $df = k - 1 = 3 - 1 = 2$.

4. P-value. The chi-square test is always right-tailed. P-value is the area to the right of the observed value χ_o^2 under the chi-square density curve with degrees of freedom $df = k - 1 = 3 - 1 = 2$. $P\text{-value} = P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 40.593)P(\chi^2 \geq 10.597) = 0.005$. That is, $P\text{-value} < 0.005$.

If the critical value approach is used, we need to find the rejection region. The chi-square test is always right-tailed. Make sure that the area of the rejection region is α , so the critical value is $\chi_\alpha^2 = \chi_{0.05}^2 = 5.991$.

5. Decision. We reject H_0 since $P\text{-value} < 0.005 < 0.05$ (α). If the critical approach is used, we reject H_0 since the observed test statistic $\chi_o^2 = 40.593 > 5.991$ falls in the rejection region.
6. Conclusion. At the 5% significance level, we have sufficient evidence that the wheel is out of balance.

2. If the die is balanced, the proportion of landing on each of 1, 2, \dots , 6 should be the

same, i.e., $p_i = \frac{1}{6}, i = 1, 2, \dots, 6$. All expected frequencies are $E_i = np_i = 150 \times \frac{1}{6} = 25 > 5$, the assumptions for a chi-square goodness-of-fit test are satisfied.

Steps:

1. Hypotheses. $H_0 : p_i = \frac{1}{6}, i = 1, 2, \dots, 6$ versus H_a : at least one proportion is equal to $\frac{1}{6}$.
2. Significance level $\alpha = 0.01$
3. Test statistic. The following table is useful.

$$\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = \frac{(23-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(23-25)^2}{25} + \frac{(21-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(26-25)^2}{25} = 2.48$$

with the degrees of freedom of the test $df = k - 1 = 6 - 1 = 5$.

4. P-value. The chi-square test is always right-tailed. P-value is the area to the right of the observed value χ_o^2 under the chi-square density curve with degrees of freedom $df = k - 1 = 6 - 1 = 5$. $P\text{-value} = P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 2.48) > P(\chi^2 \geq 9.236) = 0.1$, i.e., $P\text{-value} > 0.1$.

Or

Rejection region. The chi-square test is always right-tailed. Make sure that the area of the rejection region is α , so the critical value is $\chi_\alpha^2 = \chi_{0.01}^2 = 15.086$.

5. Decision: We cannot reject H_0 since $P\text{-value} > 0.1 > 0.01$ (α).

Or

We cannot reject H_0 since $\chi_o^2 = 2.48 < 15.086$ falls in the non-rejection region.

6. Conclusion: At the 5% significance level, we do not have sufficient evidence that the die is loaded.

3. We should use the chi-square independence test. The expected frequencies are given in the following table. All the expected frequencies are greater than 5, the assumptions for a chi-square independence test are satisfied.

Preference	Male	Female
Mail	56.84211	33.15789
Email	151.57895	88.42105
Both	69.47368	40.52632
N/A	82.10526	47.89474

Steps:

1. Hypotheses. H_0 : gender and preference are independent. versus H_a : gender and preference are associated.
2. Significance level $\alpha = 0.05$.
3. Test statistic.

$$\begin{aligned}
\chi_o^2 &= \sum_{\text{all cells}} \frac{(O - E)^2}{E} \\
&= \frac{(60 - 56.842)^2}{56.842} + \frac{(30 - 33.158)^2}{33.158} + \frac{(150 - 151.579)^2}{151.579} + \frac{(90 - 88.421)^2}{88.421} \\
&\quad + \frac{(70 - 69.474)^2}{69.474} + \frac{(40 - 40.526)^2}{40.526} + \frac{(80 - 82.105)^2}{82.105} + \frac{(50 - 47.895)^2}{47.895} \\
&= 0.678
\end{aligned}$$

with the degrees of freedom of the test $df = (r - 1) \times (c - 1) = (4 - 1) \times (2 - 1) = 3$.

4. P-value. The chi-square test is always right-tailed. P-value is the area to the right of the observed value χ_o^2 under the chi-square density curve with degrees of freedom $df = 3$. $P\text{-value} = P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 0.678) > P(\chi^2 \geq 6.251) = 0.1$, i.e., P-value > 0.10 .

Or

Rejection region. The chi-square test is always right-tailed. Make sure that the area of the rejection region is α , so the critical value is $\chi_\alpha^2 = \chi_{0.05}^2 = 7.815$.

5. Decision: Don't reject H_0 , since $P\text{-value} > 0.1 > 0.05(\alpha)$.

Or

We cannot reject H_0 since $\chi_o^2 = 0.678 < 7.815$ falls in the non-rejection region.

6. Conclusion: At the 5% significance level, we do not have sufficient evidence that gender and preference are associated.

11.8 Assignment 11

Purposes

This assignment has two parts. The first part assesses your knowledge of conducting a chi-square goodness-of-fit test and performing a chi-square independence test using the chi-square table. The second part assesses your skills in using R commander to conduct a chi-square goodness-of-fit and chi-square independence tests.

Resources

[M11_Wheel_Chisquare_Q2.xlsx](#)

[M11_BloodPressure_Age_Chisquare_Q5.xlsx](#)

Instructions

Part A

Complete the following:

1. The t -table has entries for areas of 0.10, 0.05, 0.025, 0.01, and 0.005. In contrast, the χ^2 -table has entries for those areas and for 0.995, 0.99, 0.975, 0.95, and 0.90. Explain why the t -values corresponding to these additional areas can be obtained from the existing t -table but must be provided explicitly in the χ^2 -table. (3 marks)
2. An American roulette wheel contains 18 red numbers, 18 black numbers, and 2 green numbers. The following table shows the frequency with which the ball landed on each colour in 300 trials. At the 5% significance level, do the data suggest that the wheel is out of balance? (10 marks)

Color	Red	Black	Green
Frequency	140	120	40

3. A gambler thinks a die may not be landed on the six numbers with equal chance. To

test his suspicion, he rolled the die 150 times and obtained the data in the following table. Test at the 1% significance level whether the die is balanced. (10 marks)

Number	1	2	3	4	5	6
Frequency	23	26	23	21	31	26

4. The following table reported the survey results on how members would prefer to receive ballots in annual elections. At the 5% significance level, do the data provide sufficient evidence to conclude that gender (column) and preference (row) are associated? (10 marks)

	Male	Female	Total
Mail	60	30	90
Email	150	90	240
Both	70	40	110
N/A	80	50	130
Total	360	210	570

5. The following table summarizes the age (column) and blood pressure (BP; row) status of 474 randomly selected Canadian adults in 2017. At the 10% significance level, do the data provide sufficient evidence to conclude that age and blood pressure are associated? (10 marks)

	Under 30	30-49	Over 50	Total
High BP	23	51	73	147
Normal BP	48	91	93	232
Low BP	27	37	31	95
Total	98	179	197	474

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs as required, and write down your answers in statements.

- Refer to Question 2 in Part A. The data are provided in the data file **M11_Wheel_Chisquare_Q2.xlsx**. Import the data into R commander. Re-conduct the test in Question 2 using R commander. Make sure to include all the six components of a hypothesis test. Copy and paste the computer output first and then compare the

answer with the one you obtained by hand in Question 2. (5 marks)

2. Refer to Question 4 in Part A. Input the two-way table into R commander and conduct a chi-square test using R commander. Make sure to include all the six components of a hypothesis test. Copy and paste the computer output first and then Compare the answer with the one you obtained by hand in Question 4 in Part A. (6 marks)
3. Refer to Question 5 in Part A. The data are provided in the data file **M11_BloodPressure_Age_Chisquare_Q5.xlsx**. Import the data into R commander. Re-conduct the test in Question 5 using R commander. Make sure to include all the six components of a hypothesis test. Copy and paste the computer output first and then compare the answer with the one you obtained by hand in Question 2. (6 marks)

CHAPTER 12: ONE-WAY ANOVA

Overview

Two-sample t tests are used to compare two population means based on two independent samples. When comparing $k > 2$ population means based on k independent samples, one-way ANOVA can be used. ANOVA stands for **AN**alysis **Of** **VA**riance. This chapter introduces the main idea behind the one-way ANOVA F tests and how to conduct a one-way ANOVA F test based on computer output.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- State what ANOVA stands for.
- Identify when one-way ANOVA should be used.
- Explain the main idea behind a one-way ANOVA F test.
- Write down the hypotheses for a one-way ANOVA F test.
- Conduct a one-way ANOVA F test based on computer output.

12.1 Opening Example

A student experimented to compare download speed at different times of the day. He placed a file on a remote server and then proceeded to download the file at three different periods of the day: 7 a.m., 5 p.m., and 12 a.m. He downloaded the file 48 times, 16 times at each period and recorded the download time in seconds (De Veaux, Velleman, & Bock, 2008). Does the data below provide sufficient evidence of a difference between the mean download time at 7 a.m., 5 p.m., and 12 a.m.?

Table 12.1: Download Time at 7 a.m., 5 p.m., and 12 a.m.

Time of day	Time (sec)	Time of day	Time (sec)	Time of day	Time (sec)
7 a.m.	68	5 p.m.	299	12 a.m.	216
7 a.m.	138	5 p.m.	367	12 a.m.	175
7 a.m.	75	5 p.m.	331	12 a.m.	274
7 a.m.	186	5 p.m.	257	12 a.m.	171
7 a.m.	68	5 p.m.	260	12 a.m.	187
7 a.m.	217	5 p.m.	269	12 a.m.	213
7 a.m.	93	5 p.m.	251	12 a.m.	221
7 a.m.	90	5 p.m.	200	12 a.m.	139
7 a.m.	71	5 p.m.	296	12 a.m.	226
7 a.m.	154	5 p.m.	204	12 a.m.	128
7 a.m.	166	5 p.m.	190	12 a.m.	236
7 a.m.	130	5 p.m.	240	12 a.m.	128
7 a.m.	72	5 p.m.	350	12 a.m.	217
7 a.m.	81	5 p.m.	256	12 a.m.	196
7 a.m.	76	5 p.m.	282	12 a.m.	201
7 a.m.	129	5 p.m.	320	12 a.m.	161
	$\bar{x}_1 = 113.375$		$\bar{x}_2 = 273.250$		$\bar{x}_3 = 193.063$

12.2 Main Idea Behind One-Way ANOVA

Let $\mu_1, \mu_2, \dots, \mu_k$ be k population means. The hypotheses of one-way ANOVA are formulated as

H_0 : all means are equal, i.e., $\mu_1 = \mu_2 = \dots = \mu_k$

H_a : not all the means are equal.

The alternative hypothesis H_a means there exists at least one pair of means that are not equal. **Do not** write as $H_a : \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$. **Do not** write “at least one mean is different from the others” since it sounds like at least one mean is different from the others while all the others are the same. Both are just two special cases of what “not all the means are equal” means.

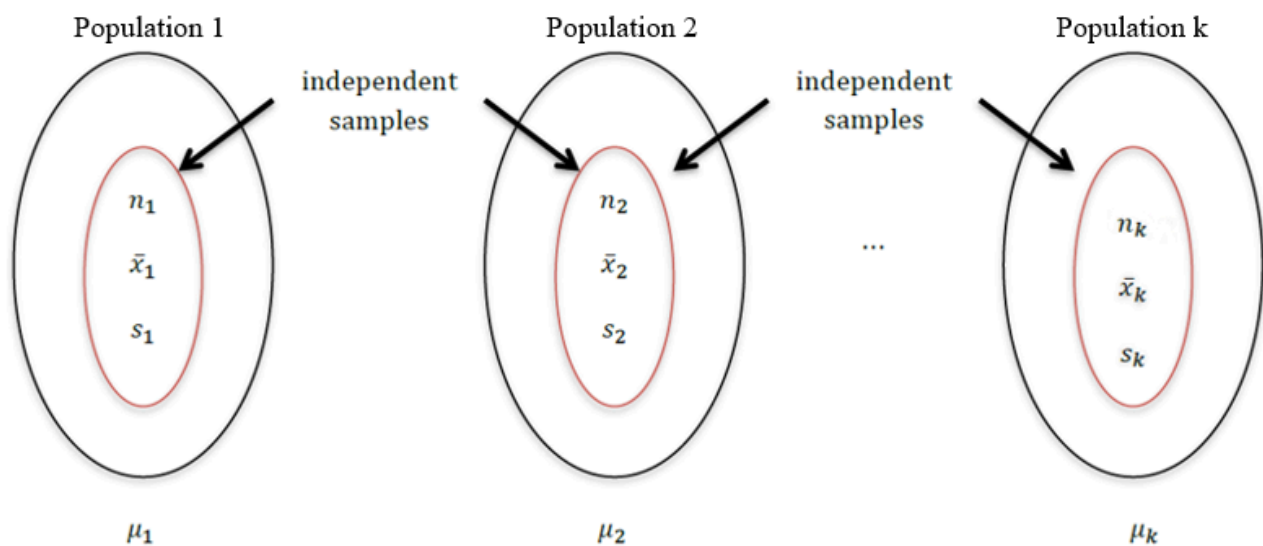


Figure 12.1: One-Way ANOVA Based on k Independent Samples. [[Image Description \(See Appendix D Figure 12.1\)\]](#)

ANOVA F tests are based on k independent, simple random samples from k populations. If $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ is true, the sample means x_1, x_2, \dots, x_k should be close to one another and hence, the variation among sample means should be small. Therefore, we should reject H_0 if the sample means are very different from one another (meaning the variation among the sample means would be large).

Quantifying Variation

The total variation of the data (SST: the total sum of squares) is quantified as the sum of squared distances from each observation to the overall mean, i.e., $SST = \sum (x_{ij} - \bar{x})^2$, where x_{ij} is the j th observation of sample i , $\bar{x} = \frac{\sum x_{ij}}{n}$ is the overall mean, and $n = n_1 + n_2 + \cdots + n_k$ is the overall sample size.

The treatment sum of squares quantifies the variation of the sample means:

$$SSTR = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2.$$

SSTR quantifies the so-called between-group variation. For this reason, we should reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ if SSTR is too large. However, SSTR is only considered “large” if it is large relative to the next measure of variation.

The within-group variation can be quantified as the sum of squared distances from each observation to the mean of its sample group, i.e.,

$$SSE = \sum (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1) s_i^2.$$

In practice, software is used to calculate all these sums of squares and other ANOVA calculations.

The total variation SST can be shown as:

$$SST = SSTR + SSE = \text{between group variation} + \text{within group variation}.$$

The relationship between SSTR (between group variation) and SSE (within-group variation) will be illustrated in the following example.

A person recorded waiting times each time he called either Uber or Taxi service from his house and again each time he called either service from work. His results are summarized in the following table (red values correspond to waiting times for Uber and blue values correspond to Taxi):

Table 12.2: Waiting Time for Uber (red) and Taxi (blue) Called from Home and Work

Home	1	2	3	3	4	5	6	7	8	8	9	10
Work	1	2	3	3	4	5	6	7	8	8	9	10

We define the waiting times called from Home as Data Set 1 and those called from Work as Data Set 2. The figure below shows two data sets. Each set consists of two populations:

waiting time for Uber (red circle with mean μ_1) and waiting time for Taxi (blue cross with mean μ_2).

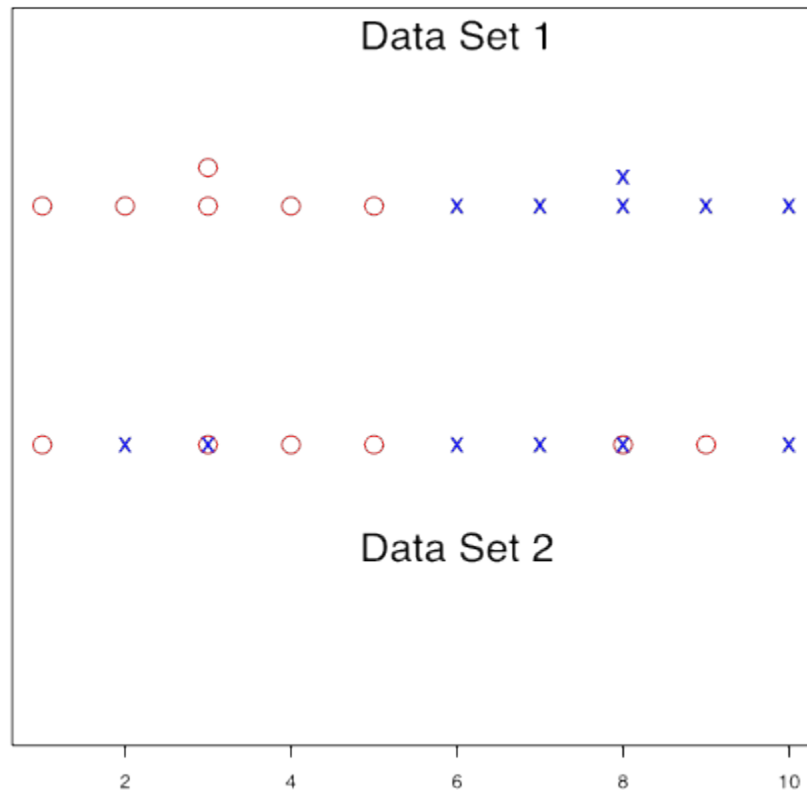


Figure 12.2: Waiting Time for Uber (red) and Taxi (blue) Called from Home (Data Set 1) and Work (Data Set 2). [[Image Description \(See Appendix D Figure 12.2\)](#)]



Activity

Exercise: Quantify Variation

Based on the two data sets shown above, answer these questions.

- The two data sets have _____(the same, different) total variation.
- Data set 1 has a _____(larger, smaller) within-group variation.
- Data set 1 has a _____(larger, smaller) between-group variation.

Support your answer by calculating the sums of squares SST , $SSTR$ and SSE for each of the two groups.

Show/Hide Answer

Answers:

- The two data sets have the same total variation.
- Data set 1 has a smaller within-group variation.
- Data set 1 has a larger between-group variation.

The two data sets have the same overall sample mean \bar{x} and the same total sum of square (SST):

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1+2+3+3+4+5+6+7+8+8+9+10}{12} = 5.5.$$

$$\begin{aligned} SST &= \sum (x_i - \bar{x})^2 \\ &= (1 - 5.5)^2 + (2 - 5.5)^2 + (3 - 5.5)^2 + \cdots + (8 - 5.5)^2 + (9 - 5.5)^2 + (10 - 5.5)^2 \\ &= 95. \end{aligned}$$

For Data Set 1, the mean waiting time for Uber is $\bar{x}_1 = \frac{1+2+3+3+4+5}{6} = 3$, and the mean waiting time for Taxi is $\bar{x}_2 = \frac{6+7+8+8+9+10}{6} = 8$. The between-group and within-group variation are:

$$SSTR = \sum n_i(\bar{x}_i - \bar{x})^2 = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 = 6(3 - 5.5)^2 + 6(8 - 5.5)^2 = 75.$$

$$\begin{aligned} SSE &= \sum (x_{ij} - \bar{x}_i)^2 \\ &= (1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 \\ &\quad + (6 - 8)^2 + (7 - 8)^2 + (8 - 8)^2 + (8 - 8)^2 + (9 - 8)^2 + (10 - 8)^2 \\ &= 10 + 10 = 20. \end{aligned}$$

For Data Set 2, the mean waiting time for Uber is $\bar{x}_1 = \frac{1+3+4+5+8+9}{6} = 5$, and the mean waiting time for Taxi is $\bar{x}_2 = \frac{2+3+6+7+8+10}{6} = 6$. The between-group and within-group variation are:

$$SSTR = \sum n_i(\bar{x}_i - \bar{x})^2 = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 = 6(5 - 5.5)^2 + 6(6 - 5.5)^2 = 3.$$

$$\begin{aligned} SSE &= \sum (x_{ij} - \bar{x}_i)^2 \\ &= (1 - 5)^2 + (3 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (8 - 5)^2 + (9 - 5)^2 \\ &\quad + (2 - 6)^2 + (3 - 6)^2 + (6 - 6)^2 + (7 - 6)^2 + (8 - 6)^2 + (10 - 6)^2 \\ &= 46 + 46 = 92. \end{aligned}$$

In both data sets, we have $SST = SSTR + SSE$. The two data sets have the same total variation $SST = 95$. Data Set 1 has a smaller within-group variation SSE (20 versus 92). Data Set 1 has a larger between-group variation $SSTR$ (75 versus 2).

For data set 1, it is clear that the data are from two populations with different means; for data set 2, however, it is hard to tell whether the data are from a single population or from two populations with similar means.

The main idea of one-way ANOVA is to decompose the total variation of the data (SST) into two parts: the variation within the samples (SSE) and the variation between sample means (SSTR). Reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ if the variation between sample means is large compared to the variation within samples. Or reject H_0 if the ratio $F = \frac{SSTR/(k-1)}{SSE/(n-k)} = \frac{MSTR}{MSE}$ is

too large, where MSTR is called the mean square of the treatments and MSE is the mean square error. The ratio follows an F distribution characterized by two degrees of freedom:

- The numerator degrees of freedom: $df_n = k - 1$,
- The denominator degrees of freedom: $df_d = n - k$.

Like chi-square tests, F tests are always right-tailed. That is both the rejection region and the p-value are upper-tailed probabilities.

12.3 F Distribution

In a two-sample t test, we use the t distribution to calculate the P -value and find the critical value. ANOVA relies on the F distribution whose density curve is unimodal, right skewed, and has two degrees of freedom. The ratio of MSTR over MSE follows an F distribution with degrees of freedom $df_n = k - 1$ and $df_d = n - k$.

The figure below shows the density curve of several F distributions with different degrees of freedom.

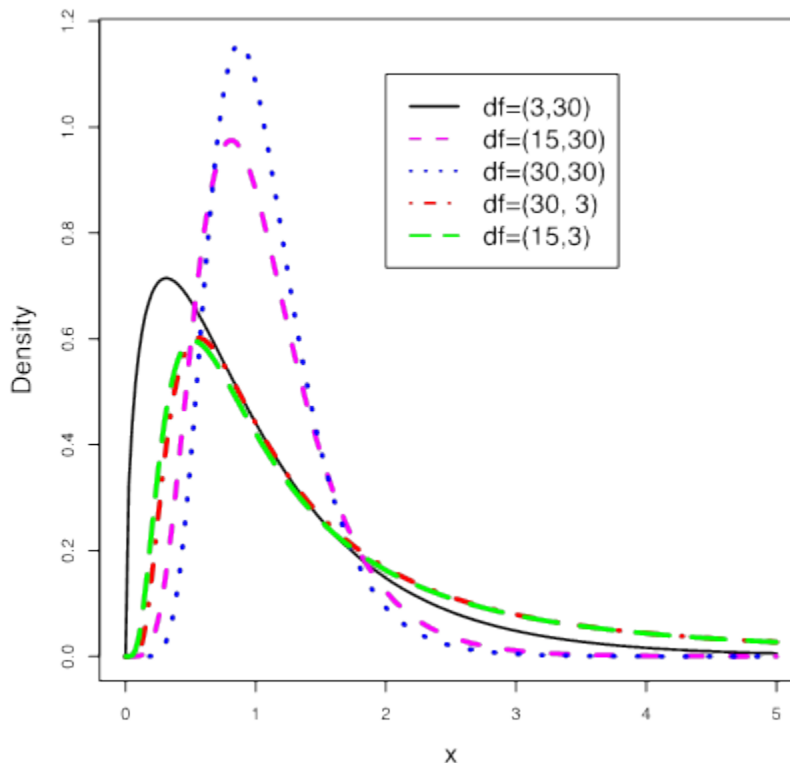


Figure 12.3: F Density Curves. [[Image Description \(See Appendix D Figure 12.3\)](#)]

The properties of F-density curves are as follows:

Key Fact: Properties of F Density Curve

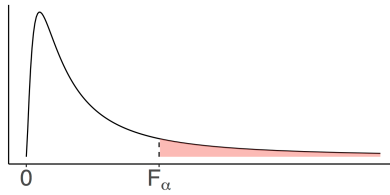
- Total area under the curve=1.
- The entire curve is above the horizontal axis.

- Right skewed.
- Horizontal axis spans from 0 to $+\infty$
- The two degrees of freedom df_n and df_d control the overall shape.

Similar to the t score table ([Table IV](#)), the F Table (see Table VI below) gives scores that correspond to right-tailed probabilities. For given numerator degrees of freedom (df_n) and denominator degrees of freedom (df_d), the F-score F_α is the value with an area α to its right. For example, for $df_n = 2$, $df_d = 3$, $F_{0.1} = 5.462$, $F_{0.01} = 30.82$, and $F_{0.005} = 49.8$.

Table 12.3: Part of the F Table (Table VI)

Table VI: Values of F_α of F -distribution



α	df_n									
	1	2	3	4	5	6	7	8	9	10
$df_d = 1$										
0.5	1.000	1.500	1.709	1.823	1.894	1.942	1.977	2.004	2.025	2.042
0.1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2
0.05	161	199	216	225	230	234	237	239	241	242
0.025	648	799	864	900	922	937	948	957	963	969
0.01	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
0.005	16211	19999	21615	22500	23056	23437	23715	23925	24091	24224
0.001	405284	499999	540379	562500	576405	585937	592873	598144	602284	605621
$df_d = 2$										
0.5	0.667	1.000	1.135	1.207	1.252	1.282	1.305	1.321	1.334	1.345
0.1	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
0.005	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40
0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40
$df_d = 3$										
0.5	0.585	0.881	1.000	1.063	1.102	1.129	1.148	1.163	1.174	1.183
0.1	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
0.05	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
0.005	55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7
0.001	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2

[\[Image Description \(See Appendix D Table 12.3\)\]](#)

12.4 One-Way ANOVA F Test

The assumptions and steps of a one-way ANOVA F test are as follows.

Assumptions:

- Normal populations: the variable of interest is normally distributed for each population.
- Equal variances: the variance of the variable of interest is the same for all populations.
- Independent samples: the samples from different populations are independent of one another.
- Simple random samples: the samples taken from the k populations are simple random samples.

Steps:

1. Set up the hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a : \text{Not all means are equal.}$$

2. State the significance level α .
3. Calculate the sums of squares SST, SSTR, SSE and the mean squares MSTR, MSE. Find the test statistic, F_o , and show the results in an ANOVA table:

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	p-value
Treatment	$k - 1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F_o = \frac{MSTR}{MSE}$	$P(F \geq F_o)$
Error	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n - 1$	SST			

4. Find the P-value **or** rejection region based on the F density curve with degrees of freedom $df_n = k - 1, df_d = n - k$.

P-value	$P(F \geq F_o)$ the area to the right of F_o under the curve
Rejection region	$F \geq F_\alpha$ the region to the right of the critical value F_α

5. Reject the null H_0 if P-value $\leq \alpha$ or F_o falls in the rejection region.
6. Conclusion.

The following ANOVA table corresponds to the download time example. Use the information in this ANOVA table to test at the 1% significance level whether there is a significant difference between the mean download times at 7 a.m., 5 p.m., and 12 a.m.

Table 12.4: ANOVA Table of Download Time Example

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	P-value
Time of day	2	204641	102320	$F_o = 46.03$	< 0.0001
Error	45	100020	2223		
Total	47	304661			

The side-by-side histograms and boxplots of all three groups can be found below. Both the side-by-side histograms and boxplots show that the downloading time at 7 AM, 5 PM and 12 AM is not normally distributed, since the histograms are not bell-shaped and the boxplots are not symmetric. The side-by-side boxplots show that the median downloading time of 7 AM, 5 PM, and 12 AM are around 90, 260, and 200 minutes respectively.

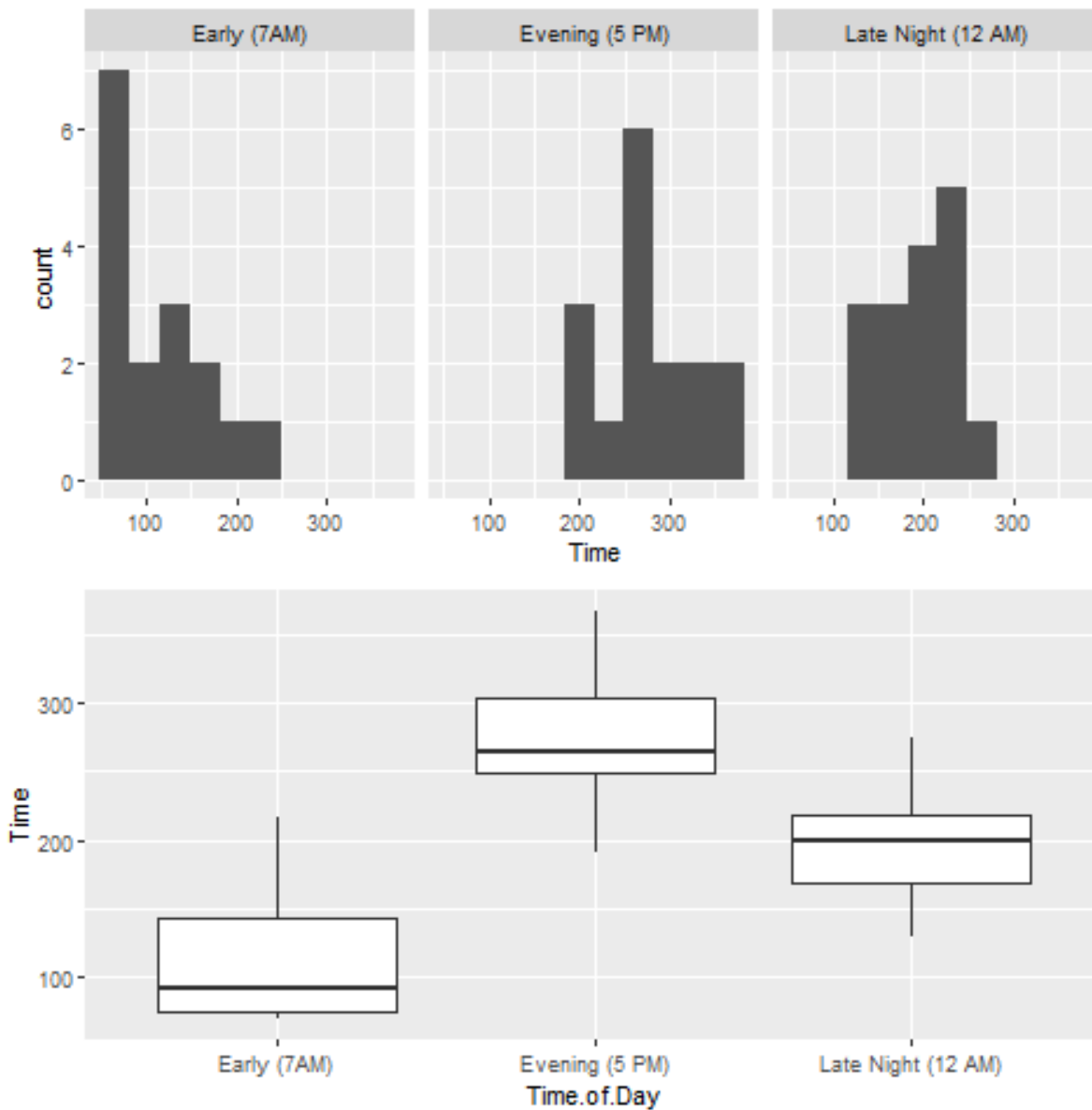


Figure 12.4: Side-by-side Histograms and Boxplots of Time of Day. [\[Image Description \(See Appendix D Figure 12.4\)\]](#)

Steps to conduct a one-way ANOVA F test:

1. Hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_a : Not all means are equal.

2. The significance level is $\alpha = 0.01$.
3. The test statistic is $F_o = 46.03$ with $df_n = k - 1 = 3 - 1 = 2$, $df_d = n - k = 48 - 3 = 45$.
4. P-value $P(F \geq F_o) = P(F \geq 46.03) < 0.0001$ (given in the ANOVA table)
5. Reject H_0 , since p-value $< 0.0001 < 0.01(\alpha)$.
6. Conclusion: At the 1% significance level, we have sufficient evidence that there is a significant

difference among the mean download times at 7 a.m., 5 p.m., and 12 a.m.



Activity

Exercises: One-Way ANOVA F Test

Many studies have suggested that there is an association between exercise and healthy bones. One study examined the effect of jumping on the bone density of growing rats. There are three treatments: a control with no jumping, a low-jump condition (the jump height was 30 centimetres), and a high-jump condition (60 centimetres). After eight weeks of 10 jumps per day, five days per week, the bone density of the rats in milligrams per cubic centimetre (mg/cm^3) was measured. The data are given in the following table.

Table 12.5: Bone Density for Three Treatments

Group	Bone density	Group	Bone density	Group	Bone density
Control	611	Low jump	635	High jump	650
Control	621	Low jump	605	High jump	622
Control	614	Low jump	638	High jump	626
Control	593	Low jump	594	High jump	626
Control	593	Low jump	699	High jump	631
Control	653	Low jump	632	High jump	622
Control	600	Low jump	631	High jump	643
Control	554	Low jump	588	High jump	674
Control	603	Low jump	607	High jump	643
Control	569	Low jump	596	High jump	650

The side-by-side histograms and boxplots are shown as follows:

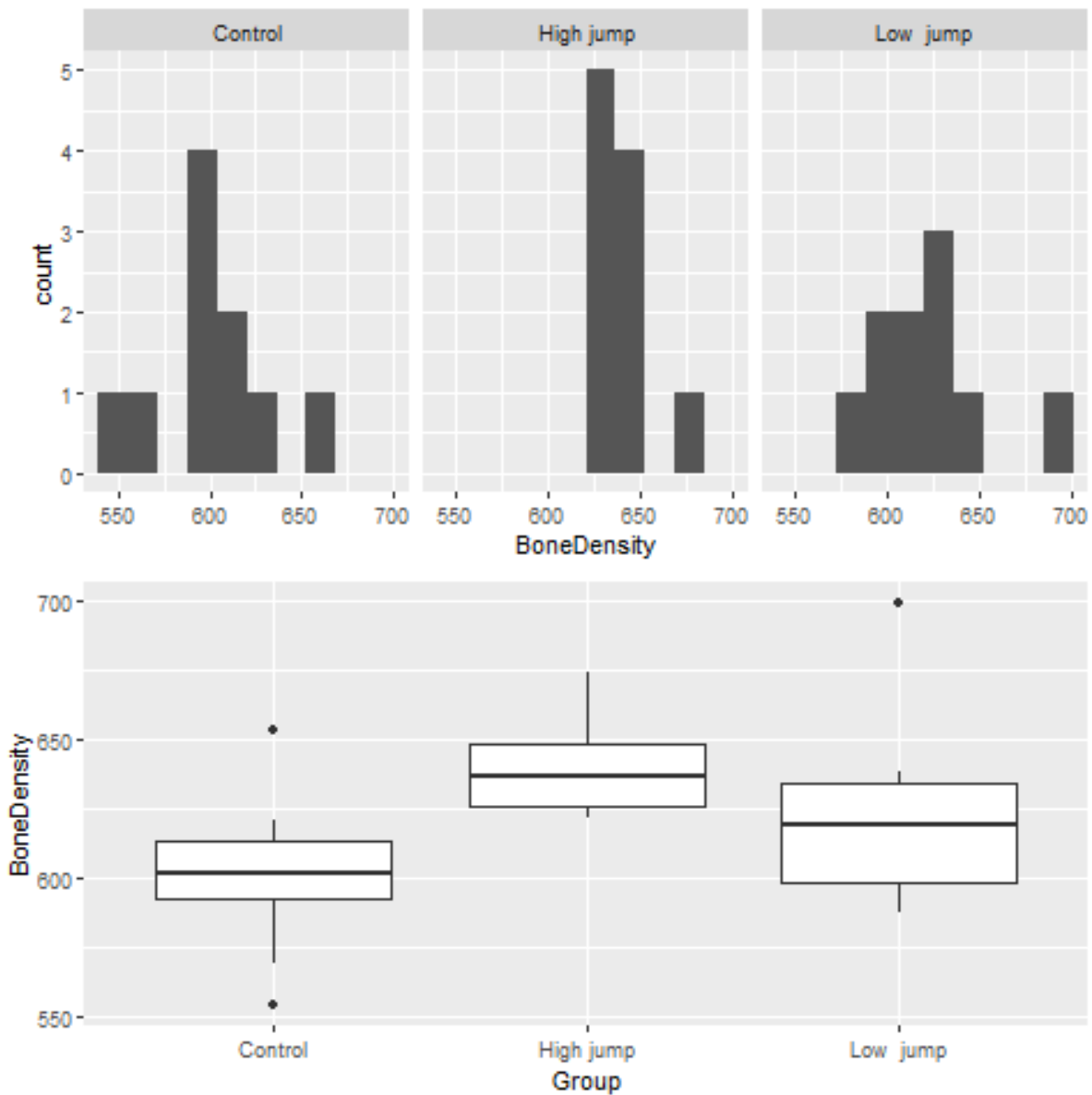


Figure 12.5: Side-by-side Histograms and Boxplots of Groups. [Image Description (See Appendix D Figure 12.5)]

Given the ANOVA table, test at the 5% significance level whether jumping strengthens the bones of rats.

Table 12.6: ANOVA Table of Bone Density Exercise

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	P-value
Group	2	7114	3557	$F_o = 5.08$	0.0133
Error	27	18879	699		
Total	29	25993			

Show/Hide Answer

Answers:

Steps to conduct a one-way ANOVA F test:

1. Hypotheses
 $H_0 : \mu_1 = \mu_2 = \mu_3$
 H_a : Not all means are equal.
2. The significance level is $\alpha = 0.05$.
3. The test statistic is $F_o = 5.08$ with $df_n = k - 1 = 3 - 1 = 2$, $df_d = n - k = 30 - 3 = 27$.
4. P-value = $P(F \geq F_o) = P(F \geq 5.08) < 0.0133$ (given in the ANOVA table)
5. Reject H_0 , since p-value = $0.0133 < 0.05(\alpha)$.
6. Conclusion: At the 5% significance level, we have sufficient evidence that the mean bone densities are different in the three treatment groups. Since the rats in the “high jump” group has the largest mean bone density, followed by the “low jump” group, we can conclude that jumping strengthens the bones of rats.

12.5 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- State what ANOVA stands for (Chapter 12 Introduction).
- Identify when one-way ANOVA should be used (Section 12.2).
- Explain the main idea behind a one-way ANOVA F test (Section 12.4).
- Write down the hypotheses for a one-way ANOVA F test (Section 12.4).
- Conduct a one-way ANOVA F test based on computer output (Section 12.4).

12.6 Review Questions

- The data on monthly rents, in dollars, for independent random samples of newly completed apartments in the four U.S. regions are presented in the following table.

Northeast	Midwest	South	West
1005	870	891	1025
898	748	630	1012
948	699	861	1090
1181	814	1036	926
1244	721		1269
	606		

Given the ANOVA table, test at the 5% significance level whether a difference exists in the mean rent of newly completed apartments in the four U.S. regions.

Source	<i>df</i>	<i>SS</i>	$MS = \frac{SS}{df}$	F-statistic	P-value
Region	3	$SSTR = 400513$	$MSTR = 133504$	$F_o = 7.541$	0.0023
Error	16	$SSE = 283265$	$MSE = 17704$		
Total	19	$SST = 683778$			

- The following table gives the salaries (in thousand dollars) for computer science (CS) majors obtaining a bachelor's degree, a master's degree, or a Ph.D.

Bachelor	Master	PhD
50.8	65.8	73.3
59.4	57.5	65.7
55.9	66.9	71.7
45.1	62.8	72.5
54.1	68.5	73
50.7	69.3	67.2
46.8	61.5	67.5

a. Fill in missing entries of the following ANOVA table.

Source	df	SS	MS = $\frac{SS}{df}$	F-statistic	P-value
Group	$k-1 = ?$	$SSTR = ?$	$MSTR = \frac{SSTR}{k-1} = 616.9$	$F_o = \frac{MSTR}{MSE} = 34.62$	<0.0001
Error	$n-k = ?$	$SSE = ?$	$MSE = \frac{SSE}{n-k} = 17.8$		
Total	$n-1 = 20$	$SST = 1554.5$			

b. Test at the 1% significance level whether a difference exists in mean salary for computer science (CS) majors obtaining a bachelor's degree, a master's degree, or a Ph.D.

Show/Hide Answer

1. From the first table, we have

$$k = 4, n_1 = 5, n_2 = 6, n_3 = 4, n_4 = 5 \longrightarrow n = n_1 + n_2 + n_3 + n_4 = 20.$$

We assume the assumptions of the one-way ANOVA F test are satisfied. Steps of one-way ANOVA F test:

Step 1: Hypotheses.

H_0 : all means are equal, $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus H_a : not all means are equal, i.e., at least one pair of means are different.

Step 2: Significance level $\alpha = 0.05$.

Step 3: Test statistic $F_o=7.541$ with degrees of freedom

$$df_{TR} = k - 1 = 4 - 1 = 3, df_E = n - k = 20 - 4 = 16.$$

Step 4: P-value=0.0023.

Step 5: Decision. Reject H_0 since P-value=0.0023 < 0.01 (α).

Step 6: Conclusion. At the 5% significance level, we have sufficient evidence that a difference exists in the mean rent of newly completed apartments in the four U.S. regions.

2.

a.	Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	P-value
	Group	k-1=?	SSTR=?	$MSTR = \frac{SSTR}{k-1} = 616.9$	$F_o = \frac{MSTR}{MSE} = 34.62$	<0.0001
	Error	n-k=?	SSE=?	$MSE = \frac{SSE}{n-k} = 17.8$		
	Total	n-1=20	SST=1554.5			

From the first table, we have $k = 3, n_1 = n_2 = n_3 = 7 \implies n = n_1 + n_2 + n_3 = 21$.

From the second table, we have

$$df_T = n - 1 = 19, SST = 1554.5, MSTR = 616.0, MSE = 17.8.$$

$$(1) df_{TR} = k - 1 = 3 - 1 = 2.$$

$$(2) df_E = n - k = 21 - 3 = 18 \text{ or } df_E = df_T - df_{TR} = 19 - 2 = 17.$$

$$(3) SSTR = MSTR \times df_{TR} = 616.9 \times 2 = 1233.8.$$

$$(4) SSE = MSE \times df_E = 17.8 \times 18 = 320.4 \text{ or}$$

$$SSE = SST - SSTR = 1554.5 - 1233.8 = 320.7. \text{ The difference is due to rounding.}$$

b. We assume the assumptions of the one-way ANOVA F test are satisfied.

Steps of one-way ANOVA F test:

Step 1: hypotheses. H_0 : all means are equal, $\mu_1 = \mu_2 = \mu_3$ versus H_a : not all means are equal, i.e., at least one pair of means are different.

Step 2: Significance level $\alpha = 0.01$.

Step 3: Test statistic $F_o = 34.62$ with degrees of freedom

$$df_{TR} = k - 1 = 3 - 1 = 2, df_E = n - k = 21 - 3 = 18.$$

Step 4: P-value < 0.0001.

Step 5: Decision. Reject H_0 since P-value < 0.0001 < 0.01 (α).

Step 6: Conclusion. At the 1% significance level, we have sufficient evidence that a difference exists in mean salary for computer science (CS) majors obtaining a bachelor's degree, a master's degree or a Ph.D.

12.7 Assignment 12

Purposes

The following questions have two parts. The first part assesses your knowledge of identifying cases where one-way ANOVA should be used, explaining the main idea of one-way ANOVA, and conducting a one-way ANOVA F test based on the computer outputs. The second part assesses your skills in using R Commander to conduct a one-way ANOVA F test.

Resources

[M12_Rent_ANOVA_Q5_FourColumnS.xlsx](#)

[M12_Salary_ANOVA_Q7.xlsx](#)

[M12_Rent_ANOVA_Q6.xlsx](#)

Instructions

Part A

Complete the following:

1. What does ANOVA stand for? (2 marks)
2. When shall we use a one-way ANOVA F test? (2 marks)
3. Suppose that a one-way ANOVA is being performed to compare the means of three populations and that the sample sizes are 10, 12, and 15. Determine the degrees of freedom for the F statistic. (2 marks)
4. We stated earlier that a one-way ANOVA test is always right-tailed because the null hypothesis is rejected only when the test statistic, F , is too large. Why is the null hypothesis rejected only when F is too large? (3 marks)
5. Fill in the missing entries in the following ANOVA table. (4 marks)

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic
Treatment	2	?	21.652	$F_o = ?$
Error	?	84.400		
Total	14	?		

6. The data on monthly rents, in dollars, for independent random samples of newly completed apartments in the four U.S. regions are presented in the following table. (See the spreadsheet **M12_Rent_ANOVA_Q5_FourColumnS.xlsx**)

Northeast	Midwest	South	West
1005	870	891	1025
898	748	630	1012
948	699	861	1090
1181	814	1036	926
1244	721		1269
	606		

Given the ANOVA table, test at a 5% significance level whether a difference exists in the mean rent of newly completed apartments in the four U.S. regions. (8 marks)

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	P-value
Region	3	400513	133504	$F_o = 7.541$	0.0023
Error	16	283265	17704		
Total	19	683778			

7. The following table gives the salaries (in thousand dollars) for computer science (CS) majors obtaining a bachelor's degree, a master's degree, or a Ph.D. (See the spreadsheet **M12_Salary_ANOVA_Q7.xlsx**)

Salary	Degree	Salary	Degree	Salary	Degree
50.8	Bachelor's	65.8	Master's	73.3	Ph.D
59.4	Bachelor's	57.5	Master's	65.7	Ph.D
55.9	Bachelor's	66.9	Master's	71.7	Ph.D
45.1	Bachelor's	62.8	Master's	72.5	Ph.D
54.1	Bachelor's	68.5	Master's	73.0	Ph.D
50.7	Bachelor's	69.3	Master's	67.2	Ph.D
46.8	Bachelor's	61.5	Master's	67.5	Ph.D

- a. Fill in missing entries of the following ANOVA table. (4 marks)

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	P-value
Degree	?	?	616.9	$F_o = 34.62$	< 0.0001
Error	?	?	17.8		
Total	20	1554.5			

- b. Test at the 1% significance level whether a difference exists in mean salary for computer science (CS) majors obtaining a bachelor's degree, a master's degree or a Ph.D. (8 marks)

Part B

Finish the following questions using R and R commander.

- Refer to Question 6 in Part A. The data are provided in the file **M12_Rent_ANOVA_Q6.xlsx**. Import the data into R commander. Regenerate the ANOVA table in Question 6 in Part A. Make sure you copy and paste the computer output. (3 marks)
- Refer to Question 7 in Part A. The data are provided in the file **M12_Salary_ANOVA_Q7.xlsx**. Import the data into R commander. Obtain the ANOVA table and compare it with Question 7 in Part A. Make sure you copy and paste the computer output. (4 marks)

Quiz 11



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2636#h5p-14>

CHAPTER 13: DESCRIPTIVE AND INFERENTIAL METHODS IN SIMPLE LINEAR REGRESSION

Overview

This chapter introduces simple linear regression, which models the relationship between two quantitative variables using a straight line; we discuss descriptive and inferential methods in simple linear regression.

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Identify situations where simple linear regression should be used.
- Explain the main idea of the method of least squares.
- Calculate the least-squares fitted line.
- Calculate and interpret the correlation coefficient r .
- Calculate and interpret the coefficient of determination r^2 .
- Explain the terms in a simple linear regression model.
- Conduct a t test and obtain a t confidence interval for the slope parameter β_1 .
- Explain the difference between confidence intervals and prediction intervals.
- Obtain a confidence interval for the conditional mean and a prediction interval for a single response.

13.1 Introduction

The following table and scatter plot show the relationship between the price (in \$1,000) and the age (in years) of 15 used cars of a particular make and model.

Table 13.1: Age and Price of Used Cars

Age (x, in year)	Price (y, in \$1000)
1	14
1	13
3	13
4	10
4	10
5	9
5	9
6	7
7	7
7	8
8	7
8	6
10	5
10	4
13	3

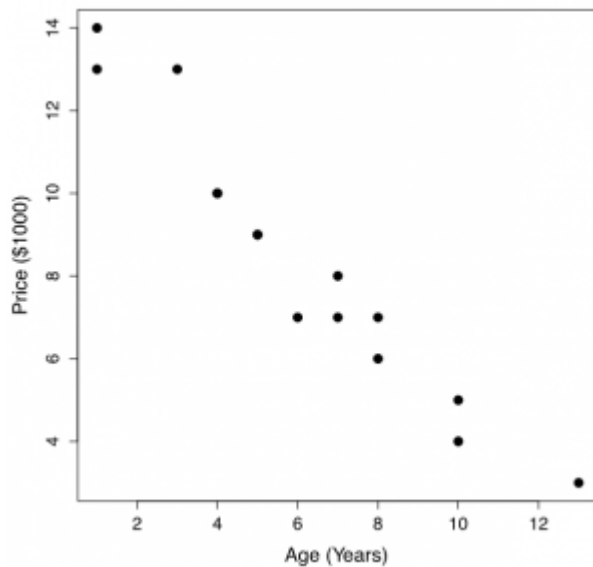


Figure 13.1: Scatter Plot of Price V.S. Age of 15 Used Cars. [\[Image Description \(See Appendix D Figure 13.1\)\]](#) Click on the image to enlarge it.



Activity

Exercise: Can We Use a Simple Linear Regression?

1. Could we use a straight line to model the relationship between the price and age of the used cars?
2. How does the price of used cars change when the age increases?

Show/Hide Answer

Answers:

1. Yes, since the points are roughly on a straight line.
2. The price of used cars tends to decrease as age increases.

13.2 Least-Squares Straight Line

We use a straight line to model the relationship between two quantitative variables y and x : $y = b_0 + b_1x$. The interpretations of the terms in the equation are given as follows:

- x : the *predictor* (independent) variable
- y : the *response* (dependent) variable
- b_0 : the intercept, it is the value of y when $x = 0$
- b_1 : the slope of the straight line. It is **the change in y when x increases by 1 unit**. If $b_1 > 0$, y increases when x increases; if $b_1 < 0$, y decreases when x increases.

The figure below illustrates the meanings of the intercept and the slope of a straight line.

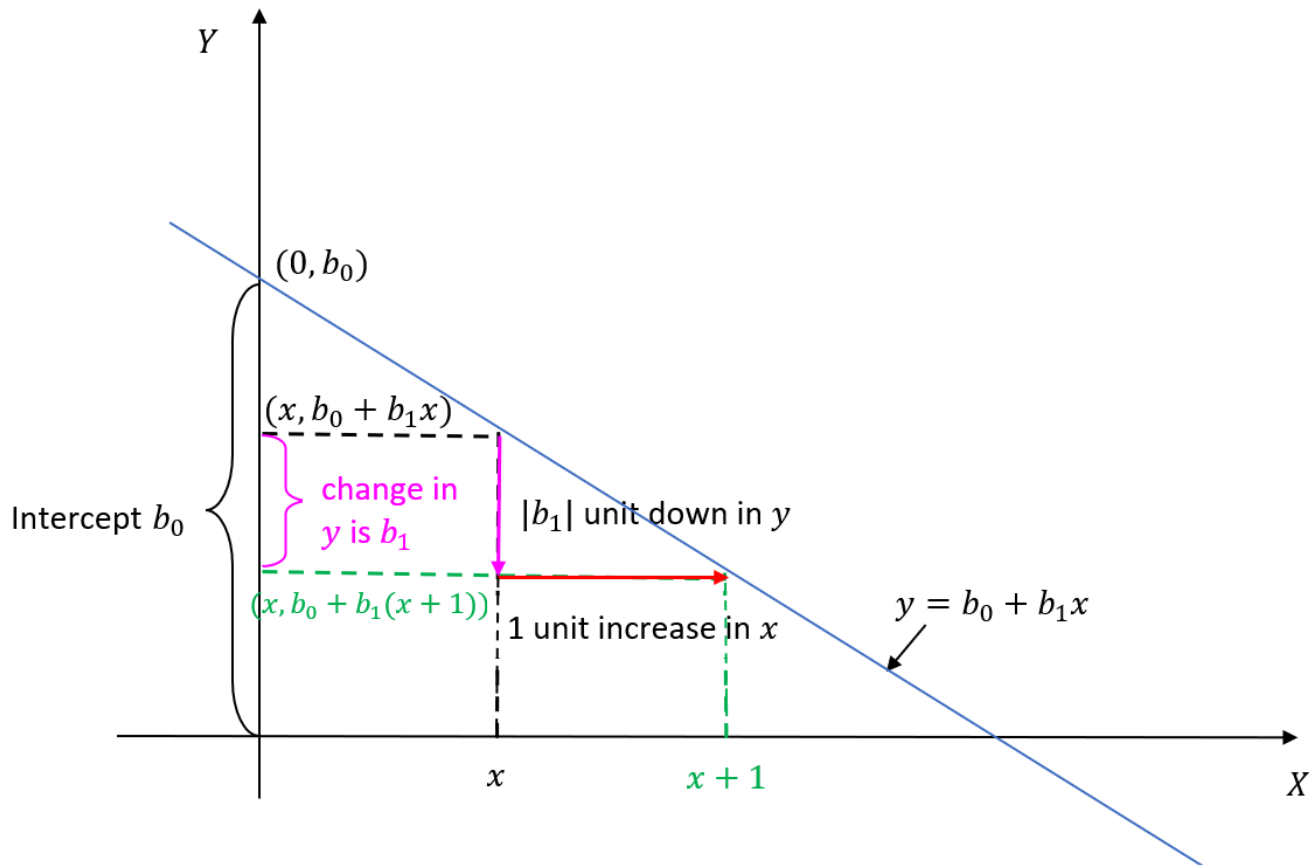


Figure 13.2: Interpretation of Intercept and Slope of a Straight Line. [[Image Description \(See Appendix D Figure 13.2\)](#)]

Our first objective is to determine the values of b_0 and b_1 that characterize the line of best fit: $\hat{y} = b_0 + b_1x$. To properly quantify what is meant by “best fit”, we introduce some definitions. The **fitted values** are $\hat{y}_i = b_0 + b_1x_i$, where x_i is the observed x -value

corresponding to y_i , the observed y -value, for $i = 1, 2, \dots, n$. Each **residual** is defined as the difference between the observed y value and the fitted value. That is, the i th residual is:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i).$$

The **least-squares regression line** is obtained by finding the values of b_0 and b_1 that minimize the residual sum of squares $SSE = \sum e_i^2 = \sum [y_i - (b_0 + b_1 x_i)]^2$. The figure below illustrates the least-squares regression line as the red line.

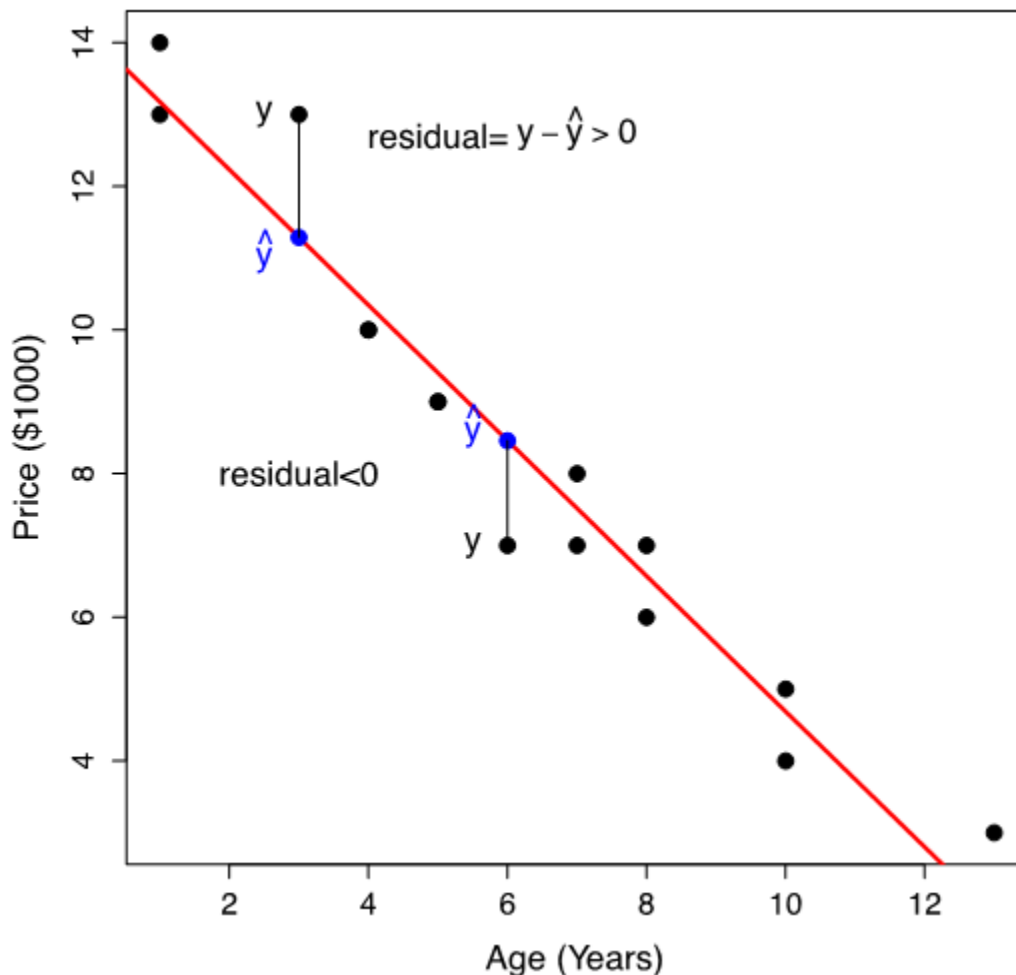


Figure 13.3: Residuals and Fitted Least-Squares Regression Line. [[Image Description \(See Appendix D Figure 13.3\)](#)]

Note that:

- Some residuals are positive, and some are negative. Note that $\sum e_i = \sum (y_i - \hat{y}_i) = 0$.
- We want a straight line closest to the data points, i.e., the total distance from the points to the line is minimized.
- We use the square of the residual e_i^2 to quantify the distance from the data point y_i to the straight line.

- The total error is the sum of the squared distances from each point to the straight line, i.e., $\sum e_i^2$.
- The straight line yielding the smallest $\sum e_i^2$ is called the least-squares line since it makes the sum of squares of the residuals the smallest.

To find the values of b_0 and b_1 that minimize the residual sum of squares

$$SSE = \sum e_i^2 = \sum [y_i - (b_0 + b_1 x_i)]^2$$

is an optimization problem. It can be shown that the solutions are

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}},$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n}.$$

Like ANOVA, the least-squares regression equation can be obtained using software in practice.

Example: Least-Squares Regression Line

Given the summaries of the 15 used cars,

$$n = 15, \sum x_i = 92, \sum x_i^2 = 724, \sum y_i = 125, \sum y_i^2 = 1193, \sum x_i y_i = 616$$

- Find the least-squares regression line to model the relationship between the used cars' price (y) and age (x).

Steps:

- Calculate the sum of squares:

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 616 - \frac{92 \times 125}{15} = -150.667,$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 724 - \frac{92^2}{15} = 159.733.$$

- Find the slope and intercept:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{-150.667}{159.733} = -0.9432,$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \frac{125}{15} - (-0.9432) \times \frac{92}{15} = 14.118.$$

Therefore, the least-squares regression line for the used cars is

$$\hat{y} = b_0 + b_1x \implies \widehat{\text{price}} = 14.118 + (-0.9432) \times \text{age} = 14.118 - 0.9432 \times \text{age}.$$

- b. Interpret the slope $b_1 = -0.9432$ (in \$1000).

On average, the price of used cars drops by \$943.2 when they get one year older.

13.3 Prediction and Extrapolation

We can use the least-squares regression line $\hat{y} = b_0 + b_1x$ to predict the value of the response variable y given the value of the predictor variable x . For example, using the least-squares straight line from the previous exercise, the predicted price of a car that is 2 years old (age=2) is: $\widehat{\text{price}} = 14.118 - 0.9432 \times 2 = 12.2316$ (\$1,000, see figure below), or \$12,231.6. The predicted price for a 10-year-old car is: $\widehat{\text{price}} = 14.118 - 0.9432 \times 10 = 4.686$ (\$1,000, see figure below), which means the price of a 10-year-old car is predicted as \$4,686.

When making a prediction, avoid **extrapolation**, in which y -values are predicted using x -values that are outside of the range of the observed x -values. For example, if we use the least-squares regression line $\widehat{\text{price}} = 14.118 - 0.9432 \times \text{age}$ to predict the price of a 20-year-old car, our estimated price is $\widehat{\text{price}} = 14.118 - 0.9432 \times 20 = -4.746$ (\$1,000, see figure below). It does not make sense for an individual to pay \$4,746 if he/she wants to sell a 20-year-old car; this is the consequence of extrapolation. This regression line was developed with used cars between 1 and 13 years old; age=20 is outside this range, and we should not use the fitted least-squares line to predict the price of a 20-year-old car. Another example of extrapolation is to predict the height of an adult based on their weight using a regression line (regress height on weight) fitted on the data of children under 10 years old.

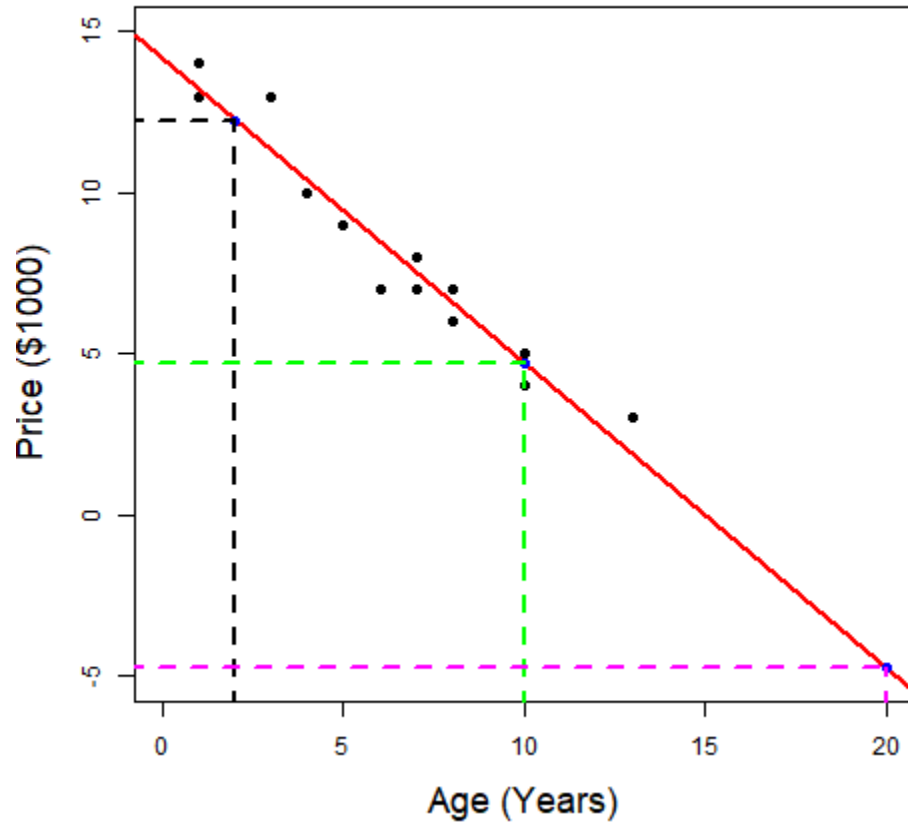


Figure 13.4: Prediction for Age=2, 10, 20 and Extrapolation. [[Image Description \(See Appendix D Figure 13.4.\)](#)]

13.4 Outliers and Influential Observations

In simple linear regression, we must also watch out for outliers and influential observations. **Outliers** are observations that are far away from the majority of the data. An **influential observation** is a data point that changes the regression equation dramatically if included. Note that an outlier might or might not be an influential observation.

Example: Outlier and Influential Observations

In the following figures, identify whether the red point is an outlier or an influential observation.

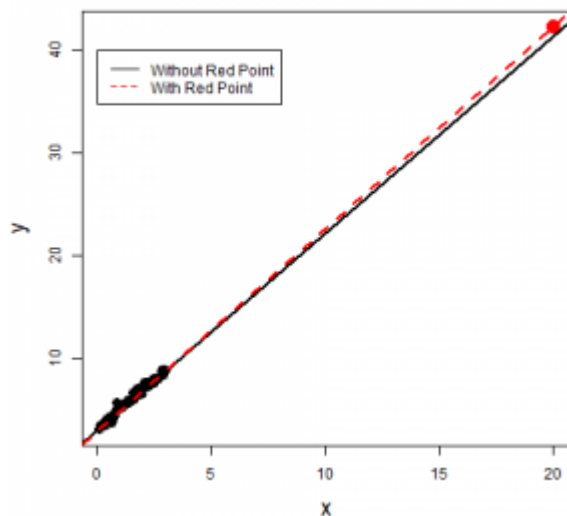


Figure 13.5: An Outlier But Not Influential. [[Image Description](#) [\(See Appendix D Figure 13.5\)](#)] Click on the image to enlarge it.

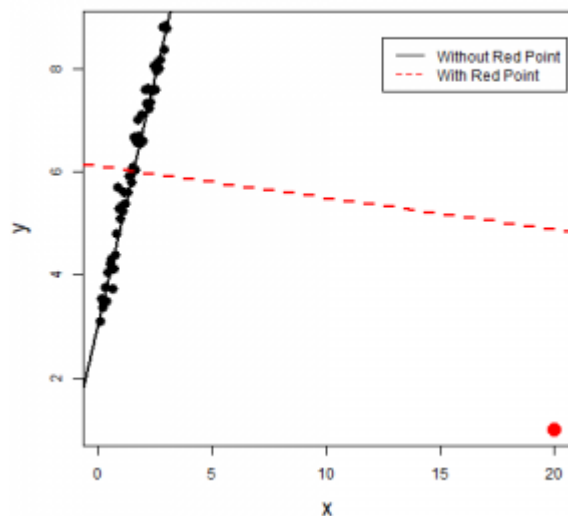


Figure 13.6: An Outlier and Influential [[Image Description](#) [\(See Appendix D Figure 13.6\)](#)] Click on the image to enlarge.

The red point on the left panel is an outlier since it is far away from the majority of the data; however, it is not an influential observation since the regression lines are almost identical with and without the red point.

The red point on the right panel is an outlier and an influential observation since including the red point dramatically changes the regression line. Without the red point, the slope of the regression line is

positive; the slope becomes negative when the red observation is included. The red observation is also far away from the majority of the data and hence is an outlier.

13.5 Correlation Coefficient r

The correlation coefficient r is calculated by

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$$

where

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}.$$

The correlation coefficient r measures the association between the response variable y and the predictor variable x in the following three aspects:

- Pattern: The correlation coefficient r measures **linear** association. Do NOT use the correlation coefficient r to describe non-linear association.
- Strength: The closer r is to either +1 or -1, the stronger the linear association. When $r = \pm 1$, y and x have a perfect linear association. That is, all the data points in the scatter plot of x versus y fall in a straight line.
- Direction: Positive or negative. Positive association ($r > 0$) means that y and x change in the same direction. That is y increases (decreases) if x increases (decreases). Negative association ($r < 0$) means that y and x change in the opposite direction, that is, y increases (decreases) if x decreases (increases).



Activity

Exercise: Correlation Coefficient

Match the following correlation coefficients with the scatter plots.

(1) 0.989 (2) 0.697 (3) -0.887 (4) -0.020

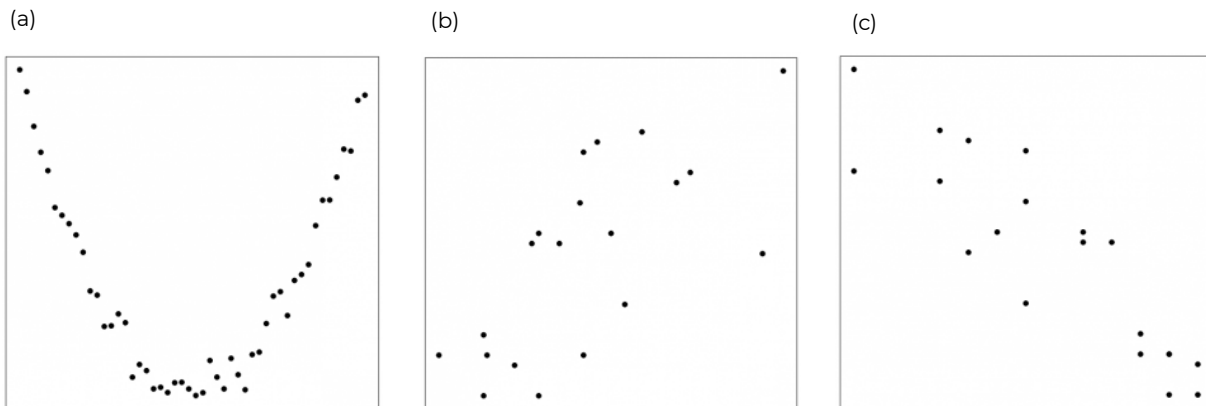


Figure 13:7: Match Correlation Coefficients and Scatter Plots. [[Image Description \(See Appendix D Figure 13.7\)](#)]

Show/Hide Answer

Answers:

- a. $r = -0.020$. There is no linear association between y and x , r should be close to 0.
- b. $r = 0.697$. When x increases, y also increases. There should be a positive linear association between y and x , i.e., $r > 0$. But it is not extremely strong since the points show little semblance of a straight line.
- c. $r = -0.887$. When x increases, y decreases. There should be a negative linear association between y and x , i.e., $r < 0$. The association is quite strong since the points are starting to resemble the rough appearance of a straight line.
- d. $r = 0.989$. When x increases, y also increases. There should be a positive linear association between y and x , i.e., $r > 0$. The association is extremely strong since the points are basically on a straight line.

Exercise: Concepts on Correlation Coefficient

Explain whether the following statements are true or false. Correct them if they are false.

1. If $r \approx 0$, there is no association between y and x .

2. The larger the value of r , the stronger the association between y and x .

Show/Hide Answer

1. False. If $r \approx 0$, there is no **linear** association between y and x .
2. False. The larger the **absolute** value of r , the stronger the **linear** association. Or the closer r is to +1 or -1, the stronger the **linear** association.

13.6 The Coefficient of Determination

Continuing the used car example, observe that the prices of the 15 cars are different, so we can quantify the variation among prices y_i by looking at the distance between each observation and the mean, i.e., $(y_i - \bar{y})^2$. Therefore, the total variation of prices is calculated by $SST = \sum_{i=1}^n (y_i - \bar{y})^2$. The term SST is called the **total sum of squares**. The prices of the 15 used cars are different; part of the reason is that their ages are different, so that means “age” explains some of the variation in “price”.

It can be shown that the total variation in the response y (SST , which is the total sum of squares) can be decomposed into two parts: variation explained by predictor variable x through the regression equation (SSR , which is the regression sum of squares) and the variation not explained by x (SSE , which is the error sum of squares), i.e.,

$$\begin{aligned} SST &= \sum (y_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \\ &= SSE + SSR. \end{aligned}$$

The sums of squares are similar to ANOVA, and have a similar decomposition. The distance from y_i to \bar{y} is composed of two parts: the distance from y_i to \hat{y}_i and the distance from \hat{y}_i to \bar{y} . The difference between y_i and \hat{y}_i is the residual $e_i = y_i - \hat{y}_i$. Then we have $SST = SSE + SSR$ with

$$SST = \sum (y_i - \bar{y})^2 = S_{yy}, SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2, SSR = \sum (\hat{y}_i - \bar{y})^2 = r^2 S_{yy} = \frac{S_{xy}^2}{S_{xx}}$$

The square of the correlation coefficient, $R^2 = r^2$, is called the **coefficient of determination**. It can be shown that

$$r^2 = \frac{S_{xy}^2}{S_{xx} \times S_{yy}} = \frac{SSR}{SST}.$$

The coefficient of determination r^2 indicates the percentage of total variation (SST) in the observed response variable that is explained by the predictor variable x through the regression equation (SSR).

Example: Correlation Coefficient and Coefficient of Determination

Given the summaries of the 15 used cars,

$$n = 15, \sum x_i = 92, \sum x_i^2 = 724, \sum y_i = 125, \sum y_i^2 = 1193, \sum x_i y_i = 616.$$

- a. Calculate and interpret the correlation coefficient r .

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 616 - \frac{92 \times 125}{15} = -150.667,$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 724 - \frac{92^2}{15} = 159.733,$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1193 - \frac{125^2}{15} = 151.333,$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{-150.667}{\sqrt{159.733 \times 151.333}} = -0.9691.$$

Interpretation: There is a **strong, negative, linear** association between the price and the age of the used cars.

- b. Calculate and interpret the coefficient of determination r^2 .

$$r^2 = (-0.9691)^2 = 0.9391.$$

Interpretation: 93.91% of the **variation in the observed price** of the used cars is due to the age of the used cars. That is, 93.91% of the variation in the price of the used cars can be explained by the age of the used cars through the regression equation.

Note: A common **mistake** in interpreting r^2 is that 93.91% of the points lie on a straight line.

Note: the geometric interpretation of r^2 is NOT required for STAT 151; the remaining material of this section is only for students interested in explaining the decomposition.

The following figure shows the geometric interpretation of the decomposition $SST = SSR + SSE$, i.e.,

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$$

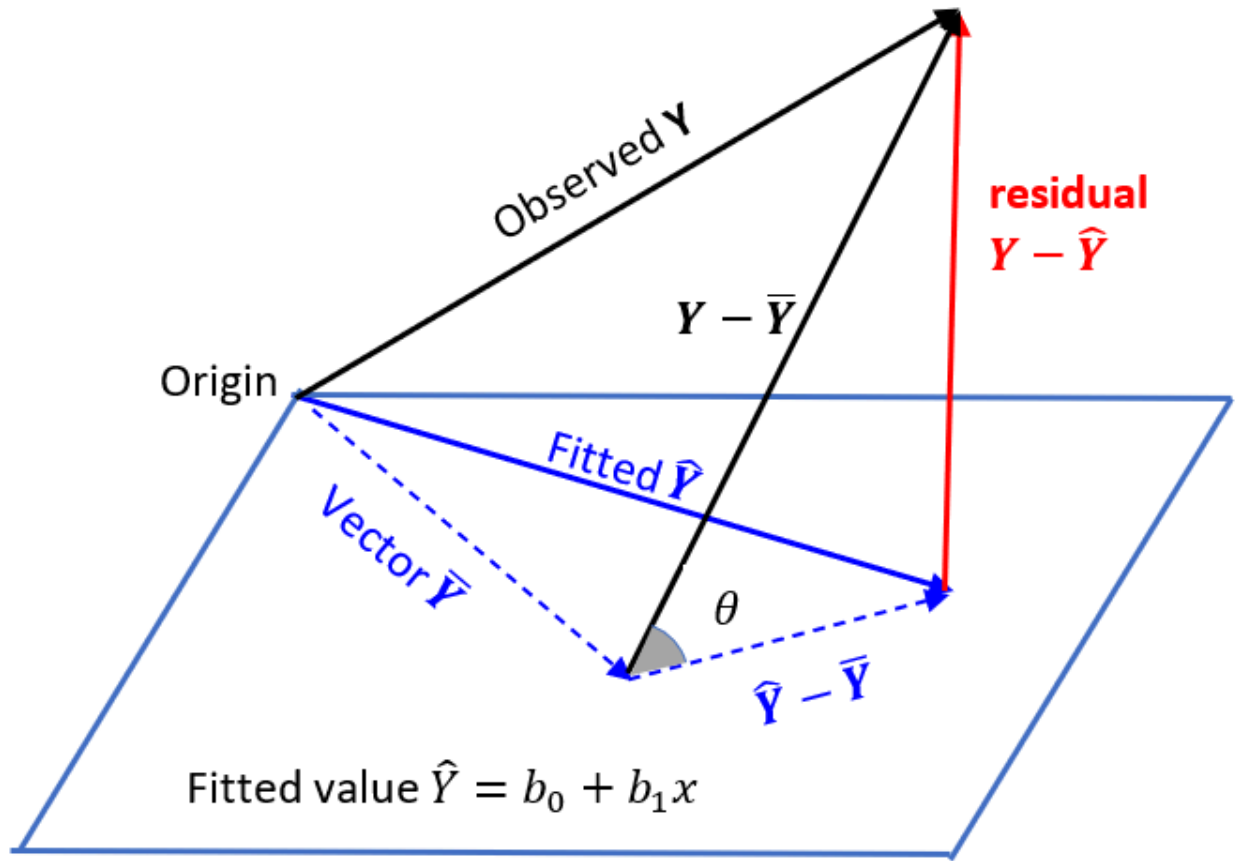


Figure 13.8: Geometric interpretation of r^2 [Image Description (See Appendix D Figure 13.8)]

Suppose there are n observations $\{x_i, Y_i\}_{i=1}^n$ in a regression problem. The responses $(Y_1, Y_2, \dots, Y_n)^T$ forms an $n \times 1$ vector \mathbf{Y} . The fitted values $(\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)^T$ is an orthogonal projection of the observed \mathbf{Y} onto the plane of the predictor variable $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$. Since the residual vector $\mathbf{Y} - \hat{\mathbf{Y}}$ is orthogonal to the plane, the triangle formed with edges $\mathbf{Y} - \bar{\mathbf{Y}}$, $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$, and $\mathbf{Y} - \hat{\mathbf{Y}}$ is a right triangle. Therefore,

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2.$$

Let θ be the angle between the two vectors $\mathbf{Y} - \bar{\mathbf{Y}}$ and $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$. It can be shown that

$$r^2 = \cos^2 \theta = \frac{\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2} = \frac{SSR}{SST}.$$

13.7 Simple Linear Regression Model (SLRM)

So far, we have focused on a sample of 15 used cars. If we would like to draw conclusions about the population (all used cars), we need some inferential methods. First, we introduce some concepts from conditional probability. Consider, for example, all five-year-old used cars; their prices follow a certain distribution. We call this distribution the **conditional distribution** of price given age is equal to 5, and the corresponding mean and standard deviation are referred to as the **conditional mean** and **conditional standard deviation**. For variables Y and X , the conditional mean and standard deviation of Y given $X = x$ are denoted with $\mu_{Y|x}$ and $\sigma_{Y|x}$. Now, our objective is to model $\mu_{Y|x}$ and the predictor variable x using a straight line (see the figure below). That is

$$\mu_{Y|x} = \beta_0 + \beta_1 x,$$

where β_0 and β_1 are the population intercept and population slope, respectively, with interpretation as follows:

- β_0 : the average of the response variable y when $x = 0$.
- β_1 : the change in the **mean value of y** when the predictor variable x **increases by 1 unit**.

For each given value of the predictor variable x , the conditional distribution of the response variable Y given x is assumed to follow a normal distribution with a mean $\mu_{Y|x} = \beta_0 + \beta_1 x$ and a standard deviation $\sigma_{y|x} = \sigma$ (that is, the conditional standard deviation of Y is assumed constant for all values of x). For example, most five-year-old cars are not equal in price to the conditional mean $\mu_{y|x=5}$; σ measures the typical difference between the conditional mean $\mu_{y|x=5}$ and the actual value of any given five-year-old car. To capture this variation (sampling error), we introduce an error term ϵ into the model. The regression model becomes

$$Y = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim N(0, \sigma).$$

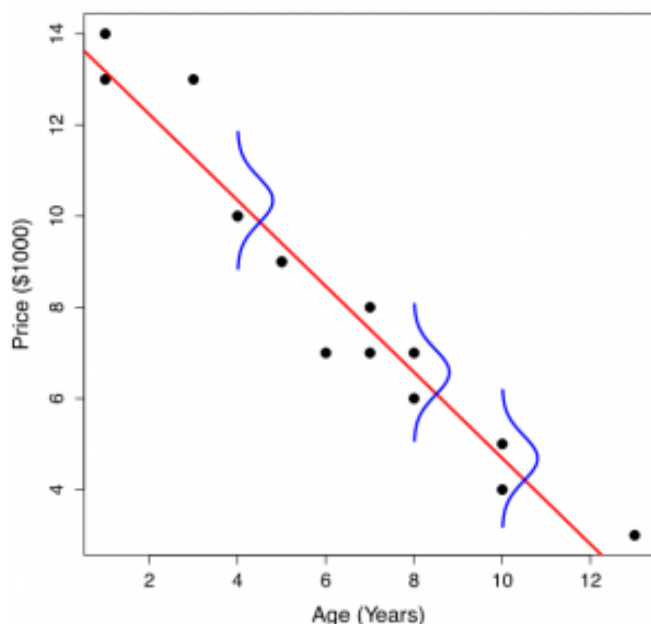


Figure 13.9: Simple Linear Regression Model. [\[Image Description \(See Appendix D Figure 13.9\)\]](#) Click on the image to enlarge it.

- This figure models the relationship between the conditional mean $\mu_{Y|x}$ (the mean of the response variable Y) and the predictor variable X using a straight line, i.e., $\mu_{Y|x} = \beta_0 + \beta_1 x$.
- In general, a single y value is not equal to the conditional mean $\mu_{Y|x}$, so we use an error term ϵ to quantify the deviation of the y values from the conditional mean $\mu_{Y|x}$. The model assumes that ϵ has a mean of 0 since the conditional mean of Y given x is assumed to be $\mu_{Y|x}$.
- Therefore, the simple linear regression model becomes

$$Y = \mu_{Y|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim N(0, \sigma).$$

Key Facts: Assumptions for Inferential Methods in Simple Linear Regression

In general, we need the following **assumptions** to apply the inferential methods in simple linear regression:

- **Linearity assumption:** There is a linear association between the conditional mean and the predictor variable. This assumption can be checked by plotting a scatter plot (Y against X). If this assumption is met, all the points should roughly lie on a straight line.
- **Normal population:** Given each value of the predictor variable x , the conditional distribution of the response variable is normally distributed. That is, $Y|x \sim N(\beta_0 + \beta_1 x, \sigma)$ or equivalently, the error term ϵ follows a normal distribution with mean 0 and standard deviation σ for all values of x .
- **Equal standard deviation:** The conditional standard deviation of the response variable is the same for all values of the predictor variable X . This common standard deviation is denoted as σ . This is equivalent to assuming that the error term ϵ has the same standard deviation for all values of x .
- **Independent observations:** The observations of the response variable are independent of one another. This is hard to check unless we know how the data were collected.

Checking whether the model assumptions are satisfied or not can be performed by **residual analysis**. Recall that the residual of the i th observation, e_i , is defined as the difference between the observed value (y_i) and the fitted value (\hat{y}), i.e., $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x_i)$. If the model assumptions are satisfied, the residuals $e_i, i = 1, 2, \dots, n$ can be regarded as a simple random sample from a normal distribution with mean 0 and standard deviation σ ; this means the residuals should follow an approximate normal distribution with standard deviations that are similar for each value of x . To check the assumptions, we draw two graphs of the residuals:

- Normal population assumption: Draw a Q-Q (normal probability) plot on the residuals. If the normality assumption is met, the points should roughly fall on a straight line.
- Equal standard deviation assumption and linearity assumption: Plot the residuals e_i (y -axis) versus fitted values $\hat{y}_i = b_0 + b_1x_i$ (x -axis). We can also plot the residuals e_i (y -axis) versus the x_i values (x -axis). If the equal standard deviation assumption is met, we will see a horizontal band centred and symmetric about the line $y = 0$. If the points form some curvature, the linearity assumption is violated.

Example: Residual Analysis

Comment on the following residual plots (first row) and normal probability plots (second row) of three sets of residuals.

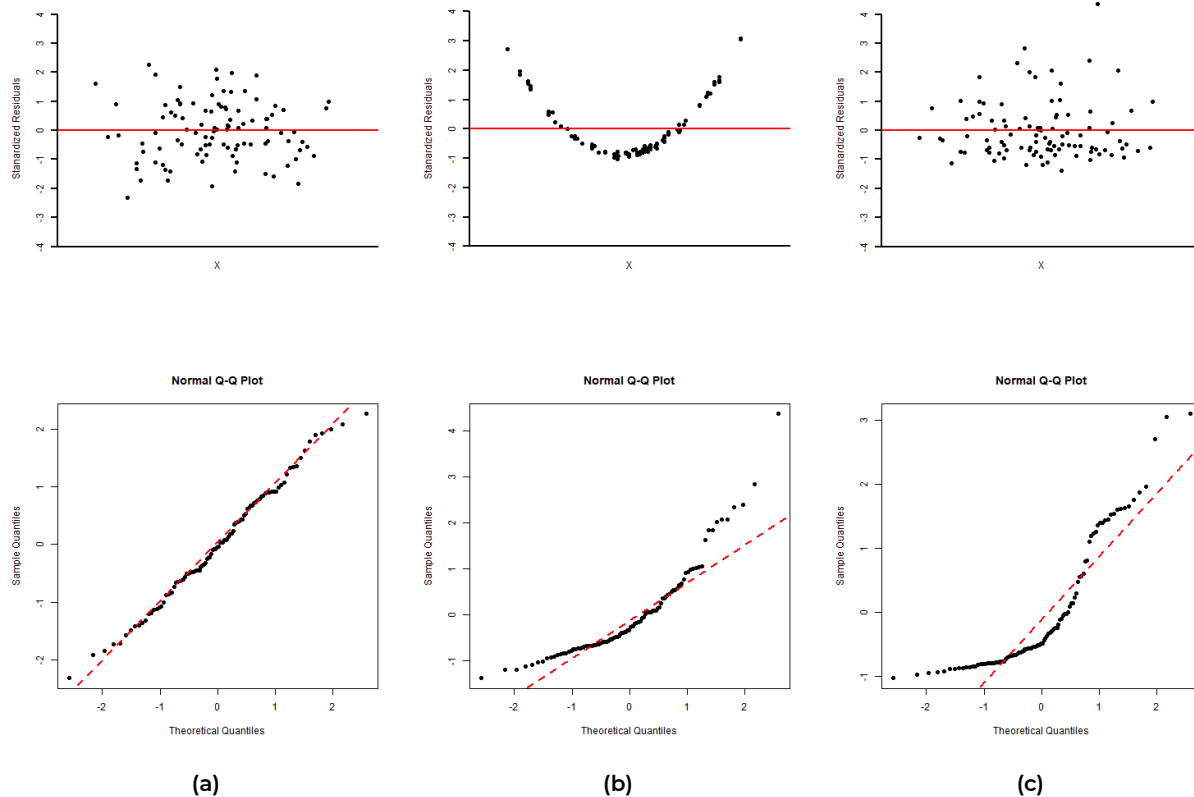


Figure 13.10: Residual Plots and Normal Probability Plots of Three Sets of Residuals. [\[Image Description \(See Appendix D Figure 13.10\)\]](#) Click on the image to enlarge it.

- a. Based on the residuals plot, both the equal standard deviation and the linearity assumptions appear met since the points form a horizontal band centred at 0. According to the normal probability (Q-Q) plot, points are roughly on a straight line and hence the normality assumption is satisfied.
- b. Based on the residual plot, the linearity assumption appears violated since the residual plot shows “U” shaped curvature. The normal probability plot provides evidence against the normality assumption.
- c. Based on the residuals plot, the equal standard deviation assumption appears violated since the residual plot does not show a horizontal band. The variation of the residuals gets larger when x increases. The normal probability plot shows that the normality assumption is violated.



Activity

Exercise: Residual Analysis for the Used Cars

The residuals for the 15 used cars are given in the following table.

- a. Given that the least-squares regression line is $\widehat{\text{price}} = 14.118 - 0.9432 \times \text{age}$. Verify the residual for the first car.

Age (x, in yr)	Price (y, in \$1,000)	\hat{y} ($\widehat{\text{price}} = 14.118 - 0.9432 \times \text{age}$)	Residual $e_i = y - \hat{y}$
1	14	13.1748	0.8252
1	13	13.1748	-0.1748
3	13	11.2884	1.7116
4	10	10.3452	-0.3452
4	10	10.3452	-0.3452
5	9	9.4020	-0.4020
5	9	9.4020	-0.4020
6	7	8.4588	-1.4588
7	7	7.5156	-0.5156
7	8	7.5156	0.4844
8	7	6.5724	0.4276
8	6	6.5724	-0.5724
10	5	4.6860	0.3140
10	4	4.6860	-0.6860
13	3	1.8564	1.1436

Table 13.2: Fitted Value and Residuals of 15 Used Cars. [\[Image Description \(See Appendix D Table 13.2\)\]](#)

- b. Comment on the following graphs based on the residuals of the 15 used cars. Explain whether the assumptions for regression inference are met or not.

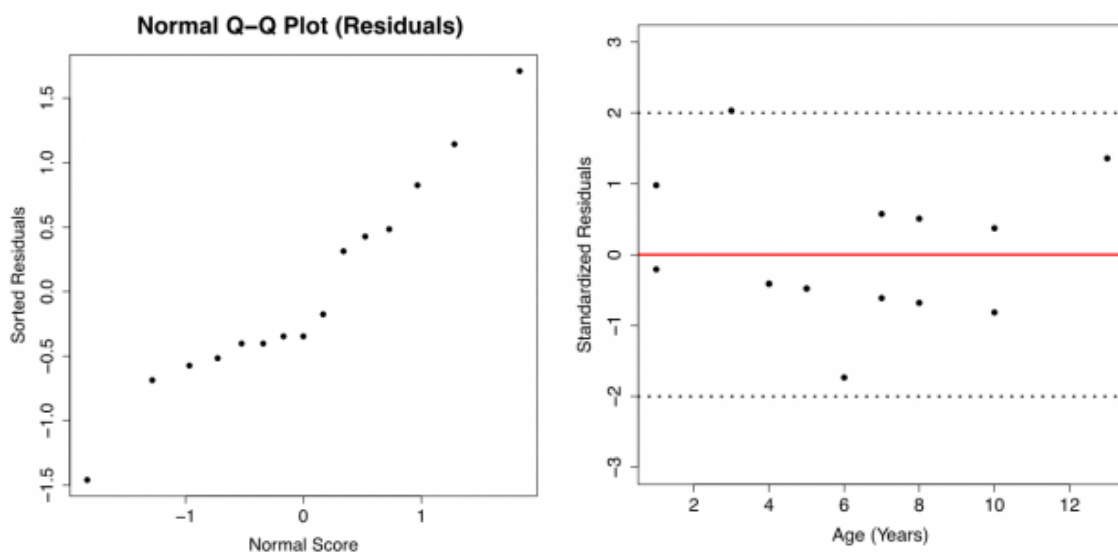


Figure 13.11: Normal Probability Plot of Residuals (Left) and Scatter Plot of Residuals V.S. Predictor Variable (Right) [\[Image Description \(See Appendix D Figure 13.11\)\]](#)

Show/Hide Answer

Answers:

- a. Given that the least-squares regression line is $\widehat{\text{price}} = 14.118 - 0.9432 \times \text{age}$, the fitted value for the first used car with age = 1 is $\hat{y}_1 = 14.118 - 0.9432 \times 1 = 13.1748$, and the residual is $e_1 = y_1 - \hat{y}_1 = 14 - 13.1748 = 0.8252$ (\$1000), which is the same as the residual given in the table.
- b. The plot in the left panel is the normal probability plot (normal QQ plot) for the residuals. The points in the QQ plot of the residuals are roughly on a straight line. Therefore, the normal population assumption appears to be met. The plot in the right panel is the scatterplot of residuals (y-axis) versus the predictor variable (x-axis). The points are roughly within a horizontal band centred at 0, without any obvious curvature. Therefore, both the linearity and the equal standard deviation assumptions appear to be met.

13.8 Inferences for the Parameters in SLRM

There are three population parameters in the simple regression model (SLRM): the population intercept β_0 , the population slope β_1 , and the standard deviation of the error σ . The three population parameters can be estimated by the least-squares estimates:

- $b_0 = \bar{y} - b_1\bar{x}$ estimates the population intercept β_0 ;
- $b_1 = \frac{S_{xy}}{S_{xx}}$ estimates the population slope β_1 ;
- the sample standard deviation of the residuals $s_e = \sqrt{\frac{\sum(e_i - \bar{e})^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$ estimates the standard deviation of the error term σ , where
$$SSE = SST - SSR = SST - r^2SST = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy} - b_1S_{xy}.$$

We are especially interested in testing whether the slope parameter β_1 differs from 0. If $\beta_1 = 0$, the predictor variable x provides no information about the conditional mean of Y , and hence there is no point fitting a regression model. Inferences about β_1 are based on the distribution of the least-squares slope b_1 .

13.8.1 Distribution of the Least-squares Slope b_1

In the previous example, a least-squares regression line was developed to predict the price of used cars with their ages; using a sample of 15 used cars, the fitted line had a slope of $b_1 = -0.9432$. Does $b_1 = -0.9432$ provide evidence of a linear association between the price and age of all used cars? In order to answer this question, we need a better understanding of the distribution of b_1 . Suppose, for example, that we repeat this experiment 1000 times by obtaining 1000 samples of 15 used cars, fitting 1000 regression lines, and as such, getting 1000 different values of b_1 , each of which is a point estimate of the population slope β_1 . The distribution of these 1000 values of b_1 , therefore, provides an estimate of the true distribution of all such b_1 . The following histogram illustrates the distribution of 1000 values of b_1 .

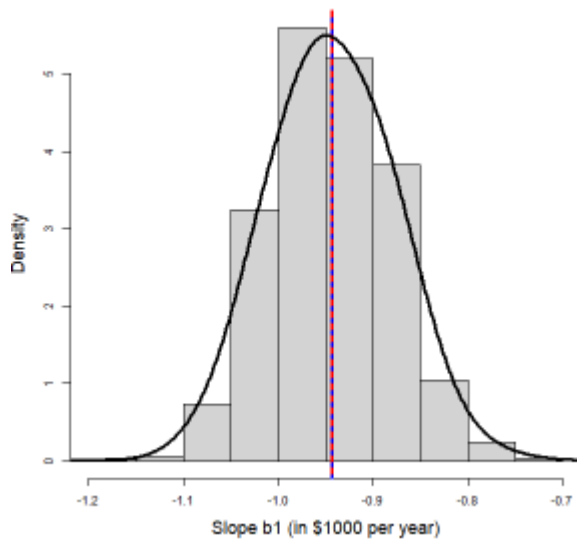


Figure 13.12: Distribution of the Least-Squares Slope b_1 . [[Image Description \(See Appendix D Figure 13.12\)](#)] Click on the image to enlarge it.

It can be shown that the distribution of b_1 is normal with

- Mean: $\mu_{b_1} = \beta_1$.
- Standard deviation: $\sigma_{b_1} = \frac{\sigma}{\sqrt{S_{xx}}}$.

That is $b_1 \sim N(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}})$. Thus, we can standardize b_1 in order to obtain: $\frac{b_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0, 1)$.

When the standard deviation of the error term σ is unknown, it can be estimated by the standard deviation of the residuals s_e . This leads to the studentized variable

$$\frac{b_1 - \beta_1}{\frac{s_e}{\sqrt{S_{xx}}}} \sim t \text{ distribution with } df = n - 2.$$

Hence, inferences about the slope parameter β_1 are based on a t distribution with degrees of freedom $n - 2$. Note that we lose two degrees of freedom in finding b_0 and b_1 .

13.8.2 t Test and t Interval for the Slope Parameter β_1

Assumptions:

1. The response variable (or the error term) is normally distributed.
2. The standard deviation of the response variable (or error term) is the same for all values of the predictor variable.
3. The observations are independent.
4. The data come from a simple random sample.

Steps to perform a test on the slope parameter β_1 :

1. Set up the hypotheses:

The predictor is useful	positive association	negative association
$H_0 : \beta_1 = 0$	$H_0 : \beta_1 \leq 0$	$H_0 : \beta_1 \geq 0$
$H_a : \beta_1 \neq 0$	$H_a : \beta_1 > 0$	$H_a : \beta_1 < 0$

2. State the significance level α .
3. Compute the test statistic: $t_o = \frac{b_1}{\frac{s_e}{\sqrt{S_{xx}}}}$ with degree of freedom $df = n - 2$.
4. Find the P-value **or** rejection region

	The predictor is useful	positive association	negative association
Null	$H_0 : \beta_1 = 0$	$H_0 : \beta_1 \leq 0$	$H_0 : \beta_1 \geq 0$
Alternative	$H_a : \beta_1 \neq 0$	$H_a : \beta_1 > 0$	$H_a : \beta_1 < 0$
P-value	$2P(t \geq t_o)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_{\alpha}$	$t \leq -t_{\alpha}$

5. Reject the null H_0 if P-value $\leq \alpha$ or t_o falls in the rejection region.
6. Conclusion.

The $(1 - \alpha) \times 100\%$ t confidence interval for β_1 corresponding to the t test:

	The predictor is useful	positive association	negative association
Null	$H_0 : \beta_1 = 0$	$H_0 : \beta_1 \leq 0$	$H_0 : \beta_1 \geq 0$
Alternative	$H_a : \beta_1 \neq 0$	$H_a : \beta_1 > 0$	$H_a : \beta_1 < 0$
CI	$\left(b_1 - t_{\alpha/2} \frac{s_e}{\sqrt{S_{xx}}}, b_1 + t_{\alpha/2} \frac{s_e}{\sqrt{S_{xx}}}\right)$	$(b_1 - t_{\alpha} \frac{s_e}{\sqrt{S_{xx}}}, \infty)$	$(-\infty, b_1 + t_{\alpha} \frac{s_e}{\sqrt{S_{xx}}})$
Decision	Reject H_0 if the interval does not contain 0.		

Example: t-Test and t Interval for the Slope Parameter β_1

Recall the used car example. We have the summaries

$$n = 15, \sum x_i = 92, \sum x_i^2 = 724, \sum y_i = 125, \sum y_i^2 = 1193, \sum x_i y_i = 616.$$

We can calculate:

$$\begin{aligned}
S_{xy} &= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 616 - \frac{92 \times 125}{15} = -150.667, \\
S_{xx} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 724 - \frac{92^2}{15} = 159.733, \\
S_{yy} &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1193 - \frac{125^2}{15} = 151.333. \\
b_1 &= \frac{S_{xy}}{S_{xx}} = \frac{-150.667}{159.733} = -0.9432; \\
b_0 &= \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \frac{125}{15} - (-0.9432) \frac{92}{15} = 14.118.
\end{aligned}$$

And the least-square straight line is $\widehat{\text{price}} = 14.118 - 0.9432 \times \text{age}$.

- a. **Test at the 5% significance level whether age is a useful predictor for the price of a used car.**

Steps:

1. Set up the hypotheses. $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.
2. The significance level is $\alpha = 0.05$.
3. Compute the value of the test statistic: $t_o = \frac{b_1}{\frac{s_e}{\sqrt{S_{xx}}}}$ with $df = n - 2$. First,

$$SSE = S_{yy} - b_1 S_{xy} = 151.333 - (-0.9432) \times (-150.667) = 9.224 \text{ so that}$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{9.224}{13}} = 0.842. \text{ Therefore,}$$

$$t_o = \frac{b_1}{\frac{s_e}{\sqrt{S_{xx}}}} = \frac{-0.9432}{\left(\frac{0.842}{\sqrt{159.733}}\right)} = -14.158, df = n - 2 = 15 - 2 = 13.$$

4. Find the P-value. For a two-tailed test with $df = 13$,
P-value = $2P(t \geq |t_o|) = 2P(t \geq 14.158) < 2 \times 0.0005 = 0.001$, since $t_{0.0005} = 4.221$.
5. Decision: Reject the null H_0 since P-value $< 0.001 < 0.05(\alpha)$.
6. Conclusion. At the 5% significance level, we have sufficient evidence that age is a **useful predictor** of the price of a used car.

- b. **Obtain a t confidence interval for the slope parameter β_1 corresponding to the test in part (a).**

A 95% two-tailed interval corresponds to a two-tailed test at the 5% significance level. Therefore, since

$$df = 13, \alpha = 0.05, \text{ and } t_{\alpha/2} = t_{0.025} = 2.160$$

It follows that a 95% confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} \frac{s_e}{\sqrt{S_{xx}}} = (-0.9432) \pm 2.160 \times \frac{0.842}{\sqrt{159.733}} = (-1.087, -0.799).$$

- c. **Interpret the interval. Does it support the conclusions of the hypothesis test in part (a)?**

We are 95% confident that β_1 is somewhere between -1.087 and -0.799 (\$1000 per year). Hence, we estimate that the **mean price** of used cars drops from \$799 to \$1087 when they get one year older.

Yes, it supports the conclusions of the hypothesis test in part (a). The interval does not contain 0; which implies $\beta_1 \neq 0$ with 95% confidence. Therefore, the interval suggests age is a useful predictor for the price of used cars, which is the conclusion of the test in part (a).

13.9 Confidence Intervals and Prediction Intervals

Suppose the model assumptions are satisfied and the regression model fits the data. In that case, we can use the fitted least-squares regression model to estimate the conditional mean μ_p and a single value of the response variable y_p , corresponding to an observed value of the predictor variable x_p . For example, we can predict the **mean price** of all used cars that are seven years old and we can use this predicted value to construct a **confidence interval** for the mean price of all used cars of this age. We can also predict the price of a single used car that is seven years old and construct a so-called **prediction interval** for the price of a single used car of this age.

Given a fixed value of the predictor variable, $x = x_p$, a point estimate for the conditional mean $\mu_p = \beta_0 + \beta_1 x_p$ is given by $\hat{\mu}_p = b_0 + b_1 x_p$. It can be shown that the conditional mean $\hat{\mu}_p$ follows a normal distribution with mean μ_p and with a standard deviation $SD(\hat{\mu}_p) = \sigma \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}}$, i.e., $\hat{\mu}_p \sim N(\mu_p, \sigma \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}})$. In the following figure, the histogram in the left panel shows the distribution of the mean price of used cars that are seven years old, based on 1,000 simple random samples of size $n = 15$. The distribution is roughly normal. The common standard deviation of the error term is estimated by the residual sample standard deviation s_e . Therefore,

$$\frac{\hat{\mu}_p - (\beta_0 + \beta_1 x_p)}{\sigma \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}} + 1} \sim N(0, 1) \text{ and } \frac{\hat{\mu}_p - (\beta_0 + \beta_1 x_p)}{s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}} + 1} \sim t \text{ with } df = n - 2.$$

A $(1 - \alpha) \times 100\%$ confidence interval for the conditional mean μ_p is

$$\hat{\mu}_p \pm t_{\alpha/2} \times SE(\hat{\mu}_p) = (b_0 + b_1 x_p) \pm t_{\alpha/2} \times s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}}.$$

A single value of the response variable, given the value of the predictor variable $x = x_p$ can be modeled by $y_p = \beta_0 + \beta_1 x_p + \epsilon$, $\epsilon \sim N(0, \sigma)$. A point estimate for y_p is given by $\hat{y}_p = b_0 + b_1 x_p$, which is the same as the point estimate for the conditional mean $\hat{\mu}_p$. It can be shown that the predicted value for a single response \hat{y}_p follows a normal distribution with mean $(\beta_0 + \beta_1 x_p)$ and with a standard deviation $SD(\hat{y}_p) = \sigma \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1}$, i.e., $\hat{y}_p \sim N\left(\beta_0 + \beta_1 x_p, \sigma \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1}\right)$.

In the following figure, the histogram in the right panel shows the distribution of the price of a randomly selected used car that is seven years old. The distribution is roughly normal. Therefore,

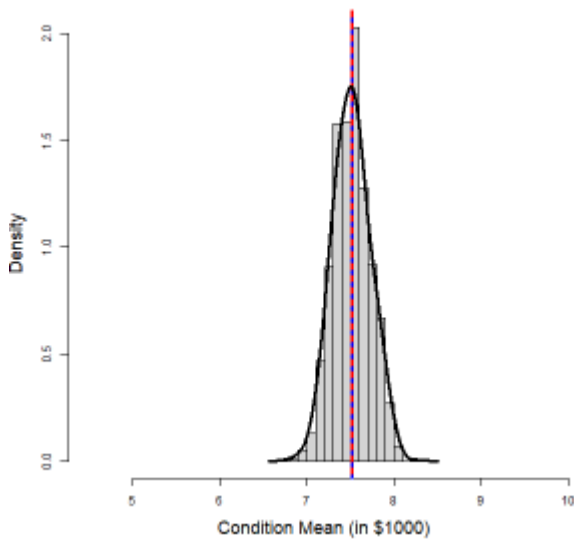
$$\frac{\hat{\mu}_p - (\beta_0 + \beta_1 x_p)}{\sigma \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1}} \sim N(0, 1) \text{ and } \frac{\hat{\mu}_p - (\beta_0 + \beta_1 x_p)}{s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1}} \sim t \text{ with } df = n - 2.$$

A $(1 - \alpha) \times 100\%$ confidence interval for a single response \hat{y}_p is

$$\hat{y}_p \pm t_{\alpha/2} \times SE(\hat{y}_p) = (b_0 + b_1 x_p) \pm t_{\alpha/2} \times s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1}.$$

We call this interval the **prediction interval** to distinguish between confidence intervals for the conditional mean versus confidence intervals for a single response.

Distribution for the Conditional Mean
 $\hat{\mu}_p = b_0 + b_1 x_p$



Distribution of a single value response
 $\hat{y}_p = b_0 + b_1 x_p$

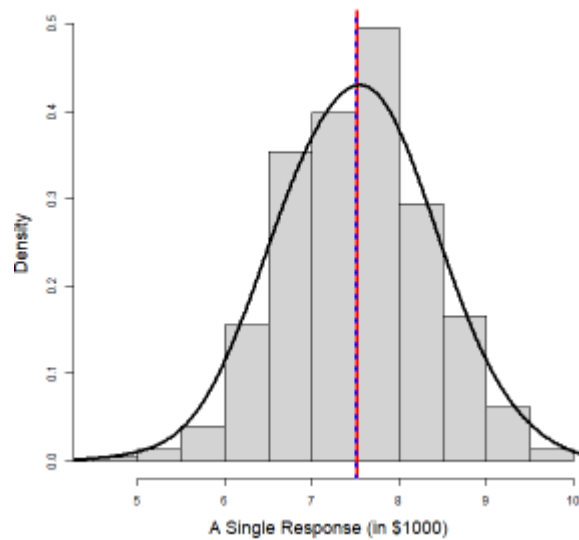


Figure 13.13: Distribution of the Conditional Mean (Left) and a Single Response (Right). [\[Image Description \(See Appendix D Figure 13.13\)\]](#) Click on the image to enlarge it.

Example: Confidence Interval and Prediction Interval

Recall the used car example. We have summaries:

$$\begin{aligned} n &= 15, \sum x_i = 92, \sum x_i^2 = 724, \sum y_i = 125, \sum y_i^2 = 1193, \sum x_i y_i = 616, \\ S_{xy} &= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 616 - \frac{92 \times 125}{15} = -150.667 \\ S_{xx} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 724 - \frac{92^2}{15} = 159.733 \end{aligned}$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1193 - \frac{125^2}{15} = 151.333$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{-150.667}{159.733} = -0.9432$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \frac{125}{15} - (-0.9432) \times \frac{92}{15} = 14.118.$$

The least-squares regression line is $\hat{y} = b_0 + b_1 x = 14.118 - 0.9432 \times \text{age}$.

$$SSE = S_{yy} - b_1 S_{xy} = 151.333 - (-0.9432) \times (-150.667) = 9.224$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{9.224}{13}} = 0.842.$$

- a. Construct a 95% confidence interval for the mean price for all used cars that are 7 years old.

$df = n - 2 = 13$, $\alpha = 0.05$, $t_{\alpha/2} = t_{0.025}$, $x_p = 7$, $\bar{x} = \frac{\sum x_i}{n} = \frac{92}{15} = 6.133$. Hence, the 95% confidence interval for the mean price is

$$\begin{aligned} & (b_0 + b_1 x_p) \pm t_{\alpha/2} \times s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}} \\ &= (14.118 - 0.9432 \times 7) \pm 2.160 \times 0.842 \times \sqrt{\frac{(7 - 6.133)^2}{159.733} + \frac{1}{15}} \\ &= (7.030, 8.002). \end{aligned}$$

Interpretation: We are 95% confident that the mean price of all 7-year-old cars is somewhere between \$7,030 and \$8,002.

- b. Obtain a 95% prediction interval for the sale price of a randomly selected used car that is 7 years old.

The 95% prediction interval for the price of a single car is

$$\begin{aligned} & (b_0 + b_1 x_p) \pm t_{\alpha/2} \times s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1} \\ &= (14.118 - 0.9432 \times 7) \pm 2.160 \times 0.842 \times \sqrt{\frac{(7 - 6.133)^2}{159.733} + \frac{1}{15} + 1} \\ &= (5.633, 9.398). \end{aligned}$$

Interpretation: We are 95% confident that the price of a randomly selected 7-year-old car is somewhere between \$5,633 and \$9,398.

- c. Which interval is wider? Explain why.

The prediction interval for a single predicted value is wider than the confidence interval for the conditional mean. This is because the error in predicting the price of a particular 7-year-old used car is due to the error in estimating the regression line plus the variation in prices of all 7-year-old cars. In contrast, the error in predicting the mean price of all 7-year-old cars is only due to the error in estimating the regression line.



Activity

Exercise: Application of Simple Linear Regression

An instructor asked a random sample of eight students to record their study times in an introductory calculus course. The total hours studied (x) over two weeks and the test scores (y) at the end of the two weeks are given in the following table.

Table 13.3: Test Score and Study Hour for Eight Students

Hours (x)	10	15	12	20	8	16	14	22
Score (y)	92	81	84	74	85	80	84	80

The summaries of the data are given by

$$n = 8, \sum x_i = 117, \sum x_i^2 = 1869, \\ \sum y_i = 660, \sum y_i^2 = 54638, \sum x_i y_i = 9519.$$

- Given the summaries of the data, find the least-squares regression equation.
- Interpret the slope of the regression equation obtained in part a) in the context of the study.
- Calculate and interpret r , the correlation coefficient between y and x .
- Calculate and interpret the coefficient of determination r^2 .
- Test at the 5% significant level, whether there is a negative linear association between study time and scores. You could use $s_e = 3.538$.
- What is the residual for the first response $y = 92$ with $x = 10$?

Show/Hide Answer

- Given the summaries of the data, find the least-squares regression equation.

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 9519 - \frac{117 \times 660}{8} = -133.5$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 1869 - \frac{117^2}{8} = 157.875$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 54638 - \frac{660^2}{8} = 188$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{-133.5}{157.875} = -0.8456$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \frac{660}{8} - (-0.8456) \times \frac{117}{8} = 94.8670.$$

And therefore, the least-square straight line is $\widehat{\text{score}} = 94.8670 - 0.8456 \times \text{hours}$.

- Interpret the slope of the regression equation obtained in part a) in the context of the study.
 $b_1 = -0.8456$ implies the average score reduces by 0.8456 for each additional hour spent studying.
- Calculate and interpret r , the correlation coefficient between y and x .

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{-133.5}{\sqrt{157.875 \times 188}} = -0.7749.$$

Interpretation: There is a moderate negative linear association between score and hours of study.

- Calculate and interpret the coefficient of determination r^2 .

$$r^2 = (-0.7749)^2 = 0.6005.$$

Interpretation: 60.05% of the variation in the observed exam scores can be explained by hours of study through the regression line $\widehat{\text{score}} = 94.8670 - 0.8456 \times \text{hour}$.

- Test at the 5% significant level, whether there is a negative linear association between study time

and scores. You could use $s_e = 3.538$.

Steps:

1. Set up the hypotheses. $H_0 : \beta_1 \geq 0$ versus $H_a : \beta_1 < 0$.
2. The significance level is $\alpha = 0.05$.
3. Compute the value of the test statistic: $t_o = \frac{b_1}{\frac{s_e}{\sqrt{S_{xx}}}}$ with $df = n - 2$. Therefore,

$$t_o = \frac{b_1}{\frac{s_e}{\sqrt{S_{xx}}}} = \frac{-0.8456}{\left(\frac{3.538}{\sqrt{157.875}}\right)} = -3.003, df = n - 2 = 8 - 2 = 6.$$

4. Find the P-value. For a left-tailed test with $df = 6$,
P-value = $P(t \leq t_o) = P(t \leq -3.003) = P(t \geq 3.003)$. Thus, since
 $2.447(t_{0.025}) < 3.003 < 3.143(t_{0.01})$, it follows that
 $0.01 < \text{P-value} = P(t \geq 3.003) < 0.025$.
 5. Decision: Reject the null H_0 since P-value $< 0.025 < 0.05(\alpha)$.
 6. Conclusion: At the 5% significance level, we have sufficient evidence of a negative linear association between study time and scores.
- f. What is the residual for the first response $y = 92$?

The first residual is

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 = y_1 - (b_0 + b_1 \times x_1) = 92 - (94.8670 - 0.8456 \times 10) = 92 - 86.411 \\ &= 5.589. \end{aligned}$$

13.10 Learning Objectives Revisited

Learning Objectives

As a result of completing this chapter, you will be able to do the following:

- Identify situations where simple linear regression should be used (Section 13.1).
- Explain the main idea of the method of least squares (Section 13.2).
- Calculate the least-squares fitted line (Section 13.2).
- Calculate and interpret the correlation coefficient r (Section 13.5).
- Calculate and interpret the coefficient of determination r^2 (Section 13.6).
- Explain the terms in a simple linear regression model (Section 13.7).
- Conduct a t test and obtain a t confidence interval for the slope parameter β_1 (Section 13.8).
- Explain the difference between confidence intervals and prediction intervals (Section 13.9).
- Obtain a confidence interval for the conditional mean and a prediction interval for a single response (Section 13.9).

13.11 Review Questions

Researchers examined the controversial issue of the human vomeronasal organ regarding its structure, function, and identity. The following table shows the age of fetuses (x) in weeks and the length of crown-rump (y) in millimeters.

Age (x)	10	10	13	13	18	19	19	23	25	28
Length (y)	66	66	108	106	161	166	177	228	235	280

The summaries of the data are given by
 $n = 10, \sum x_i = 178, \sum x_i^2 = 3522, \sum y_i = 1593, \sum y_i^2 = 302027, \sum x_i y_i = 32476$.

- Given the summaries of the data, find the least-squares regression equation.
- Graph the regression equation and the data points.
- Interpret the slope of the regression equation obtained in part (a) in the context of the study.
- Calculate r , the correlation coefficient between y and x . Interpret the number.
- Calculate the coefficient of determination r^2 . Interpret the number.
- Test at the 1% significant level whether the age of fetuses is a useful predictor for the length of the crown-rump. You could use $s_e = 5.518$.
- Predict the crown-rump length of a 19-week-old fetus.
- What is the residual for the last observation with response $y = 280$ and $x = 28$?

Show/Hide Answer

$$\begin{aligned}
 S_{xx} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 3522 - \frac{(178)^2}{10} = 353.6; \\
 S_{xy} &= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 32476 - \frac{(178)(1593)}{10} = 4120.6; \\
 S_{yy} &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 302027 - \frac{(1593)^2}{10} = 48262.1; \\
 b_1 &= \frac{S_{xy}}{S_{xx}} = \frac{4120.6}{353.6} = 11.65328; \\
 b_0 &= \frac{\sum y_i}{n} - b_1 \times \frac{\sum x_i}{n} = \frac{1593}{10} - 11.65328 \times \frac{178}{10} = -48.12838.
 \end{aligned}$$

a.

Therefore, the least-squares regression equation $\hat{y} = b_0 + b_1 x = -48.12838 + 11.65328x$ or $\widehat{\text{length}} = -48.12838 + 11.65328 \text{age}$.

- Left as an exercise for the reader.

- c. The slope is $b_1 = 11.65328$.

Interpretation: The average length of the crown rump increases by 11.65328 millimeters when the age of the fetus increases by 1 week. In other words, for each week the fetus ages, the expected increase in crown-rump length is 11.65328 mm.

- d. The correlation coefficient r is given by $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{4120.6}{\sqrt{(353.6)(48262.1)}} = 0.9974732$.

Interpretation: there is a very strong, positive, linear association between the length of crown-rump (y) and the age (x) of the fetus.

- e. The coefficient of determination is $r^2 = 0.9974732^2 = 0.9949528$.

Interpretation: 99.50% of the variation in the length of the crown rump is due to the age of the fetus. Or 99.50% of the variation in the length of crown-rump can be explained by the age of the fetus through the fitted regression line

$$\hat{y} = b_0 + b_1x = -48.12838 + 11.65328x.$$

- f. We assume all assumptions for inference on simple linear regression are satisfied.

Step 1: Hypotheses. $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

Step 2: Significance level $\alpha = 0.01$.

Step 3: Test statistic $t_o = \frac{b_1}{(\frac{s_e}{\sqrt{S_{xx}}})} = \frac{11.65328}{(\frac{5.518}{\sqrt{353.6}})} = 39.71208$ with $df = n - 2 = 10 - 2 = 8$.

Step 4: P-value. It is a two-tailed test,

$$\text{p-value} = 2P(t \geq |t_o|) = 2P(t \geq 39.71208) < 2 \times 0.0005 = 0.001.$$

Step 5: Decision. We reject H_0 since $\text{p-value} < 0.001 < 0.01(\alpha)$.

Step 6: Conclusion. At the 1% significant level, we have sufficient evidence that the age of fetuses is a useful predictor for the crown-rump length.

$$\hat{y} = b_0 + b_1x = -48.12838 + 11.65328x = -48.12838 + 11.65328 \times 19 =$$

- g. 173.2839

The predicted crown-rump length of a 19-week-old fetus is 173.2839 mm.

$$e = y - \hat{y} = y - (b_0 + b_1x) = 280 - (-48.12838 + 11.65328 \times 28) =$$

- h. Residual $280 - 278.1635 = 1.8365$.

13.12 Assignment 13

Purposes

This assignment has two parts. The first part assesses your knowledge of explaining the idea of least-squares method, obtaining the least-squares regression equation, calculating and interpreting the correlation coefficient r and the coefficient of determination r^2 , interpreting the terms in the simple linear regression model, conducting a t test and obtaining a t confidence interval for the slope parameter β_1 , predicting the value of the response variable given the value of the predictor variable, and calculating the fitted value and the residual of an observation. The second part assesses your skills in using R Commander to fit a least-squares simple linear regression model and conduct a t test.

Resources

[M13_HomePrice_Regression_Q1.xlsx](#)

[M13_CrownRump_Length_Regression_Q2.xlsx](#)

Instructions

Part A

Complete the following:

1. A random sample of nine custom homes currently listed for sale provided the following information on size and price. Here, x denotes size, in hundreds of square feet, rounded to the nearest hundred, and y denotes price, in thousands of dollars, rounded to the nearest thousand.

x	26	27	33	29	29	34	30	40	22
y	540	555	575	577	606	661	738	804	496

The summaries of the data are given by

$n = 9$, $\sum x_i = 270$, $\sum x_i^2 = 8316$, $\sum y_i = 5552$, $\sum y_i^2 = 3504412$, $\sum x_i y_i = 169993$.

- Given the summaries of the data, find the least-squares regression equation. (5 marks)
 - Graph the regression equation and the data points. (4 marks)
 - Interpret the slope of the regression equation obtained in part (a) in the context of the study. (2 marks)
 - Calculate r , the correlation coefficient between y and x . Interpret the number. (4 marks)
 - Calculate the coefficient of determination r^2 . Interpret the number. (3 marks)
 - Test at the 5% significant level whether the size is a useful predictor for the price of the custom homes. You could use $s_e = 59.62$. (8 marks)
 - Obtain a 95% confidence interval for the slope of the population regression line that relates price to size for custom homes. (4 marks)
 - Interpret the confidence interval obtained in part (g). Does this interval support the results of the hypothesis test in part (f)? (4 marks: 2+2)
2. Researchers examined the controversial issue of the human vomeronasal organ, an auxiliary olfactory sense organ located at the base of the nasal cavity for detecting chemical stimuli, regarding its structure, function, and identity. The following table shows the age of fetuses (x) in weeks and the length of crown-rump (y) in millimeters.

x	10	10	13	13	18	19	19	23	25	28
y	66	66	108	106	161	166	177	228	235	280

The summaries of the data are given by

$n = 10$, $\sum x_i = 178$, $\sum x_i^2 = 3,522$, $\sum y_i = 1,593$, $\sum y_i^2 = 302,027$, $\sum x_i y_i = 32,476$.

- Given the summaries of the data, find the least-squares regression equation. (5 marks)
- Interpret the slope of the regression equation obtained in part (a) in the context of the study. (2 marks)
- Calculate r , the correlation coefficient between y and x . Interpret the number. (4 marks)
- Calculate the coefficient of determination r^2 . Interpret the number. (3 marks)
- Test at the 1% significant level whether the age of fetuses is a useful predictor for the length of crown-rump. You could use $s_e = 5.518$. (8 marks)
- Predict the crown-rump length of a 19-week-old fetus. (2 marks)
- What is the residual for the last observation with response $y = 280$ and $x = 28$? (3 marks)

Part B

Finish the following questions using R and R commander. Make sure that you copy and paste the computer outputs as required and write down your answers in statements.

1. Refer to Question 1 in Part A. The data are provided in the file **M13_HomePrice_Regression_Q1.xlsx**. Import the data into R commander and complete the following questions and tasks.
 - a. Could we use a straight line to model the relationship between size and price of the custom homes? Use a proper graphical method to justify your answer. (3 marks)
 - b. How does the price of custom homes change when the size increases? Would the slope be positive or negative? (2 marks)
 - c. Obtain the least-squares regression equation using R commander. First copy and paste the computer and then compare the answer obtained by hand in Question 1 part (a). (3 marks)
 - d. Obtain the coefficient of determination r^2 and the correlation of coefficient r from the computer output. First copy and paste the computer output and then compare the answers with the ones you obtained by hand in Question 1 (d) and (e). (5 marks)
 - e. Re-conduct the hypothesis test in Question 1 (f) using R commander. Make sure to include all the six components of a hypothesis test. First copy and paste the computer output and then compare the answer with the one you obtained by hand in Question 1 (f). (5 marks)
2. Refer to Question 2 in Part A. The data are provided in the file **M13_CrownRump_Length_Regression_Q2.xlsx**. Import the data into R commander and complete the following questions or tasks.
 - a. Is it reasonable to model the relationship between age and crown-rump length for fetuses using a straight line? Justify your answer by a proper graphical tool using R commander. (3 marks)
 - b. Obtain the least-squares regression equation using R commander. Copy and paste the computer output first and then compare the answer obtained by hand in Question 2 (a). (3 marks)
 - c. Obtain the coefficient of determination r^2 and the correlation of coefficient r from the computer output. Copy and paste the computer output and then compare the answers with the ones you obtained by hand in Question 2 (c) and (d). (5 marks)
 - d. Re-conduct the hypothesis test in Question 2 (e) using R commander. Make sure to include all the six components of a hypothesis test. Copy and paste the

computer output first and then compare the answer with the one you obtained by hand in Question 2 (e). (5 marks)

Quiz 12



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://openbooks.macewan.ca/introstats/?p=2643#h5p-15>

Appendix A: Formula Sheet

Important Notations

Measures	Population	Sample
Sample Size	N	n
Mean	μ	$\bar{\mu}$
Standard Deviation	σ	s
Proportion	p	\hat{p}
Slope	β_1	b_1

Descriptive Measures

- Five-number summary: minimum, Q_1 , Q_2 , Q_3 , and maximum
- Outliers: lowerlimit = $Q_1 - 1.5 \times IQR$; upperlimit = $Q_3 + 1.5 \times IQR$; $IQR = Q_3 - Q_1$
- Sample mean: $\frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}$
- Sample standard deviation: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}{n-1}}$

Probability Concepts

- **Equal-likely outcome model: Probability of event E**

$$P(E) = \frac{\text{of sample points in event } E}{\text{of sample points in sample space } S} = \frac{\text{of ways event } E \text{ can occur}}{\text{of possible outcomes}} = \frac{f}{N}$$

- Complement rule: $P(\text{not } E) = 1 - P(E)$
- Special addition rule: $P(A \text{ or } B) = P(A) + P(B)$ if events A and B are mutually exclusive.
More generally, if events A, B, C, \dots are mutually exclusive, then
 $P(A \text{ or } B \text{ or } C \text{ or } \dots) = P(A) + P(B) + P(C) + \dots$
- General addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$
- Conditional probability of A given B : $P(A|B) = \frac{P(A \& B)}{P(B)}$ for $P(B) > 0$
- General multiplication rule: $P(A \& B) = P(B)P(A|B) = P(A)P(B|A)$
- Special multiplication rule: $P(A \& B) = P(A)P(B)$ if events A and B are independent.
More generally, if events A, B, C, \dots are independent,
 $P(A \& B \& C \& \dots) = P(A) \times P(B) \times P(C) \times \dots$
- Two events A and B are **independent** if **ANY** of the following is true:

$$P(A|B) = P(A) \text{ OR } P(B|A) = P(B) \text{ OR } P(A \& B) = P(A) \times P(B)$$

- Permutation: $nP_r = \frac{n!}{(n-r)!}$
- Combination: $nC_r = \frac{n!}{r!(n-r)!}$

Discrete Random Variables

- The mean (expected value) of a discrete random variable X : $\mu = \sum xP(X = x)$
- Standard deviation of X : $\sigma = \sqrt{\sum (x - \mu)^2 P(X = x)} = \sqrt{\sum x^2 P(X = x) - \mu^2}$

Binomial Distribution

Among n independent Bernoulli trials with probability of success p , let X be the number of successes. The probability of observing x successes is:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} = nC_x p^x (1 - p)^{n-x}, x = 0, 1, \dots, n$$

The mean and standard deviation of a Binomial distribution are $\mu = np$, $\sigma = \sqrt{np(1-p)}$, respectively.

Normal Distribution

- If random variable $X \sim N(\mu, \sigma)$, then the standardized variable $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$.
- Given an x value, its z-score is $z = \frac{x - \mu}{\sigma}$
- Given the z-score, find the x value: $x = \mu + z \times \sigma$.
- If sample mean $\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}})$, the standardized variable $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

Sampling Distributions

- Mean and standard deviation of the sample mean \bar{X} : $\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- Mean and standard deviation of a sample proportion \hat{p} : $\mu_{\hat{p}} = p; \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- Mean and standard deviation of $\bar{X}_1 - \bar{X}_2$: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2; \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Mean and standard deviation of $\hat{p}_1 - \hat{p}_2$: $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2; \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Confidence Intervals and Hypothesis Tests

Parameter	Estimate	Test Statistic	$(1 - \alpha) \times 100\%$ Confidence Interval
μ	\bar{x}	$t_o = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$	$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ with $df = n - 1$
p	\hat{p}	$z_o = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} = \frac{x}{n}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ with $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ with $df = n_1 + n_2 - 2$
$\mu_1 - \mu_2$	\bar{d}	$t_o = \frac{\bar{d} - \delta_0}{\left(\frac{s_d}{\sqrt{n}}\right)}$	$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$ with $df = n - 1, n = \# \text{ of pairs}$ $\bar{d} = \frac{\sum d_i}{n}, s_d = \sqrt{\frac{(\sum d_i^2) - \frac{(\sum d_i)^2}{n}}{n-1}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ $\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}, \hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$
β_1	b_1	$t_o = \frac{b_1}{\left(\frac{s_e}{\sqrt{S_{xx}}}\right)}$	$b_1 \pm t_{\alpha/2} \frac{s_e}{\sqrt{S_{xx}}}$ with $df = n - 2$

Margin of Error and Sample Size Calculation

- Margin of error for the estimate of μ : $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- Sample size calculation for μ : $n = \left(\frac{\sigma \times z_{\alpha/2}}{E}\right)^2$ round up to the nearest integer
- Margin of error for the estimate of p : $E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Sample size calculation for p without guessing \hat{p} : $n \leq 0.05(1 - 0.05) \left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.25 \left(\frac{z_{\alpha/2}}{E}\right)^2$
- Sample size calculation for p with guessing \hat{p} : $n = p_g(1 - p_g) \left(\frac{z_{\alpha/2}}{E}\right)^2$ round up

Chi-Square Test

- Chi-square goodness-of-fit test for one categorical/discrete variable:
 - Expected frequency: $E = np$
 - Test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$ with $df = k - 1$, where k is number of possible values of the variable

- Chi-square independence (or homogeneity) test of two variables:
 - Expected frequency: $E = \frac{(\text{rth row total}) \times (\text{cth column total})}{n}$
 - Test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$ with $df = (r-1) \times (c-1)$ where r is the number of rows and c is number of columns of the cells.

Regression Analysis

- Sums of squares

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}; S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}; S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

- The least-squares straight line: $\hat{y} = b_0 + b_1 x$, where $b_1 = \frac{S_{xy}}{S_{xx}}$ and $b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n}$
- Total sum of squares: $SST = \sum (y_i - \bar{y})^2 = S_{yy}$
- Regression sum of squares: $SSR = \sum (\hat{y} - \bar{y})^2 = r^2 S_{yy} = \frac{S_{xy}^2}{S_{xx}}$
- Error sum of squares: $SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2 = SST - SSR = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$
- Regression identify: $SST = SSE + SSR$
- Residual: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$
- Correlation coefficient: $r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$
- Coefficient of determination: $R^2 = r^2 = \frac{S_{xy}^2}{S_{xx} \times S_{yy}} = \frac{SSR}{SST}$
- Standard error of the estimate: $s_e = \sqrt{\frac{\sum (e_i - \bar{e})^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$
- Test statistic for β_1 : $t_o = \frac{b_1}{\left(\frac{s_e}{\sqrt{S_{xx}}}\right)}$ with $df = n - 2$
- A $(1 - \alpha) \times 100\%$ confidence interval for β_1 : $b_1 \pm t_{\alpha/2} \frac{s_e}{\sqrt{S_{xx}}}$ with $df = n - 2$
- A $(1 - \alpha) \times 100\%$ confidence interval for the conditional mean μ_p is

$$\hat{\mu}_p \pm t_{\alpha/2} \times SE(\hat{\mu}_p) = (b_0 + b_1 x_p) \pm t_{\alpha/2} \times s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n}} \text{ with } df = n - 2$$

- A $(1 - \alpha) \times 100\%$ confidence interval for a single response y_p is

$$\hat{y}_p \pm t_{\alpha/2} \times SE(\hat{y}_p) = (b_0 + b_1 x_p) \pm t_{\alpha/2} \times s_e \sqrt{\frac{(x_p - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1} \text{ with } df = n - 2$$

Analysis of Variance (One-Way ANOVA F Test)

Compare k population means: $\mu_1, \mu_2, \dots, \mu_k$. Denote sample sizes as n_1, n_2, \dots, n_k , sample means as $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ and sample standard deviations as s_1, s_2, \dots, s_k . Let $n = n_1 + n_2 + \dots + n_k$ and $\bar{x} = \frac{\sum x_{ij}}{n}$ where x_{ij} is the j th observation of sample i .

- Test statistic: $F_o = \frac{SSTR/(k-1)}{SSE/(n-k)} = \frac{MSTR}{MSE}$ with $df_n = k - 1$ and $df_d = n - k$
- Total sum of squares: $SST = \sum (x_{ij} - \bar{x})^2 = \sum x_{ij}^2 - \frac{(\sum x_{ij})^2}{n}$
- Treatment sum of squares: $SSTR = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$

- Error sum of squares: $SSE = \sum (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1) s_i^2 = SST - SSTR$
- ANOVA identity: $SST = SSE + SSTR$

Appendix B: Statistical Tables

[Table I: Random Number Table](#)

[Table II: Area under the standard normal curve to the left of a given \$z\$ score](#)

[Table III: Normal scores \(theoretical quantiles\) for normal Q-Q plot given a certain sample size \$n\$](#)

[Table IV: Values of \$t_{\alpha}\$ of \$t\$ -distribution for a given degrees of freedom \$df\$](#)

[Table V: Values of \$\chi^2_{\alpha}\$ of \$\chi^2\$ -distribution for a given degrees of freedom \$df\$](#)

[Table VI: Values of \$F_{\alpha}\$ of \$F\$ -distribution for given degrees of freedom \$df_n\$ and \$df_d\$](#)

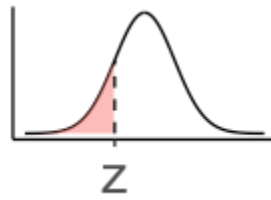
Acknowledgment

Table II (area under the standard normal), Table IV (t -score table), Table V (χ^2 table), and Table VI (F table) were adopted and modified from the statistical tables created by Haziq Jamil from <https://haziqj.github.io/stat-tables>.

Table I: Random Number Table

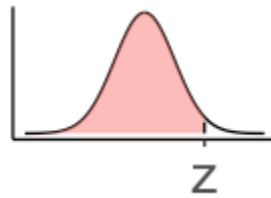
Line	Column Number				
Number	00-09	10-19	20-29	30-39	40-49
00	16704 99757	02084 13701	91085 22010	84870 18477	65610 81453
01	87858 50308	47345 91528	28483 00467	72419 98125	50436 40454
02	30868 11973	79360 01976	78239 77854	70053 02503	06774 74584
03	36079 34997	53913 19964	22889 04348	19640 84223	86641 87187
04	31972 18313	87798 60455	82994 27904	54207 24843	64660 70822
05	82528 67314	12700 92323	19513 23922	47133 45420	45622 65822
06	65623 58908	51813 02385	32989 50597	93006 18262	52978 57243
07	57678 08569	15198 88216	66438 29008	84161 50120	63153 78982
08	67357 87763	72548 63577	47562 74495	07752 30648	41034 10823
09	68229 84134	91022 36078	73441 85333	95723 90445	34364 89746
10	18429 88041	42363 80299	05241 83520	48786 77428	47528 15818
11	91677 04920	24220 76025	88296 88237	99147 71691	08498 91990
12	97320 44164	36087 02974	78646 10845	64450 39205	49446 89839
13	71952 71990	96952 89645	21953 95674	03307 94580	69395 53166
14	74434 23251	69387 11435	85297 29360	91167 31999	91952 99976
15	75625 54072	15596 19760	82205 38602	81571 42905	73072 21498
16	78068 61799	04149 60182	29886 98331	16522 02877	88431 89780
17	16904 45998	32476 66193	61188 58177	13377 93954	92140 97713
18	02425 00548	59490 27868	36700 41390	57153 77361	24628 57530
19	74813 19215	44605 21467	62776 22892	00394 42249	96697 02241

Table II: Area under the standard normal curve for negative z



Second decimal place in z										z
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-3.9
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	-3.8
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	-3.7
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	-3.6
0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	-3.5
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	-3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	-3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	-3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	-3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	-2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	-2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	-2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	-2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	-2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	-2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	-2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	-2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	-2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	-2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	-1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	-1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	-1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	-1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	-1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	-1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	-1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	-1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	-0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	-0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	-0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	-0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	-0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	-0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	-0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	-0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	-0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	-0.0

Table II: Area under the standard normal curve for positive z



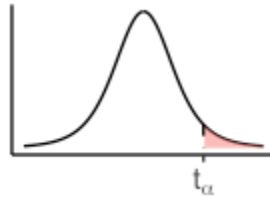
z	Second decimal place in z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table III: Normal scores for normal Q-Q plot

Rank	Sample size n												
	5	6	7	8	9	10	11	12	13	14	15	16	17
1	-1.18	-1.28	-1.36	-1.43	-1.49	-1.55	-1.59	-1.64	-1.67	-1.71	-1.74	-1.77	-1.80
2	-0.50	-0.64	-0.76	-0.85	-0.93	-1.00	-1.06	-1.11	-1.16	-1.21	-1.25	-1.28	-1.32
3	0.00	-0.20	-0.35	-0.47	-0.57	-0.66	-0.73	-0.79	-0.85	-0.90	-0.95	-0.99	-1.03
4	0.50	0.20	0.00	-0.15	-0.27	-0.38	-0.46	-0.54	-0.60	-0.66	-0.71	-0.76	-0.81
5	1.18	0.64	0.35	0.15	0.00	-0.12	-0.22	-0.31	-0.39	-0.45	-0.51	-0.57	-0.62
6		1.28	0.76	0.47	0.27	0.12	0.00	-0.10	-0.19	-0.27	-0.33	-0.40	-0.45
7			1.36	0.85	0.57	0.38	0.22	0.10	0.00	-0.09	-0.17	-0.23	-0.29
8				1.43	0.93	0.66	0.46	0.31	0.19	0.09	0.00	-0.08	-0.15
9					1.49	1.00	0.73	0.54	0.39	0.27	0.17	0.08	0.00
10						1.55	1.06	0.79	0.60	0.45	0.33	0.23	0.15
11							1.59	1.11	0.85	0.66	0.51	0.40	0.29
12								1.64	1.16	0.90	0.71	0.57	0.45
13									1.67	1.21	0.95	0.76	0.62
14										1.71	1.25	0.99	0.81
15											1.74	1.28	1.03
16												1.77	1.32
17													1.80

Rank	Sample size n												
	18	19	20	21	22	23	24	25	26	27	28	29	30
1	-1.82	-1.85	-1.87	-1.89	-1.91	-1.93	-1.95	-1.96	-1.98	-2.00	-2.01	-2.03	-2.04
2	-1.35	-1.38	-1.40	-1.43	-1.45	-1.48	-1.50	-1.52	-1.54	-1.56	-1.58	-1.59	-1.61
3	-1.06	-1.10	-1.13	-1.16	-1.19	-1.21	-1.24	-1.26	-1.28	-1.30	-1.32	-1.34	-1.36
4	-0.85	-0.88	-0.92	-0.95	-0.98	-1.01	-1.04	-1.06	-1.09	-1.11	-1.13	-1.16	-1.18
5	-0.66	-0.71	-0.74	-0.78	-0.81	-0.85	-0.88	-0.90	-0.93	-0.96	-0.98	-1.00	-1.02
6	-0.50	-0.55	-0.59	-0.63	-0.67	-0.70	-0.73	-0.76	-0.79	-0.82	-0.84	-0.87	-0.89
7	-0.35	-0.40	-0.45	-0.49	-0.53	-0.57	-0.60	-0.64	-0.67	-0.70	-0.72	-0.75	-0.78
8	-0.21	-0.26	-0.31	-0.36	-0.41	-0.45	-0.48	-0.52	-0.55	-0.58	-0.61	-0.64	-0.67
9	-0.07	-0.13	-0.19	-0.24	-0.29	-0.33	-0.37	-0.41	-0.44	-0.48	-0.51	-0.54	-0.57
10	0.07	0.00	-0.06	-0.12	-0.17	-0.22	-0.26	-0.30	-0.34	-0.38	-0.41	-0.44	-0.47
11	0.21	0.13	0.06	0.00	-0.06	-0.11	-0.16	-0.20	-0.24	-0.28	-0.32	-0.35	-0.38
12	0.35	0.26	0.19	0.12	0.06	0.00	-0.05	-0.10	-0.14	-0.19	-0.22	-0.26	-0.29
13	0.50	0.40	0.31	0.24	0.17	0.11	0.05	0.00	-0.05	-0.09	-0.13	-0.17	-0.21
14	0.66	0.55	0.45	0.36	0.29	0.22	0.16	0.10	0.05	0.00	-0.04	-0.09	-0.12
15	0.85	0.71	0.59	0.49	0.41	0.33	0.26	0.20	0.14	0.09	0.04	0.00	-0.04
16	1.06	0.88	0.74	0.63	0.53	0.45	0.37	0.30	0.24	0.19	0.13	0.09	0.04
17	1.35	1.10	0.92	0.78	0.67	0.57	0.48	0.41	0.34	0.28	0.22	0.17	0.12
18	1.82	1.38	1.13	0.95	0.81	0.70	0.60	0.52	0.44	0.38	0.32	0.26	0.21
19		1.85	1.40	1.16	0.98	0.85	0.73	0.64	0.55	0.48	0.41	0.35	0.29
20			1.87	1.43	1.19	1.01	0.88	0.76	0.67	0.58	0.51	0.44	0.38
21				1.89	1.45	1.21	1.04	0.90	0.79	0.70	0.61	0.54	0.47
22					1.91	1.48	1.24	1.06	0.93	0.82	0.72	0.64	0.57
23						1.93	1.50	1.26	1.09	0.96	0.84	0.75	0.67
24							1.95	1.52	1.28	1.11	0.98	0.87	0.78
25								1.96	1.54	1.30	1.13	1.00	0.89
26									1.98	1.56	1.32	1.16	1.02
27										2.00	1.58	1.34	1.18
28											2.01	1.59	1.36
29												2.03	1.61
30													2.04

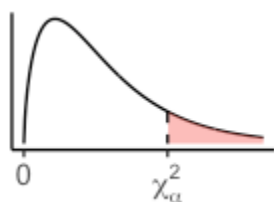
Table IV: Values of t_α of t -distribution



df	α : Area to the Right of t_α											
	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.010	0.0075	0.0050	0.0025	0.0005
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	31.821	42.433	63.657	127.321	636.619
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	6.965	8.073	9.925	14.089	31.599
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	4.541	5.047	5.841	7.453	12.924
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	3.747	4.088	4.604	5.598	8.610
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	3.365	3.634	4.032	4.773	6.869
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	3.143	3.372	3.707	4.317	5.959
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.998	3.203	3.499	4.029	5.408
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.896	3.085	3.355	3.833	5.041
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.821	2.998	3.250	3.690	4.781
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.764	2.932	3.169	3.581	4.587
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.718	2.879	3.106	3.497	4.437
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.681	2.836	3.055	3.428	4.318
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.650	2.801	3.012	3.372	4.221
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.624	2.771	2.977	3.326	4.140
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.602	2.746	2.947	3.286	4.073
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.583	2.724	2.921	3.252	4.015
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	2.567	2.706	2.898	3.222	3.965
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.552	2.689	2.878	3.197	3.922
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	2.539	2.674	2.861	3.174	3.883
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.528	2.661	2.845	3.153	3.850
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	2.518	2.649	2.831	3.135	3.819
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	2.508	2.639	2.819	3.119	3.792
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	2.500	2.629	2.807	3.104	3.768
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	2.492	2.620	2.797	3.091	3.745
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.485	2.612	2.787	3.078	3.725
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	2.479	2.605	2.779	3.067	3.707
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	2.473	2.598	2.771	3.057	3.690
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	2.467	2.592	2.763	3.047	3.674
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	2.462	2.586	2.756	3.038	3.659
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.457	2.581	2.750	3.030	3.646
31	0.256	0.530	0.853	1.054	1.309	1.696	2.040	2.453	2.576	2.744	3.022	3.633
32	0.255	0.530	0.853	1.054	1.309	1.694	2.037	2.449	2.571	2.738	3.015	3.622
33	0.255	0.530	0.853	1.053	1.308	1.692	2.035	2.445	2.566	2.733	3.008	3.611
34	0.255	0.529	0.852	1.052	1.307	1.691	2.032	2.441	2.562	2.728	3.002	3.601
35	0.255	0.529	0.852	1.052	1.306	1.690	2.030	2.438	2.558	2.724	2.996	3.591
36	0.255	0.529	0.852	1.052	1.306	1.688	2.028	2.434	2.555	2.719	2.990	3.582
37	0.255	0.529	0.851	1.051	1.305	1.687	2.026	2.431	2.551	2.715	2.985	3.574
38	0.255	0.529	0.851	1.051	1.304	1.686	2.024	2.429	2.548	2.712	2.980	3.566
39	0.255	0.529	0.851	1.050	1.304	1.685	2.023	2.426	2.545	2.708	2.976	3.558
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.423	2.542	2.704	2.971	3.551
41	0.255	0.529	0.850	1.050	1.303	1.683	2.020	2.421	2.539	2.701	2.967	3.544
42	0.255	0.528	0.850	1.049	1.302	1.682	2.018	2.418	2.537	2.698	2.963	3.538
43	0.255	0.528	0.850	1.049	1.302	1.681	2.017	2.416	2.534	2.695	2.959	3.532
44	0.255	0.528	0.850	1.049	1.301	1.680	2.015	2.414	2.532	2.692	2.956	3.526
45	0.255	0.528	0.850	1.049	1.301	1.679	2.014	2.412	2.529	2.690	2.952	3.520
46	0.255	0.528	0.850	1.048	1.300	1.679	2.013	2.410	2.527	2.687	2.949	3.515
47	0.255	0.528	0.849	1.048	1.300	1.678	2.012	2.408	2.525	2.685	2.946	3.510
48	0.255	0.528	0.849	1.048	1.299	1.677	2.011	2.407	2.523	2.682	2.943	3.505
49	0.255	0.528	0.849	1.048	1.299	1.677	2.010	2.405	2.521	2.680	2.940	3.500
50	0.255	0.528	0.849	1.047	1.299	1.676	2.009	2.403	2.519	2.678	2.937	3.496

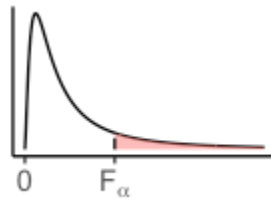
df	α : Area to the Right of t_α											
	0.40	0.30	0.20	0.15	0.10	0.05	0.025	0.010	0.0075	0.0050	0.0025	0.0005
51	0.255	0.528	0.849	1.047	1.298	1.675	2.008	2.402	2.518	2.676	2.934	3.492
52	0.255	0.528	0.849	1.047	1.298	1.675	2.007	2.400	2.516	2.674	2.932	3.488
53	0.255	0.528	0.848	1.047	1.298	1.674	2.006	2.399	2.514	2.672	2.929	3.484
54	0.255	0.528	0.848	1.046	1.297	1.674	2.005	2.397	2.513	2.670	2.927	3.480
55	0.255	0.527	0.848	1.046	1.297	1.673	2.004	2.396	2.511	2.668	2.925	3.476
56	0.255	0.527	0.848	1.046	1.297	1.673	2.003	2.395	2.510	2.667	2.923	3.473
57	0.255	0.527	0.848	1.046	1.297	1.672	2.002	2.394	2.508	2.665	2.920	3.470
58	0.255	0.527	0.848	1.046	1.296	1.672	2.002	2.392	2.507	2.663	2.918	3.466
59	0.254	0.527	0.848	1.046	1.296	1.671	2.001	2.391	2.506	2.662	2.916	3.463
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.390	2.504	2.660	2.915	3.460
61	0.254	0.527	0.848	1.045	1.296	1.670	2.000	2.389	2.503	2.659	2.913	3.457
62	0.254	0.527	0.847	1.045	1.295	1.670	1.999	2.388	2.502	2.657	2.911	3.454
63	0.254	0.527	0.847	1.045	1.295	1.669	1.998	2.387	2.501	2.656	2.909	3.452
64	0.254	0.527	0.847	1.045	1.295	1.669	1.998	2.386	2.500	2.655	2.908	3.449
65	0.254	0.527	0.847	1.045	1.295	1.669	1.997	2.385	2.499	2.654	2.906	3.447
66	0.254	0.527	0.847	1.045	1.295	1.668	1.997	2.384	2.498	2.652	2.904	3.444
67	0.254	0.527	0.847	1.045	1.294	1.668	1.996	2.383	2.497	2.651	2.903	3.442
68	0.254	0.527	0.847	1.044	1.294	1.668	1.995	2.382	2.496	2.650	2.902	3.439
69	0.254	0.527	0.847	1.044	1.294	1.667	1.995	2.382	2.495	2.649	2.900	3.437
70	0.254	0.527	0.847	1.044	1.294	1.667	1.994	2.381	2.494	2.648	2.899	3.435
71	0.254	0.527	0.847	1.044	1.294	1.667	1.994	2.380	2.493	2.647	2.897	3.433
72	0.254	0.527	0.847	1.044	1.293	1.666	1.993	2.379	2.492	2.646	2.896	3.431
73	0.254	0.527	0.847	1.044	1.293	1.666	1.993	2.379	2.491	2.645	2.895	3.429
74	0.254	0.527	0.847	1.044	1.293	1.666	1.993	2.378	2.490	2.644	2.894	3.427
75	0.254	0.527	0.846	1.044	1.293	1.665	1.992	2.377	2.490	2.643	2.892	3.425
76	0.254	0.527	0.846	1.044	1.293	1.665	1.992	2.376	2.489	2.642	2.891	3.423
77	0.254	0.527	0.846	1.043	1.293	1.665	1.991	2.376	2.488	2.641	2.890	3.421
78	0.254	0.527	0.846	1.043	1.292	1.665	1.991	2.375	2.487	2.640	2.889	3.420
79	0.254	0.527	0.846	1.043	1.292	1.664	1.990	2.374	2.487	2.640	2.888	3.418
80	0.254	0.526	0.846	1.043	1.292	1.664	1.990	2.374	2.486	2.639	2.887	3.416
81	0.254	0.526	0.846	1.043	1.292	1.664	1.990	2.373	2.485	2.638	2.886	3.415
82	0.254	0.526	0.846	1.043	1.292	1.664	1.989	2.373	2.485	2.637	2.885	3.413
83	0.254	0.526	0.846	1.043	1.292	1.663	1.989	2.372	2.484	2.636	2.884	3.412
84	0.254	0.526	0.846	1.043	1.292	1.663	1.989	2.372	2.483	2.636	2.883	3.410
85	0.254	0.526	0.846	1.043	1.292	1.663	1.988	2.371	2.483	2.635	2.882	3.409
86	0.254	0.526	0.846	1.043	1.291	1.663	1.988	2.370	2.482	2.634	2.881	3.407
87	0.254	0.526	0.846	1.043	1.291	1.663	1.988	2.370	2.482	2.634	2.880	3.406
88	0.254	0.526	0.846	1.043	1.291	1.662	1.987	2.369	2.481	2.633	2.880	3.405
89	0.254	0.526	0.846	1.043	1.291	1.662	1.987	2.369	2.481	2.632	2.879	3.403
90	0.254	0.526	0.846	1.042	1.291	1.662	1.987	2.368	2.480	2.632	2.878	3.402
91	0.254	0.526	0.846	1.042	1.291	1.662	1.986	2.368	2.479	2.631	2.877	3.401
92	0.254	0.526	0.846	1.042	1.291	1.662	1.986	2.368	2.479	2.630	2.876	3.399
93	0.254	0.526	0.846	1.042	1.291	1.661	1.986	2.367	2.478	2.630	2.876	3.398
94	0.254	0.526	0.845	1.042	1.291	1.661	1.986	2.367	2.478	2.629	2.875	3.397
95	0.254	0.526	0.845	1.042	1.291	1.661	1.985	2.366	2.477	2.629	2.874	3.396
96	0.254	0.526	0.845	1.042	1.290	1.661	1.985	2.366	2.477	2.628	2.873	3.395
97	0.254	0.526	0.845	1.042	1.290	1.661	1.985	2.365	2.476	2.627	2.873	3.394
98	0.254	0.526	0.845	1.042	1.290	1.661	1.984	2.365	2.476	2.627	2.872	3.393
99	0.254	0.526	0.845	1.042	1.290	1.660	1.984	2.365	2.476	2.626	2.871	3.392
100	0.254	0.526	0.845	1.042	1.290	1.660	1.984	2.364	2.475	2.626	2.871	3.390
200	0.254	0.525	0.843	1.039	1.286	1.653	1.972	2.345	2.454	2.601	2.839	3.340
400	0.254	0.525	0.843	1.038	1.284	1.649	1.966	2.336	2.443	2.588	2.823	3.315
1000	0.253	0.525	0.842	1.037	1.282	1.646	1.962	2.330	2.437	2.581	2.813	3.300
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.326	2.432	2.576	2.807	3.291

Table V: Values of χ^2_α of χ^2 -distribution



df	α : Area to the Right of χ^2_α									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.0 ⁴ 393	0.0 ³ 157	0.0 ³ 982	0.0 ² 393	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table VI: Values of F_α of F -distribution



α	df_n									
	1	2	3	4	5	6	7	8	9	10
$df_d = 1$										
0.5	1.000	1.500	1.709	1.823	1.894	1.942	1.977	2.004	2.025	2.042
0.1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2
0.05	161	199	216	225	230	234	237	239	241	242
0.025	648	799	864	900	922	937	948	957	963	969
0.01	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
0.005	16211	19999	21615	22500	23056	23437	23715	23925	24091	24224
0.001	405284	499999	540379	562500	576405	585937	592873	598144	602284	605621
$df_d = 2$										
0.5	0.667	1.000	1.135	1.207	1.252	1.282	1.305	1.321	1.334	1.345
0.1	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
0.005	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40
0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40
$df_d = 3$										
0.5	0.585	0.881	1.000	1.063	1.102	1.129	1.148	1.163	1.174	1.183
0.1	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
0.05	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
0.005	55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7
0.001	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2
$df_d = 4$										
0.5	0.549	0.828	0.941	1.000	1.037	1.062	1.080	1.093	1.104	1.113
0.1	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
0.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
0.005	31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0
0.001	74.1	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.5	48.1
$df_d = 5$										
0.5	0.528	0.799	0.907	0.965	1.000	1.024	1.041	1.055	1.065	1.073
0.1	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
0.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
0.005	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6
0.001	47.2	37.1	33.2	31.1	29.8	28.8	28.2	27.6	27.2	26.9

α	df_n									
	11	12	13	14	15	20	30	60	120	∞
$df_d = 1$										
0.5	2.06	2.07	2.08	2.09	2.09	2.12	2.15	2.17	2.18	2.20
0.1	60.5	60.7	60.9	61.1	61.2	61.7	62.3	62.8	63.1	63.3
0.05	243	244	245	245	246	248	250	252	253	254
0.025	973	977	980	983	985	993	1001	1010	1014	1018
0.01	6083	6106	6126	6143	6157	6209	6261	6313	6339	6366
0.005	24334	24426	24505	24572	24630	24836	25044	25253	25359	25464
0.001	608368	610668	612622	614303	615764	620908	626099	631337	633972	636619
$df_d = 2$										
0.5	1.354	1.361	1.367	1.372	1.377	1.393	1.410	1.426	1.434	1.443
0.1	9.401	9.408	9.415	9.420	9.425	9.441	9.458	9.475	9.483	9.491
0.05	19.40	19.41	19.42	19.42	19.43	19.45	19.46	19.48	19.49	19.50
0.025	39.41	39.41	39.42	39.43	39.43	39.45	39.46	39.48	39.49	39.50
0.01	99.41	99.42	99.42	99.43	99.43	99.45	99.47	99.48	99.49	99.50
0.005	199.41	199.42	199.42	199.43	199.43	199.45	199.47	199.48	199.49	199.50
0.001	999.41	999.42	999.42	999.43	999.43	999.45	999.47	999.48	999.49	999.50
$df_d = 3$										
0.5	1.191	1.197	1.203	1.207	1.211	1.225	1.239	1.254	1.261	1.268
0.1	5.222	5.216	5.210	5.205	5.200	5.184	5.168	5.151	5.143	5.134
0.05	8.763	8.745	8.729	8.715	8.703	8.660	8.617	8.572	8.549	8.526
0.025	14.37	14.34	14.30	14.28	14.25	14.17	14.08	13.99	13.95	13.90
0.01	27.13	27.05	26.98	26.92	26.87	26.69	26.50	26.32	26.22	26.13
0.005	43.5	43.4	43.3	43.2	43.1	42.8	42.5	42.1	42.0	41.8
0.001	128.7	128.3	128.0	127.6	127.4	126.4	125.4	124.5	124.0	123.5
$df_d = 4$										
0.5	1.120	1.126	1.131	1.135	1.139	1.152	1.165	1.178	1.185	1.192
0.1	3.91	3.90	3.89	3.88	3.87	3.84	3.82	3.79	3.78	3.76
0.05	5.94	5.91	5.89	5.87	5.86	5.80	5.75	5.69	5.66	5.63
0.025	8.79	8.75	8.71	8.68	8.66	8.56	8.46	8.36	8.31	8.26
0.01	14.5	14.4	14.3	14.2	14.2	14.0	13.8	13.7	13.6	13.5
0.005	20.8	20.7	20.6	20.5	20.4	20.2	19.9	19.6	19.5	19.3
0.001	47.7	47.4	47.2	46.9	46.8	46.1	45.4	44.7	44.4	44.1
$df_d = 5$										
0.5	1.080	1.085	1.090	1.094	1.098	1.111	1.123	1.136	1.143	1.149
0.1	3.28	3.27	3.26	3.25	3.24	3.21	3.17	3.14	3.12	3.10
0.05	4.70	4.68	4.66	4.64	4.62	4.56	4.50	4.43	4.40	4.36
0.025	6.57	6.52	6.49	6.46	6.43	6.33	6.23	6.12	6.07	6.02
0.01	10.0	9.9	9.8	9.8	9.7	9.6	9.4	9.2	9.1	9.0
0.005	13.5	13.4	13.3	13.2	13.1	12.9	12.7	12.4	12.3	12.1
0.001	26.6	26.4	26.2	26.1	25.9	25.4	24.9	24.3	24.1	23.8

α	df_n									
	1	2	3	4	5	6	7	8	9	10
$df_d = 6$										
0.5	0.515	0.780	0.886	0.942	0.977	1.000	1.017	1.030	1.040	1.048
0.1	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
0.05	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
0.025	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461
0.01	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
0.005	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.391	10.250
0.001	35.507	27.000	23.703	21.924	20.803	20.030	19.463	19.030	18.688	18.411
$df_d = 7$										
0.5	0.506	0.767	0.871	0.926	0.960	0.983	1.000	1.013	1.022	1.030
0.1	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
0.05	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
0.025	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761
0.01	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
0.005	16.236	12.404	10.882	10.050	9.522	9.155	8.885	8.678	8.514	8.380
0.001	29.245	21.689	18.772	17.198	16.206	15.521	15.019	14.634	14.330	14.083
$df_d = 8$										
0.5	0.499	0.757	0.860	0.915	0.948	0.971	0.988	1.000	1.010	1.018
0.1	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
0.05	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
0.025	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295
0.01	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
0.005	14.688	11.042	9.596	8.805	8.302	7.952	7.694	7.496	7.339	7.211
0.001	25.415	18.494	15.829	14.392	13.485	12.858	12.398	12.046	11.767	11.540
$df_d = 9$										
0.5	0.494	0.749	0.852	0.906	0.939	0.962	0.978	0.990	1.000	1.008
0.1	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416
0.05	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
0.025	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964
0.01	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
0.005	13.614	10.107	8.717	7.956	7.471	7.134	6.885	6.693	6.541	6.417
0.001	22.857	16.387	13.902	12.560	11.714	11.128	10.698	10.368	10.107	9.894
$df_d = 10$										
0.5	0.490	0.743	0.845	0.899	0.932	0.954	0.971	0.983	0.992	1.000
0.1	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
0.05	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
0.025	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717
0.01	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
0.005	12.826	9.427	8.081	7.343	6.872	6.545	6.302	6.116	5.968	5.847
0.001	21.040	14.905	12.553	11.283	10.481	9.926	9.517	9.204	8.956	8.754

α	df_n									
	11	12	13	14	15	20	30	60	120	∞
$df_d = 6$										
0.5	1.054	1.060	1.065	1.069	1.072	1.084	1.097	1.109	1.116	1.122
0.1	2.920	2.905	2.892	2.881	2.871	2.836	2.800	2.762	2.742	2.722
0.05	4.027	4.000	3.976	3.956	3.938	3.874	3.808	3.740	3.705	3.669
0.025	5.410	5.366	5.329	5.297	5.269	5.168	5.065	4.959	4.904	4.849
0.01	7.790	7.718	7.657	7.605	7.559	7.396	7.229	7.057	6.969	6.880
0.005	10.133	10.034	9.950	9.877	9.814	9.589	9.358	9.122	9.001	8.879
0.001	18.182	17.989	17.824	17.682	17.559	17.120	16.672	16.214	15.981	15.745
$df_d = 7$										
0.5	1.037	1.042	1.047	1.051	1.054	1.066	1.079	1.091	1.097	1.103
0.1	2.684	2.668	2.654	2.643	2.632	2.595	2.555	2.514	2.493	2.471
0.05	3.603	3.575	3.550	3.529	3.511	3.445	3.376	3.304	3.267	3.230
0.025	4.709	4.666	4.628	4.596	4.568	4.467	4.362	4.254	4.199	4.142
0.01	6.538	6.469	6.410	6.359	6.314	6.155	5.992	5.824	5.737	5.650
0.005	8.270	8.176	8.097	8.028	7.968	7.754	7.534	7.309	7.193	7.076
0.001	13.879	13.707	13.561	13.434	13.324	12.932	12.530	12.119	11.909	11.696
$df_d = 8$										
0.5	1.024	1.029	1.034	1.038	1.041	1.053	1.065	1.077	1.083	1.089
0.1	2.519	2.502	2.488	2.475	2.464	2.425	2.383	2.339	2.316	2.293
0.05	3.313	3.284	3.259	3.237	3.218	3.150	3.079	3.005	2.967	2.928
0.025	4.243	4.200	4.162	4.130	4.101	3.999	3.894	3.784	3.728	3.670
0.01	5.734	5.667	5.609	5.559	5.515	5.359	5.198	5.032	4.946	4.859
0.005	7.104	7.015	6.938	6.872	6.814	6.608	6.396	6.177	6.065	5.951
0.001	11.352	11.194	11.060	10.943	10.841	10.480	10.109	9.727	9.532	9.334
$df_d = 9$										
0.5	1.014	1.019	1.024	1.028	1.031	1.043	1.055	1.067	1.073	1.079
0.1	2.396	2.379	2.364	2.351	2.340	2.298	2.255	2.208	2.184	2.159
0.05	3.102	3.073	3.048	3.025	3.006	2.936	2.864	2.787	2.748	2.707
0.025	3.912	3.868	3.831	3.798	3.769	3.667	3.560	3.449	3.392	3.333
0.01	5.178	5.111	5.055	5.005	4.962	4.808	4.649	4.483	4.398	4.311
0.005	6.314	6.227	6.153	6.089	6.032	5.832	5.625	5.410	5.300	5.188
0.001	9.718	9.570	9.443	9.334	9.238	8.898	8.548	8.187	8.001	7.813
$df_d = 10$										
0.5	1.006	1.012	1.016	1.020	1.023	1.035	1.047	1.059	1.064	1.070
0.1	2.302	2.284	2.269	2.255	2.244	2.201	2.155	2.107	2.082	2.055
0.05	2.943	2.913	2.887	2.865	2.845	2.774	2.700	2.621	2.580	2.538
0.025	3.665	3.621	3.583	3.550	3.522	3.419	3.311	3.198	3.140	3.080
0.01	4.772	4.706	4.650	4.601	4.558	4.405	4.247	4.082	3.996	3.909
0.005	5.746	5.661	5.589	5.526	5.471	5.274	5.071	4.859	4.750	4.639
0.001	8.586	8.445	8.324	8.220	8.129	7.804	7.469	7.122	6.944	6.762

α	df_n									
	1	2	3	4	5	6	7	8	9	10
$df_d = 11$										
0.5	0.486	0.739	0.840	0.893	0.926	0.948	0.964	0.977	0.986	0.994
0.1	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
0.05	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
0.025	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526
0.01	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
0.005	12.226	8.912	7.600	6.881	6.422	6.102	5.865	5.682	5.537	5.418
0.001	19.687	13.812	11.561	10.346	9.578	9.047	8.655	8.355	8.116	7.922
$df_d = 12$										
0.5	0.484	0.735	0.835	0.888	0.921	0.943	0.959	0.972	0.981	0.989
0.1	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
0.05	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
0.025	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374
0.01	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
0.005	11.754	8.510	7.226	6.521	6.071	5.757	5.525	5.345	5.202	5.085
0.001	18.643	12.974	10.804	9.633	8.892	8.379	8.001	7.710	7.480	7.292
$df_d = 13$										
0.5	0.481	0.731	0.832	0.885	0.917	0.939	0.955	0.967	0.977	0.984
0.1	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
0.05	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
0.025	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250
0.01	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
0.005	11.374	8.186	6.926	6.233	5.791	5.482	5.253	5.076	4.935	4.820
0.001	17.815	12.313	10.209	9.073	8.354	7.856	7.489	7.206	6.982	6.799
$df_d = 14$										
0.5	0.479	0.729	0.828	0.881	0.914	0.936	0.952	0.964	0.973	0.981
0.1	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
0.05	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
0.025	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147
0.01	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
0.005	11.060	7.922	6.680	5.998	5.562	5.257	5.031	4.857	4.717	4.603
0.001	17.143	11.779	9.729	8.622	7.922	7.436	7.077	6.802	6.583	6.404
$df_d = 15$										
0.5	0.478	0.726	0.826	0.878	0.911	0.933	0.949	0.960	0.970	0.977
0.1	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
0.05	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
0.025	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060
0.01	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
0.005	10.798	7.701	6.476	5.803	5.372	5.071	4.847	4.674	4.536	4.424
0.001	16.587	11.339	9.335	8.253	7.567	7.092	6.741	6.471	6.256	6.081

α	df_n									
	11	12	13	14	15	20	30	60	120	∞
$df_d = 11$										
0.5	1.000	1.005	1.010	1.013	1.017	1.028	1.040	1.052	1.058	1.064
0.1	2.227	2.209	2.193	2.179	2.167	2.123	2.076	2.026	2.000	1.972
0.05	2.818	2.788	2.761	2.739	2.719	2.646	2.570	2.490	2.448	2.404
0.025	3.474	3.430	3.392	3.359	3.330	3.226	3.118	3.004	2.944	2.883
0.01	4.462	4.397	4.342	4.293	4.251	4.099	3.941	3.776	3.690	3.602
0.005	5.320	5.236	5.165	5.103	5.049	4.855	4.654	4.445	4.337	4.226
0.001	7.761	7.626	7.509	7.409	7.321	7.008	6.684	6.348	6.175	5.998
$df_d = 12$										
0.5	0.995	1.000	1.004	1.008	1.012	1.023	1.035	1.046	1.052	1.058
0.1	2.166	2.147	2.131	2.117	2.105	2.060	2.011	1.960	1.932	1.904
0.05	2.717	2.687	2.660	2.637	2.617	2.544	2.466	2.384	2.341	2.296
0.025	3.321	3.277	3.239	3.206	3.177	3.073	2.963	2.848	2.787	2.725
0.01	4.220	4.155	4.100	4.052	4.010	3.858	3.701	3.535	3.449	3.361
0.005	4.988	4.906	4.836	4.775	4.721	4.530	4.331	4.123	4.015	3.904
0.001	7.136	7.005	6.892	6.794	6.709	6.405	6.090	5.762	5.593	5.420
$df_d = 13$										
0.5	0.990	0.996	1.000	1.004	1.007	1.019	1.030	1.042	1.048	1.054
0.1	2.116	2.097	2.080	2.066	2.053	2.007	1.958	1.904	1.876	1.846
0.05	2.635	2.604	2.577	2.554	2.533	2.459	2.380	2.297	2.252	2.206
0.025	3.197	3.153	3.115	3.082	3.053	2.948	2.837	2.720	2.659	2.595
0.01	4.025	3.960	3.905	3.857	3.815	3.665	3.507	3.341	3.255	3.165
0.005	4.724	4.643	4.573	4.513	4.460	4.270	4.073	3.866	3.758	3.647
0.001	6.647	6.519	6.409	6.314	6.231	5.934	5.626	5.305	5.138	4.967
$df_d = 14$										
0.5	0.987	0.992	0.996	1.000	1.003	1.015	1.026	1.038	1.044	1.050
0.1	2.073	2.054	2.037	2.022	2.010	1.962	1.912	1.857	1.828	1.797
0.05	2.565	2.534	2.507	2.484	2.463	2.388	2.308	2.223	2.178	2.131
0.025	3.095	3.050	3.012	2.979	2.949	2.844	2.732	2.614	2.552	2.487
0.01	3.864	3.800	3.745	3.698	3.656	3.505	3.348	3.181	3.094	3.004
0.005	4.508	4.428	4.359	4.299	4.247	4.059	3.862	3.655	3.547	3.436
0.001	6.256	6.130	6.023	5.930	5.848	5.557	5.254	4.938	4.773	4.604
$df_d = 15$										
0.5	0.983	0.989	0.993	0.997	1.000	1.011	1.023	1.034	1.040	1.046
0.1	2.037	2.017	2.000	1.985	1.972	1.924	1.873	1.817	1.787	1.755
0.05	2.507	2.475	2.448	2.424	2.403	2.328	2.247	2.160	2.114	2.066
0.025	3.008	2.963	2.925	2.891	2.862	2.756	2.644	2.524	2.461	2.395
0.01	3.730	3.666	3.612	3.564	3.522	3.372	3.214	3.047	2.959	2.868
0.005	4.329	4.250	4.181	4.122	4.070	3.883	3.687	3.480	3.372	3.260
0.001	5.935	5.812	5.707	5.615	5.535	5.248	4.950	4.638	4.475	4.307

α	df_n									
	1	2	3	4	5	6	7	8	9	10
$df_d = 20$										
0.5	0.472	0.718	0.816	0.868	0.900	0.922	0.938	0.950	0.959	0.966
0.1	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
0.05	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
0.025	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774
0.01	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
0.005	9.944	6.986	5.818	5.174	4.762	4.472	4.257	4.090	3.956	3.847
0.001	14.819	9.953	8.098	7.096	6.461	6.019	5.692	5.440	5.239	5.075
$df_d = 30$										
0.5	0.466	0.709	0.807	0.858	0.890	0.912	0.927	0.939	0.948	0.955
0.1	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
0.05	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
0.025	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511
0.01	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
0.005	9.180	6.355	5.239	4.623	4.228	3.949	3.742	3.580	3.450	3.344
0.001	13.293	8.773	7.054	6.125	5.534	5.122	4.817	4.581	4.393	4.239
$df_d = 60$										
0.5	0.460	0.701	0.798	0.849	0.880	0.901	0.917	0.928	0.937	0.945
0.1	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707
0.05	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
0.025	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270
0.01	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
0.005	8.495	5.795	4.729	4.140	3.760	3.492	3.291	3.134	3.008	2.904
0.001	11.973	7.768	6.171	5.307	4.757	4.372	4.086	3.865	3.687	3.541
$df_d = 120$										
0.5	0.458	0.697	0.793	0.844	0.875	0.896	0.912	0.923	0.932	0.939
0.1	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652
0.05	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910
0.025	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222	2.157
0.01	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
0.005	8.179	5.539	4.497	3.921	3.548	3.285	3.087	2.933	2.808	2.705
0.001	11.380	7.321	5.781	4.947	4.416	4.044	3.767	3.552	3.379	3.237
$df_d = \infty$										
0.5	0.455	0.693	0.789	0.839	0.870	0.891	0.907	0.918	0.927	0.934
0.1	2.706	2.303	2.084	1.945	1.847	1.774	1.717	1.670	1.632	1.599
0.05	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831
0.025	5.024	3.689	3.116	2.786	2.567	2.408	2.288	2.192	2.114	2.048
0.01	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321
0.005	7.879	5.298	4.279	3.715	3.350	3.091	2.897	2.744	2.621	2.519
0.001	10.828	6.908	5.422	4.617	4.103	3.743	3.475	3.266	3.097	2.959

Appendix C: Lab Manual

[STAT 151 Lab Manual in R Commander Version 3](#)

Appendix D: Image Descriptions

Figure 1.1 Image Description: A larger ellipse in blue is labelled “Population.” Inside the blue ellipse is a smaller, white ellipse outlined in orange labelled “Sample.” The smaller ellipse is entirely inside the larger, showing that a sample is a portion of a population. [[Return to Figure 1.1](#)]

Table 1.1 Image Description: The table has 50 columns and 20 rows. The columns are divided into groups of 10 and the rows are in groups of 5. Each cell has a random number from 0 to 9. The number 82 at the intersection of 05 with 00 and 01 is highlighted, and a set of red arrows shows the numbers selected after 82. [[Return to Table 1.1](#)]

Snapshot 1.1 Image Description: A screenshot of the R commander window. The section for input, called R Script, has the following lines: “set dot seed (4061) line-break sample (1:10 comma 10).” The output window repeats the above lines with a new line underneath that says “[1] 53 14 57 13 8 45 11 50 59 25.” [[Return to Snapshot 1.1](#)]

Snapshot 1.2 Image Description: A screenshot of the R commander window. The section for input, called R Script, has the following lines: “set dot seed (6194) line-break sample (1:100 comma 5).” The output window repeats the above lines with a new line underneath that says “[1] 59 9 1 40 77.” [[Return to Snapshot 1.2](#)]

Figure 1. 2 image description: A bar graph in the left panel shows the relative frequency of how students came to school. The vertical axis marks relative frequency ranging from 0 to 0.5 in intervals of 0.1. The horizontal axis shows the categories of variable “Transport”. The values are car (a relative frequency of 0.194), public (at 0.498), bike (at 0.033), walk (at 0.271), and others (and 0.004). The right panel shows a pie chart on the right panel showing percentages of the same data. One slice for one category. Clockwise from the top, the slices show that 19.4% of students by car, 0.4% by other means, 27.1% walked, 3.3% by bike, and 49.8% by public transportation. [[Return to Figure 1.2](#)]

Figure 1.3 Image Description: The left panel is a pie chart showing percentages of how female students came to school. Clockwise from the top, the slices show that 16.9% of female students by car, 0% by other means, 25.7% walked, 3.4% by bike, and 54.1% by public transportation. The right panel is also a pie chart showing percentages of how male students came to school. Clockwise from the top, the slices show that 22.4% of male students by car, 0.8% by other means, 28.8% walked, 3.2% by bike, and 44.8% by public transportation. [[Return to Figure 1.3](#)]

Figure 1.4 Image Description: A side-by-side bar graph comparing relative frequency of how female and male students came to school. The vertical axis marks relative frequency ranging from 0 to 1 in intervals of 0.2. The horizontal axis shows the categories of variable “Transport”. The values are car (at 0.169 for females and 0.224 for males), public (at 0.541 for females and 0.448 for males), bike (at 0.034 for females and 0.032 for males), walk (at 0.257 for females and 0.288 for males), others (at 0 for females and 0.008 for males). [[Return to Figure 1.4](#)]

Figure 1.5 Image Description: A histogram of number of siblings. The y-axis (vertical) is frequency, going from 0 to 35 in increments of 5. The x-axis (horizontal) is the number of siblings with labelled values of 0, 1, 2, 3, and larger than 3. Five bars are shown as follows: x at 0 has a height of 10, x at 1 has a height of 30, x at 2 has a height of 35, x at 3 has a height of 15, and x at more than 3 has a height of 10. [[Return to Figure 1.5](#)]

Figure 1.6 Image Description: A histogram of grade. The y-axis is frequency, going from 0 to 14 in increments of 2. The x-axis is grades from 0 to 100 with 10 bars taking up intervals of width 10. The heights of the bars are as follows: x at 0 to 10 has a height of 1, x at 10 to 20 has a height of 0, x at 20 to 30 has a height of 2, x at 30 to 40 has a height of 4, x at 40 to 50 has a height of 8, x at 50 to 60 has a height of 14, x at 60 to 70 has a height of 10, x at 70 to 80 has a height of 6, x at 80 to 90 has a height of 3, and x at 90 to 100 has a height of 2. [[Return to Figure 1.6](#)]

Figure 1.7 Image Description: A stem and leaf diagram of grade. The stems are listed vertically to the left of a vertical line, starting from 0 to 9. The leaves of each stem are listed horizontally on the right of the vertical line. The leaf of 0 is 9, no leaf for stem 1. Leaves of 2 are 4 and 9, leaves of 3 are 4, 7, 7, and 9. Leaves of 4 are 0, 2, 6, 7, 8, 8, 9, 9. Leaves of 5 are 0, 1, 2, 3, 4, 5, 5, 5, 6, 6, 7, 9, 9 and 9. Leaves of 6 are 0, 0, 1, 2, 5, 5, 8, 8, 8, and 8. Leaves of 7 are 1, 2, 4, 5, 6, and 9. Leaves of 8 are 1, 3 and 5. The last row are the leaves of 9: 0 and 2. [[Return to Figure 1.7](#)]

Figure 1.8 Image Description: Nine special shapes of distributions presented in three rows and three columns. The three figures in the first row are as follows: figure (a) is a bell-shape curve, (b) is an isosceles triangle, and (c) is a rectangle called uniform distribution. The three figures in the second row are as follows: figure (d) is non-symmetrical curve does have a peak and a longer tail at the right end, it is called right skewed; figure (e) is non-symmetrical curve does have a peak and a longer tail at the left end, it is called left skewed; and figure (f) is non-symmetrical and slopes upward continually, looks like a capital letter J. The three figures on the third row are as follows: figure (g) is non-symmetrical and slopes downward continually, looks like a reversed capital letter J; figure (h) is a symmetrical curve with two peaks, it is called bimodal distribution; and figure (i) is a non-symmetrical curve with three peaks, it is called multi-modal. [[Return to Figure 1.8](#)]

Figure 1.8.1 Image Description: A histogram of grade. The y-axis is frequency from 0 to 14 in increments of 2. The x-axis is grades from 0 to 100 with 10 bars taking up intervals of width 10. The heights of the bars are as follows: x at 0 to 10 has a height of 1, x at 10 to 20 has a height of 0, x at 20 to 30 has a height of 2, x at 30 to 40 has a height of 4, x at 40 to 50 has a height of 8, x at 50 to 60 has a height of 14, x at 60 to 70 has a height of 10, x at 70 to 80 has a height of 6, x at 80 to 90 has a height of 3, and x at 90 to 100 has a height of 2. [[Return to Figure 1.8.1](#)]

Figure 1.9 Image Description: A histogram of survival time after diagnosis of cancer. The y-axis is frequency from 0 to 1000 in increments of 200. The x-axis is survival time in years from 0 to 32 with 32 bars in increments of 1 year. The heights of the bars are as follows: x at 1 is close to 550, x at 2 is close to 1000, x at 3 is close to 1050 (this is the peak of the histogram), and x at 4 is close to 900. Following 4, the height of the bars decreases as x increases. The bars after x at 15 have a height very close to zero. [[Return to Figure 1.9](#)]

Figure 1.10 Image Description: A histogram of salary. The y-axis is frequency from 0 to 8 in increments of 2. The x-axis is salary from 20 to 60 with 9 bars or increment 5. The heights of the bars are as follows: x at 20 to 25 has a height of 1, x at 25 to 30 has a height of 3, x at 30 to 35 has a height of 8, x at 35 to 40 has a height of 7, x at 40 to 45 has a height of 5, x at 45 to 50 has a height of 9, x at 50 to 55 has a height of 7, x at 55 to 60 has a height of 1, and x at 60 to 65 has a height of 1. [[Return to Figure 1.10](#)]

Assignment 1 Question 2 Image Description: The table shows the first thirty entries of the home sale spreadsheet. Each row is numbered, representing a single house. The columns are labelled "Size" in square feet, "Pool" yes or no, "Area" in square feet, "Age" in years, "Bath" in number of bathrooms, "Stories" in number of stories, "Garage" yes or no, "Traffic" yes or no, "Roof" tile or non-tile, and "Price" in dollars. The first entry (number one) has the following data: Size is 1865, Area is 9509.4, Age is 18, Bath is 2.5, Stories is 1, Garage is 2, Traffic is no, Roof is non-tile, and Price is 145950. For more entries, please download M01_SaleHome.xlsx from the top of the assignment page. [[Return to Question 2](#)]

Figure 2.1 Image Description: Three histograms in a row. The histogram on the left panel is roughly symmetric, the mean (red solid vertical line) and the median (the blue dashed vertical line) are almost identical. The histogram in the middle is right skewed with a longer tail on the right, the mean (red solid vertical line) is on the right of the median (the blue dashed vertical line). The histogram on the right panel is left skewed with a longer tail on the left, the mean (red solid vertical line) is on the left of the median (the blue dashed vertical line). [[Return to Figure 2.1](#)]

Figure 2.2 Image Description: A histogram of survival time after diagnosis of cancer. The y-axis is frequency from 0 to 1000 in increments of 200. The x-axis is survival time in years

from 0 to 32 with 32 bars in increments of 1 year. The heights of the bars are as follows: x at 1 is close to 550, x at 2 is close to 1000, x at 3 is close to 1050 (this is the peak of the histogram), and x at 4 is close to 900. Following 4, the height of the bars decreases as x increases. The bars after x at 15 have a height very close to zero. [\[Return to Figure 2.2\]](#)

Figure 2.3 Image Description: A vertical box plot for the given data. The y-axis is in increments of 5 from 0 to 20. The boxplot's bottom adjacent value is 1, the smallest value within the lower and upper limits. The lower whisker (a short dashed line) extends from the bottom adjacent value, 1, to the first quartile which is 4. The box begins at the first quartile and extends to the third quartile which is 10. A horizontal line is drawn at the median which is 7. And the upper whisker (a short dashed line) extends from the third quartile which is 10 to the top adjacent value which is 11. The observation 21 is an outlier indicated as a circle. [\[Return to Figure 2.3\]](#)

Figure 2.4 Image Description: Three box plots with the same y-axis in a row. The y-axis is in increments of 0.2 from 0 to 1. The leftmost panel is a box plot of a right skewed distribution ranging from 0 to 0.5. The lower whisker (a dashed line extending from the smallest observation to the first quartile) is shorter than the upper whisker (a dash line extending from the third quartile to the upper adjacent value). The distance between the first quartile and the median is also shorter than the distance between the median and the third quartile. There are also several outliers on the top. The middle panel presents a box plot of a symmetric distribution ranging from 0.2 to 0.8. The lower whisker and the upper whisker are roughly of the same length. The distance between the first quartile and the median and the distance between the median and the third quartile are roughly the same. There is one outlier on the top. The rightmost panel presents a box plot of a left skewed distribution ranging from 0.5 to 1. The lower whisker is longer than the upper whisker. The distance between the first quartile and the median is also larger than the distance between the median and the third quartile. There are also several outliers at the bottom. [\[Return to Figure 2.4\]](#)

Figure 2.5 Image Description: A pair of boxplots, titled "Boxplot of Non-attendees & Attendees". The y-axis is labelled "Final grade" and is in increments of 5 from 35 to 95. There are two boxplots represented here, the one on the left is labelled "Non-attendee" and the one on the right is labelled "Attendee". There are no outliers for either group. The range of the boxplot for "Non-attendee" is roughly between 35 and 87 with a median of 65, and the box (IQR) goes from about 52 to 78. The range of the box plot for "Attendee" is roughly between 48 and 97 with a median of 78, and the box (IQR) goes from about 70 to 85. [\[Return to Figure 2.5\]](#)

Figure 2.6 Image Description: A vertical box plot for the given data. The y-axis is in increments of 1 from negative 5 to 1. The boxplot's bottom adjacent value is 0.05, the

smallest value within the lower and upper limits. The lower whisker extends from the bottom adjacent value, 0.05, to the first quartile which is 0.2. The box begins at the first quartile and extends to the third quartile which is 0.7. A horizontal line is drawn at the median which is indicated as Q_2 equals 0.5. And the upper whisker goes from the third quartile which is 0.7 to the top adjacent value which is 0.95. The observation negative 5 is an outlier indicated as a circle at the bottom. [[Return to Figure 2.6](#)]

Figure 3.1 Image Description: A scatter plot with y-axis labelled “Proportion of Heads” and x-axis labelled “# of Experiments”. The y-axis is in increments of 0.02 from 0.44 to 0.54 and the x-axis is in increments of 20000 from 0 to 100000. As the number of experiments increases the proportion of heads starts with 0.45 and then goes up to 0.5 and then fluctuates under a red horizontal dashed line at 0.5. The points are getting closer to 0.5 when the number of experiments goes above 70000. [[Return to Figure 3.1](#)]

Figure 3.2 Image Description: Four venn diagrams are presented in a 2 by 2 matrix. Each venn diagram has a rectangle with a label “S” on the top-left corner representing the sample space. The top-left venn diagram is labelled “E” under the rectangle and has an oval in the center. The oval has a letter “E” in the center and is filled with light blue. The top-right venn diagram is labelled “not E” under the rectangle and has an oval in the center. The oval has a letter “E” in the center and the whole rectangle except the oval is filled with light blue. The bottom-left venn diagram is labelled “A & B” under the rectangle and has two overlapped ovals in the center, the left oval is labelled “A” and the right oval is labelled “B” and their overlap is labelled “A & B”. The overlap is filled with blue. The bottom-right venn diagram is labelled “A or B” under the rectangle and has two overlapped ovals in the center, the left oval is labelled “A” and the right oval is labelled “B”. Both two ovals and their overlap are filled with blue. [[Return to Figure 3.2](#)]

Figure 3.3 Image Description: A venn diagram shows event A as a subset of event B. A rectangle with a label “S” on the top-left corner represents the sample space. There is a big oval labelled “B” in the center of the rectangle and filled with light blue. There is a small oval labelled “A” inside the big oval “B” and filled with dark blue. [[Return to Figure 3.3](#)]

Figure 3.4 Image Description: This tree diagram grows from the left to the right and has two levels. The first level has two branches representing two possible outcomes of Midterm I. The upper branch is labelled “ $P(A_1)$ equals 0.15” above and “greater than or equal to 90” below the branch. The lower branch reads “ $P(B_1)$ equals 0.85” below and “greater than 90” above the branch. Both the upper and lower branches of the first level have two sub-branches representing the outcomes of Midterm II given the result of Midterm I. The topmost branch is connected to the first upper branch and reads “ $P(A_2 \text{ given } A_1)$ equals 0.8” above and “less than or equal to 90” below the sub-branch. The next branch down is connected to the first upper branch and reads “ $P(B_2 \text{ given } A_1)$ equals 0.2” below and

“greater than 90” above the sub-branch. The next branch down is connected to the first lower branch and reads “ $P(A_2 \text{ given } B_1) \text{ equals } 0.1$ ” above and “greater than or equal to 90” below the sub-branch. The bottom most branch is connected to the first lower branch and reads “ $P(B_2 \text{ given } B_1) \text{ equals } 0.9$ ” below and “greater than 90” above the sub-branch. The tree has four paths, the outcomes and their associated probabilities are listed in two columns to the right of the tree. The outcome of the first path is “ $A_2 \text{ \& } A_1$ ” with probability “ $0.15 \text{ times } 0.8 \text{ equals } 0.12$ ”. The outcome of the second path is “ $B_2 \text{ \& } A_1$ ” with probability “ $0.15 \text{ times } 0.2 \text{ equals } 0.03$ ”. The outcome of the third path is “ $A_2 \text{ \& } B_1$ ” with probability “ $0.85 \text{ times } 0.1 \text{ equals } 0.085$ ”. The outcome of the fourth path is “ $B_2 \text{ \& } B_1$ ” with probability “ $0.85 \text{ times } 0.9 \text{ equals } 0.765$ ”. [\[Return to Figure 3.4\]](#)

Example 3.1 Image Description: This tree diagram grows from the left to the right and has two levels. The first level has two branches representing two possible outcomes of smoking status. The upper first branch reads “ $P(S) \text{ equals } 0.2$ ” above the branch. The lower first branch reads “ $P(\text{not } S)$ ” below the branch. Both the upper and lower branches of the first level have two sub-branches representing the outcomes of breast cancer status given their smoking status. The topmost branch is connected to the first upper branch and reads “ $P(B \text{ given } S) \text{ equals one-third}$ ” above the sub-branch. The next branch down is connected to the first upper branch and reads “ $P(\text{not } B \text{ given } S) \text{ equals two-thirds}$ ” below the sub-branch. The next branch down is connected to the first lower branch and reads “ $P(B \text{ given not } S) \text{ equals three-seventeenths}$ ” above the sub-branch. The bottom most branch is connected to the first lower branch and reads “ $P(\text{not } B \text{ given not } S) \text{ equals fourteen-seventeenths}$ ” below the sub-branch. The tree has four paths, the outcome events and their associated probabilities are listed in two columns to the right of the tree. The event of the first path is “ $B \text{ \& } S$ ” with probability “ $0.2 \text{ times one-third equals } 0.0667$ ”. The event of the second path is “ $\text{not } B \text{ \& } S$ ” with probability “ $0.2 \text{ times two-thirds equals } 0.1333$ ”. The event of the third path is “ $B \text{ \& not } S$ ” with probability “ $0.8 \text{ times three-seventeenths equals } 0.1412$ ”. The event of the fourth path is “ $\text{not } B \text{ \& not } S$ ” with probability “ $0.8 \text{ times fourteen-seventeenths equals } 0.6588$ ”. [\[Return to Example 3.1\]](#)

Figure 4.1 Image Description: A mapping for the random variable X , number of tails that occur if we flip a coin twice. There are two ovals. The one labelled S shows the possible outcomes of the two coin flips and the one labelled X shows all possible numeric values of the random variable x , the number of tails. The possible outcomes listed in S are: HH , HT , TH , and TT . The possible values listed in X are 0, 1, and 2. The value HH from S is connected to X with an arrow pointing to 0. The values HT and TH in S are connected to X an arrow each, both pointing to 1. The value TT in S is connected X with an arrow pointing to 2. [\[Return to Figure 4.1\]](#)

Figure 4.2 Image Description: A mapping for the random variable X , number of siblings

if we randomly select a student. There are two ovals. The one representing S shows the randomly selected student and the one labelled X shows all possible values of the random variable x , the number of siblings. The possible outcomes in S are the five selected students: Mark, John, Rebecca, Sarah, and Mary. The possible values of X are 0, 1, 2, and 3. Mark (in S) is connected to X with an arrow pointing to 0. John (in S) is connected to X with an arrow pointing to 1. Rebecca and Sarah (in S) are connected to X with an arrow each, both pointing to 2. Mary (in S) is connected to X with an arrow pointing to 3. [[Return to Figure 4.2](#)]

Figure 4.3 Image Description: A histogram of number of siblings. The y-axis is probability (i.e., relative frequency) from 0 to 0.4 in increments of 0.1. The x-axis is the number of siblings with bars at 0, 1, 2, and 3. The heights of the bars are as follows: x at 0 has a height of 0.2, x at 1 has a height of 0.2, x at 2 has a height of 0.4, and x at 3 has a height of 0.2. [[Return to Figure 4.3](#)]

Figure 5.1 Image Description: The graph on the left is titled “Probability Histogram of Grade”. The y-axis is Probability from 0 to 0.4 in increments of 0.1. The x-axis is Grades with 7 bars in increments of 10 starting at 30 going to 100. The heights of the bars are as follows: x between 30 and 40 has a height of 0.003, x between 40 and 50 has a height of 0.019, x between 50 and 60 has a height of 0.15, x between 60 and 70 has a height of 0.33, x between 70 and 80 has a height of 0.332, x between 80 and 90 has a height of 0.144 and x between 90 and 100 has a height of 0.022. This is the same as the values given in table 5.1. The three left-most bins are filled with black. The Density Curve of Grade on the right panel is almost identical to the Probability Histogram on the left panel except that the y-axis labelled “Density” in increments of 0.01 from 0 to 0.04. There is a red bell-shaped curve on the top of the bars in the density graph. [[Return to Figure 5.1](#)]

Figure 5.1.1 Image Description: Three identical density curves are presented in a row. The leftmost curve has the area to the left of a vertical line x equals a under the density curve shaded in grey. In the middle panel, the area to the right of a vertical line x equals b under the density curve is shaded in grey. The right panel has the area between vertical lines x equals a and x equals b (a less than b) under the density curve is shaded in grey. [[Return to Figure 5.1.1](#)]

Figure 5.2 Image Description: This graph shows three normal density curves. The x-axis labelled as “ z ” is in increments of 5 from negative 5 to 5; The y-axis labelled as “ $f(z)$ ” is in increments of 0.1 from 0 to 5. The black solid bell-shaped curve centred at 0 is the density curve for a normal distribution with mean 0 and standard deviation 2. The red dashed bell-shaped curve centred at 0 and taller than the black solid curve is the density curve for a normal distribution with mean 0 and standard deviation 1. The blue dotted bell-shaped curve centred at 4 and of the same shape as the red dashed curve is the density curve for a normal distribution with mean 4 and standard deviation 1. [[Return to Figure 5.2](#)]

Figure 5.3 Image Description: This figure illustrates the empirical rule of a normal distribution. The normal curve is shown over a horizontal axis. The axis is labelled “X” with points $\mu - 3\sigma$, $\mu - 2\sigma$, $\mu - \sigma$, μ , $\mu + \sigma$, $\mu + 2\sigma$, and $\mu + 3\sigma$. Two red vertical lines connect the axis to the curve at labelled points $\mu - 3\sigma$ and $\mu + 3\sigma$. A red horizontal line connects these two points on the curve and reads “99.74%” indicating that the area under the normal curve between these points contains 99.74% of observations. Two blue vertical lines connect the axis to the curve at labelled points $\mu - 2\sigma$ and $\mu + 2\sigma$. A blue horizontal line connects these two points on the curve and reads “95.44%” indicating that the area under the normal curve between these points contains 95.44% of observations. Two green vertical lines connect the axis to the curve at labelled points $\mu - \sigma$ and $\mu + \sigma$. A green line connects these two points on the curve and reads “68.26%” indicating that the area under the normal curve contains 68.26% of observations. [[Return to Figure 5.3](#)]

Figure 5.4 Image Description: There are two normal density curves: the one on the left is the normal density curve for a normal distribution with mean μ and standard deviation σ ; the one on the right is the normal density curve for a standard normal distribution with mean 0 and standard deviation 1. The normal curve has labelled points a and b, with the area between a and b under the curve shaded in grey. The standard curve has points labelled $(a - \mu) / \sigma$ and $(b - \mu) / \sigma$. The area between these points is shaded grey. Arrows point to both grey areas and they are labelled “equal areas.” [[Return to Figure 5.4](#)]

Figure 5.5 Image Description: Three normal density curves are shown on a single horizontal axis. The x-axis is in increments of 1 from negative 6 to 18. The green normal density curve has a mean μ equals negative 2 and a standard deviation σ equals 1. The red normal density curve has a mean μ equals 4 and a standard deviation σ equals 0.5; a smaller standard deviation makes the red density curve taller and slimmer than the green density curve. The blue normal density curve has a mean μ equals 10 and a standard deviation σ equals 2; a larger standard deviation makes the blue density curve flatter than the green density curve. All these three normal density curves can be converted to the standard normal density curve through the standardisation method z equals $(X - \mu) / \sigma$. The standard normal density curve with mean μ equals 0 and standard deviation σ equals 1 is in black and shown over a horizontal axis goes from negative 4 to 4 with increments of 1. The black curve has the same shape as the green normal density curve, since they have the same standard deviation. [[Return to Figure 5.5](#)]

Figure 5.6 Image Description: This figure shows part of the second page of Table II (area under the standard normal curve for positive z scores). The first column of the table (labelled “z”) gives the first decimal place of the z-score in increments of 0.1 from 0.0

to 1.9. The first row of the table reads “Second decimal place in z” and the second row gives the second decimal place of the z-score in increments of 0.01 from 0.00 to 0.09. The elements of the main body of the table are the area (in four decimal places) to the left of the corresponding z-scores under the standard normal curve. The graph also shows that the area to the left of 1.96 is 0.9750. [[Return to Figure 5.6](#)]

Figure 5.7 Image Description: Three standard normal density curves are presented in a row. The left-hand graph has the area to the left of a vertical line z equals 1.96 under the standard normal density curve shaded in grey. This area is labelled as equalling 0.975. The middle graph has the area to the right of a vertical line z equals 1.96 under the standard normal density curve shaded in grey. This area is labelled area equalling $1 - P(Z \text{ less than } 1.96) = 1 - 0.975 = 0.025$. The rightmost graph has the area between vertical lines z equals negative 1.96 and z equals 1.96 under the standard normal density curve shaded in grey. This area is labelled “area equals $P(\text{negative } 1.96 \text{ less than } Z \text{ less than } 1.96) = P(Z \text{ less than } 1.96) - P(Z \text{ less than negative } 1.96) = 0.975 - 0.025 = 0.95$ ”. [[Return to Figure 5.7](#)]

Figure 5.8 Image Description: Four standard normal density curves are presented in a row. Each curve is shown over a horizontal axis labelled “Z” and has a dashed vertical line at the center z equals 0. The first graph corresponds to question 1 in this example. It is titled “Area to the Left of negative 2: $P(Z \text{ less than negative } 2)$ ”. The area to the left of a vertical line z equals negative 2 under the standard normal density curve is shaded in grey. The second graph corresponds to question 2 in this example. It is titled “Area to the Right of 2: $P(Z \text{ greater than } 2)$ ”. The area to the right of a vertical line z equals 2 under the standard normal density curve is shaded in grey. The third graph corresponds to question 3 in this example. It is titled “Area Beyond negative 2 and 2: $P(\text{absolute } Z \text{ greater than } 2)$ ”. The area to the left of a vertical line z equals negative 2 and the area to the right of a vertical line z equals 2 under the standard normal density curve is shaded in grey. The fourth graph corresponds to question 4 in this example. It is titled “Area Between negative 4 and 5: $P(\text{negative } 4 \text{ less than } Z \text{ less than } 5)$ ”. The area between a vertical line z equals negative 4 and a vertical line z equals 5 under the standard normal density curve is shaded in grey. [[Return to Figure 5.8](#)]

Example 5.1 Image Description: The leftmost image is a standard normal density curve with a labelled “ z equals ?” under the horizontal axis on the left end. A vertical line at the label “ z equals ?” is drawn and the area to its left under the bell-shaped curve is shaded in grey. On the top left corner of the graph, it reads “area equals 0.1”. Pause here to answer. The middle image is a standard normal density curve with a labelled “ z equals ?” under the horizontal axis on the right end. A vertical line at the label “ z equals ?” is drawn and the area to its right under the bell-shaped curve is shaded in grey. On the top right corner of the graph, it reads “area equals 0.1”. Pause here to answer. The rightmost image is a standard

normal density curve with a labelled “z equals ?” under the horizontal axis on the right end, which is a little bit further to the right end than the one in the previous example. A vertical line at the label “z equals ?” is drawn and the area to its right under the bell-shaped curve is shaded in grey. On the top right corner of the graph, it reads “area equals 0.05”. [\[Return to Example 5.1\]](#)

Figure 5.9 Image Description: From left to right, it reads “X approximates Normal (μ , σ)” on the first line and “x value” on the second line. Then there are parallel horizontal lines with an arrow at the end. The arrow of the horizontal line on the top is at the right end and pointing to the right, it writes “Z equals (X minus μ) divided σ ” above the line. The arrow of the horizontal line at the bottom is at the left end and pointing to the left, it writes “x equals μ plus z times σ ” below the line. Then it reads “z equals (x minus μ) divided by σ ”. Then two horizontal lines with arrows at both ends. It reads “Table II (standard normal table)” above the straight line on the top. Finally, it reads “area under the curve” on the first line and “probability/percentage” on the second line. [\[Return to Figure 5.9\]](#)

Figure 5.10 Image Description: This graph titled “Normal Probability Plot” is a scatter plot for the data given in Table 5.2. The x-axis labelled “Normal Score” is in increments of 0.5 from negative 1 to 1. The y-axis labelled “Sorted Grade” is in increments of 10 from 40 to 90. There are six points plotted, at (negative 1.28, 40), (negative 0.64, 75), (negative 0.20, 75), (0.20, 80), (0.64, 85), and (1.28, 90). [\[Return to Figure 5.10\]](#)

Figure 5.11 Image Description: Three normal probability plots titled “(a)”, “(b)”, and “(c)” respectively are presented in a row. They share the same x-axis and y-axis. The x-axis labelled “theoretical” is in increments of 1 from negative 2 to 2. The y-axis labelled “sample” is in increments of 0.25 from 0 to 1.00. The points in panel (a) are roughly on a straight line. The points in panel (b) show a “J” shape. The points in panel (c) also show a “J” shape. Most of the points in panel (c) are roughly on a straight line, except several extremely small and large observations. [\[Return to Figure 5.11\]](#)

Figure 5.12 Image Description: A total of eighteen graphs are shown here, in two groups of nine. The first nine graphs are presented in a 3 by 3 matrix. Histograms of a left skewed distribution, a normal distribution and a right skewed distribution are in the first row. Their corresponding normal probability plots and boxplot are given in the second and third row respectively. For a left skewed distribution, the histogram has a longer tail on the left-hand side; the normal probability plot is concave; the boxplot has a longer lower whisker and larger distance between Q1 and Q2. For a normal distribution, the histogram is roughly symmetric and bell-shaped; the normal probability plot shows a strong linear pattern; the lower and upper whiskers are roughly of the same length in the boxplot. For a right skewed distribution, the histogram has a longer tail on the right-hand side; the normal probability

plot has a “J” shape; the boxplot has a shorter lower whisker and smaller distance between Q1 and Q2.

The second set of nine graphs are also presented in a 3 by 3 matrix. Histograms of a multimodal distribution, a normal distribution with outliers and a uniform distribution are in the first row. Their corresponding normal probability plots and boxplot are given in the second and third row respectively. For a multimodal distribution, the histogram shows a roughly symmetric distribution with three peaks; the normal probability plot shows a flattening “S” shape; the lower and upper whiskers are roughly of the same length in the boxplot. For a normal distribution with outliers, the histogram is roughly symmetric and bell-shaped except for some observations on both ends; except for several observations on both ends, all the remaining points are roughly on a straight line in the normal probability plot; the boxplot is symmetric except for the outliers at both ends. For a uniform distribution, the histogram looks like a rectangle; the normal probability plot has a flattening “S” shape; the lengths of the lower and upper whiskers, the distance between Q1 and Q2, and the distance between Q2 and Q3 are all roughly the same. [[Return to Figure 5.12](#)]

Table 6.1 Image Description: There are five students in the population, if we randomly pick two students, there are 10 (5 choose 2) different samples and the sample means are 160, 165, 170, 175, 170, 175, 180, 180, 185, and 190. The mean and standard deviation of these ten sample means are 175 and 8.66 respectively. If we randomly pick three students, there are 10 (5 choose 3) different samples and the sample means are 165, 168.33, 171.67, 171.67, 175, 178.33, 175, 178.33, 181.67, and 185. The mean and standard deviation of these ten sample means are 175 and 5.77 respectively. If we randomly pick four students, there are 5 (5 choose 4) different samples and the sample means are 170, 172.5, 175, 177.5, and 180. The mean and standard deviation of these ten sample means are 175 and 3.54 respectively. [[Return to Table 6.1](#)]

Figure 6.1 Image Description: The probability distribution of the sample mean for n equals 2 is given on the left-hand side in a table with two columns. The first column labelled as “ \bar{x} ” lists all possible values of the sample mean as values of \bar{x} . Their associated probabilities (relative frequency) are given in the second column labelled as “ $P(\text{upper case } X \text{ bar equals lower case } x \text{ bar})$ ”. The values are as follows: for \bar{x} at 160, one-tenth equals 0.1; for \bar{x} at 165, one-tenth equals 0.1; for \bar{x} at 170, two-tenths equals 0.2; for \bar{x} at 175, two-tenths equals 0.2; for \bar{x} at 180, one-tenth equals 0.1; for \bar{x} at 185, one-tenth equals 0.1; and for \bar{x} at 190, one-tenth equals 0.1. The probability histogram is shown on the right-hand side. The y-axis is probability (i.e., relative frequency) from 0 to 0.2 in increments of 0.1 and x-axis is “Average Height of Two” with bars for each \bar{x} . The heights of the bars are the same as their second column values. [[Return to Figure 6.1](#)]

Figure 6.2 Image Description: Two graphs are presented in a row. The one on the left is

titled “Population Distribution of Grade”. The y-axis labelled “Density” is in increments of 0.05 from 0 to 0.20. The x-axis labelled “Grade” is increment of 20 from 20 to 100. The graph consists of 4 bars symmetric at grade equals 70, and a black bell-shaped curve ranging from 20 to 120 is drawn on the top of the bars. A red vertical line is drawn at grade equals 70. In the middle of the left-hand side of the graph, it reads “population mean equals 70” in the first line and “population SD equals 10” in the second line. The one on the right is titled “Q-Q Plot of Grade”. The y-axis labelled “Observed Quantiles: Grade” is in increments of 20 from 40 to 100. The x-axis labelled “Theoretical Quantiles: Normal Score” is increment of 2 from negative 4 to 4. The points show an almost perfect straight line. [[Return to Figure 6.2](#)]

Figure 6.3 Image Description: Three density curves of the sample mean for sample size n equals 2, 5 and 30 are presented in a row. These three graphs have identical x- and y-axis. Their corresponding normal probability plots are shown below. The first graph on the first row has the title “Distribution of Sample Mean With n equals 2”. The y-axis labelled “Density” is in increments of 0.05 from 0 to 0.20. The x-axis labelled “Average Grade” is in increment of 20 from 20 to 100. The graph consists of bars, and a black bell-shaped curve ranging from 35 to 105 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the left-hand side of the graph, it reads “mean of sample mean equals 70” in the first line and “SD of sample mean equals 7.2” in the second line.

The second graph on the first row has the title “Distribution of Sample Mean With n equals 5”. The y-axis labelled “Density” is in increments of 0.05 from 0 to 0.20. The x-axis labelled “Average Grade” is in increment of 20 from 20 to 100. The graph consists of bars, and a black bell-shaped curve ranging from 50 to 90 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the left-hand side of the graph, it reads “mean of sample mean equals 70” in the first line and “SD of sample mean equals 4.5” in the second line.

The third graph on the first row has the title “Distribution of Sample Mean With n equals 30”. The y-axis labelled “Density” is in increments of 0.05 from 0 to 0.20. The x-axis labelled “Average Grade” is in increments of 20 from 20 to 100. The graph consists of bars, and a black bell-shaped curve ranging from 60 to 80 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the left-hand side of the graph, it reads “mean of sample mean equals 70” in the first line and “SD of sample mean equals 1.8” in the second line.

Three normal probability plots corresponding to the density curves above are presented

in the second row. The first probability plot has the title “Q-Q Plot of Sample Mean With n equals 2”. The y-axis labelled “Observed Quantiles: Average Grade” is in increments of 10 from 40 to 100. The x-axis labeled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show an almost perfect straight line. The second probability plot has the title “Q-Q Plot of Sample Mean With n equals 5”. The y-axis labelled “Observed Quantiles: Average Grade” is in increments of 5 from 55 to 85. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show an almost perfect straight line. The third probability plot has the title “Q-Q Plot of Sample Mean With n equals 30”. The y-axis labelled “Observed Quantiles: Average Grade” is in increments of 2 from 62 to 76. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show an almost perfect straight line. [\[Return to Figure 6.3\]](#)

Figure 6.4 Image Description: Density curve of outcome of rolling a die with a title “Population Distribution”. The y-axis labelled “Density” is in increments of 0.5 from 0 to 1.5. The x-axis labelled “Outcome of one die” is in increments of intervals of 1 from 1 to 6. The graph consists of 6 bars, the height of each bars is one-sixth. A red vertical line is drawn at outcome equals 3.5. In the middle of the left-hand side of the graph, it reads “population mean equals 3.5” in the first line and “population SD equals 1.71” in the second line. [\[Return to Figure 6.4\]](#)

Figure 6.5 Image Description: Three density curves of the sample mean for sample size n equals 2, 5 and 30 are presented in a row, these three graphs have identical x- and y-axis. Their corresponding normal probability plots are shown below. The first graph in the first row is titled “Distribution of Sample Mean With n equals 2”. The y-axis labelled “Density” is in increments of 0.5 from 0 to 1.5. The x-axis labelled “Average of n equals 2 Dice” is in increments of 1 from 1 to 6. The graph consists of bars, and a black symmetric triangular curve ranging from negative 0.5 to 6.5 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the left-hand side of the graph, it reads “mean of sample mean equals 3.5” in the first line and “SD of sample mean equals 1.2” in the second line.

The second graph in the first row is titled “Distribution of Sample Mean With n equals 5”. The y-axis labelled “Density” is in increments of 0.5 from 0 to 1.5. The x-axis labelled “Average of n equals 5 Dice” is in increments of 1 from 1 to 6. The graph consists of bars, and a black bell-shaped curve ranging from negative 0.5 to 6.5 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost

identical. In the middle of the left-hand side of the graph, it reads “mean of sample mean equals 3.5” in the first line and “SD of sample mean equals 0.8” in the second line.

This is the third graph in the first row with a title “Distribution of Sample Mean With n equals 30”. The y-axis labelled “Density” is in increments of 0.5 from 0 to 1.5. The x-axis labelled “Average of n equals 30 Dice” is in increments of 1 from 1 to 6. The graph consists of bars, and a black bell-shaped curve ranging from 2 to 5 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the left-hand side of the graph, it reads “mean of sample mean equals 3.5” in the first line and “SD of sample mean equals 0.31” in the second line.

Three normal probability plots corresponding to the density curves above are presented in the second row. The first probability plot is titled “Q-Q Plot of Sample Mean With n equals 2”. The y-axis labelled “Observed Quantiles: Average of n equals 2 Dice” is in increments of 1 from 1 to 6. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show a stair with 11 steps going upward from the left to the right. The second probability plot is titled “Q-Q Plot of Sample Mean With n equals 5”. The y-axis labelled “Observed Quantiles: Average of n equals 5 Dice” is in increments of 1 from 1 to 6. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show a stair with 26 steps going upward from the left to the right. The points are roughly on a straight line. The third probability plot is titled “Q-Q Plot of Sample Mean With n equals 30”. The y-axis labelled “Observed Quantiles: Average of n equals 30 Dice” is in increments of 1 from 1 to 6. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show a stair with many steps going upward from the left to the right. Therefore, the points appear to form a straight line. [[Return to Figure 6.5](#)]

Figure 6.6 Image Description: Two graphs are presented in a row. The one on the left is titled “Population Distribution of Survival Time”. The y-axis labelled “Density” is in increments of 0.1 from 0.0 to 0.5. The x-axis labelled “Survival Time (year)” is in increments of 5 from 0 to 20. The graph consists of 4 bars. The heights are as follows: x between 0 and 5 has a height of roughly 0.13, x between 5 and 10 has a height of roughly 0.5, x between 10 and 15 has a height of roughly 0.2, and x between 15 and 20 has a height of roughly 0.1. A black reversed “J” shaped curve ranging from negative 1 to 21 is drawn on the top of the bars. A red vertical line is drawn at survival time equals 5. In the middle of the right-hand side of the graph, it reads “population mean equals 5” in the first line and “population SD equals 5” in the second line.

The graph on the right panel is titled “Q-Q Plot of Survival Time”. The y-axis labelled “Observed Quantiles: Survival Time” is in increments of 10 from 0 to 40. The x-axis labelled

“Theoretical Quantiles: Normal Score” is an increment of 2 from negative 4 to 4. The points show a “J” shaped curve. [[Return to Figure 6.6](#)]

Figure 6.7 Image Description: Three density curves of the sample mean for sample size n equals 2, 5 and 30 are presented in a row. These three graphs have identical x- and y-axis. Their corresponding normal probability plots are shown below. The first graph is titled “Distribution of Sample Mean With n equals 2”. The y-axis labelled “Density” is in increments of 0.1 from 0.0 to 0.5. The x-axis labelled “Survival Time (year)” is incremented in intervals of 5 from 0 to 20. The graph consists of bars showing a right-skewed distribution, and a black curve ranging from negative 1 to 21 is drawn on the top of the bars. The curve goes upward from survival time equals negative 1 to the peak at survival time equals 3 and goes downward until survival equals 21. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the right-hand side of the graph, it reads “mean of sample mean equals 5” in the first line and “SD of sample mean equals 3.6” in the second line.

The second graph is titled “Distribution of Sample Mean With n equals 5”. The y-axis labelled “Density” is in increments of 0.1 from 0.0 to 0.5. The x-axis labelled “Survival Time (year)” is incremented in intervals of 5 from 0 to 20. The graph consists of bars showing a right-skewed distribution, and a black curve ranging from negative 1 to 19 is drawn on the top of the bars. The curve goes upward from survival time equals negative 1 to the peak at survival time equals 4 and goes downward till survival equals 19. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the right-hand side of the graph, it reads “mean of sample mean equals 5” in the first line and “SD of sample mean equals 2.2” in the second line.

The third graph is titled “Distribution of Sample Mean With n equals 30”. The y-axis labelled “Density” is in increments of 0.1 from 0.0 to 0.5. The x-axis labelled “Survival Time (year)” is incremented in intervals of 5 from 0 to 20. The graph consists of bars showing a symmetric distribution, and a black bell-shaped (roughly) curve ranging from 1.5 to 9.5 is drawn on the top of the bars. A red solid vertical line representing the population mean and a blue dashed vertical line representing the mean of the sample means are drawn. The red and blue lines are almost identical. In the middle of the right-hand side of the graph, it reads “mean of sample mean equals 5” in the first line and “SD of sample mean equals 0.9” in the second line.

Three normal probability plots corresponding to the density curves above are presented in the second row. The first probability plot is titled “Q-Q Plot of Sample Mean With n equals 2”. The y-axis labelled “Observed Quantiles: Average Survival Time” is in increments of 5 from

0 to 25. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show a “J” shaped curve. The second probability plot is titled “Q-Q Plot of Sample Mean With n equals 5”. The y-axis labelled “Observed Quantiles: Average Survival Time” is in increments of 5 from 0 to 15. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from negative 4 to 4. The points show a flattened “J” shaped curve. The third probability plot is titled “Q-Q Plot of Sample Mean With n equals 30”. The y-axis labelled “Observed Quantiles: Average Survival Time” is in increments of 1 from 3 to 9. The x-axis labelled “Theoretical Quantiles: Normal Score” is in increments of 2 from -4 to 4. The points are roughly on a straight line. [[Return to Figure 6.7](#)]

Exercise 6.1 Image Description: The graph is titled “Density Curve of Rent”. The y-axis labelled “Density” is in increments of 0.01 from 0.00 to 0.12. The x-axis labelled “Rent of One-Bedroom Apartment (\$100)” is incremented in intervals of 5 from 0 to 30. The curve goes upward from rent equals 0 to the peak at rent equals 5 and then goes downward until rent equals 30. [[Return to Exercise 6.1](#)]

Chapter 6 Review Question 7 Image Description: The graph is titled “Density Curve of Rent”. The y-axis labelled “Density” is in increments of 0.01 from 0.00 to 0.12. The x-axis labelled “Rent of One-Bedroom Apartment (\$100)” is incremented in intervals of 5 from 0 to 30. The curve goes upward from rent equals 0 to the peak at rent equals 5 and then goes downward until rent equals 30. [[Return to Question 7](#)]

Assignment 6 Question 7 Image Description: The graph is titled “Density Curve of Rent”. The y-axis labelled “Density” is in increments of 0.01 from 0.00 to 0.12. The x-axis labelled “Rent of One-Bedroom Apartment (\$100)” is incremented in intervals of 5 from 0 to 30. The curve goes upward from rent equals 0 to the peak at rent equals 5 and then goes downward until rent equals 30. [[Return to Question 7](#)]

Figure 7.1 Image Description: The graph is titled “Sample Mean”. A horizontal line is drawn with labels at 339, 340, 341, 342 and 343. A green vertical line is drawn at 341, and two blue vertical lines are drawn at 340.02 and 341.98. Twenty horizontal lines with very short vertical bars at each end represent twenty 95% confidence intervals. The center of each interval (the sample mean) is shown as a red diamond. All intervals are of the same length which is 1.96. There is one interval that does not intersect the green vertical line. [[Return to Figure 7.1](#)]

Figure 7.2 Image Description: A horizontal line representing a confidence interval centred at 339 (indicated as \bar{x} below the number) is drawn. The left end point of the interval is calculated as $339 - 0.98$ equals 338.02. The formula to calculate the value is given below the equation: $\bar{x} - z_{\alpha/2} \times \sigma / \sqrt{n}$. The right end point of the interval is calculated as $339 + 0.98$ equals 339.98. The formula to calculate the value is given below the equation: $\bar{x} + z_{\alpha/2} \times \sigma / \sqrt{n}$.

of n . The interval is divided into two halves indicated by two horizontal lines with arrows at both ends. An equation “capital E equals z sub alpha over 2 times sigma over root of n equals 0.98” is written above the horizontal line of each half. A red vertical line indicating the population mean μ is drawn. A green vertical line indicating value 341 is shown to be outside of the interval. [\[Return to Figure 7.2\]](#)

Figure 7.3 Image Description: Several t-curves at varying degrees of freedom are compared to the standard normal curve. The y-axis of the graph is “Density” in increments of 0.1 from 0 to 0.4 and the x-axis is “X” in increment of 1 from negative 4 to 4. Four bell-shaped curves are shown in this figure. The blue dashed curve representing a t distribution with 1 degrees of freedom is flattest curve. The purple dashed curve representing a t distribution with 3 degrees of freedom is the second flattest curve. The red dashed-dotted curve representing a t distribution with 15 degrees of freedom is slightly flatter but looks almost the same as the black density curve representing the standard normal distribution. [\[Return to Figure 7.3\]](#)

Figure 7.4 Image Description: This figure shows part of the first page of Table IV (Values of t sub alpha of t distribution). The first column of the table (labelled “df”) gives the degrees of freedom of a t distribution. The first row of the table reads “alpha: Area to the Right of t sub alpha”. The second row gives the values of alpha: 0.40, 0.30, 0.20, 0.15, 0.10, 0.05, 0.025, 0.010, 0.0075, 0.005, 0.0025 and 0.0005. The elements of the main body of the table are t-scores (in three decimal places) having an area of alpha (given as the column name) to its right for a given degrees of freedom (given as the row name). The graph shows that the t-score has an area of 0.025 to its right under the t distribution with df equals 9 is 2.262. The graph also shows that the area to the right of a t-score of 1.5 under the t distribution with df equals 9 is between 0.10 and 0.05. [\[Return to Figure 7.4\]](#)

Figure 7.5 Image Description: The bell-shaped curve shows the density of a t distribution with 14 degrees of freedom (df equals 14). Several coloured lines show important critical values according to Table IV. The largest area equalling 0.1 is shown to the right of a purple line at 1.345 under the density curve. The second largest area equalling 0.05 is shown to the right of a light blue line at 1.761. The middling area equalling 0.025 is shown to the right of a light green line at 2.145. The second smallest area equalling 0.01 is shown to the right of a red line at 2.624. The smallest area equalling 0.005 is shown to the right of a brown line at 2.977. [\[Return to Figure 7.5\]](#)

Example 7.1 Image Description: This figure shows part of Table IV (Values of t sub alpha of t distribution) for df=1 to 5 and df=31 to 50. [Complete Table IV \(see Appendix B Table IV\).](#) [\[Return to Example 7.1\]](#)

Figure 8.1 Image Description: Two density curves with overlap are shown over a horizontal

axis labelled “Sugar level in blood”. The red curve on the left indicates the blood sugar level for patients without diabetes and the green curve on the right indicates the blood sugar level for patients with diabetes. There is a black vertical line labeled “cut-off C”. The area to the left of the cut-off C under the density curve for diabetes is shaded in blue and is labeled “False negative, Type II error”. The area to the right of the cut-off C under the density curve for diabetes free patients is shaded in gold and is labelled “False positive, Type I error”. [\[Return to Figure 8.1\]](#)

Figure 8.2 Image Description: The figure summarises a table illustrating the main idea of a hypothesis testing. We should reject the null hypothesis $H_0: \mu = \mu_0$ and claim the alternative $H_a: \mu \neq \mu_0$ if the sample mean \bar{x} is either too large or too small. We should reject the null hypothesis $H_0: \mu \leq \mu_0$ and claim the alternative $H_a: \mu > \mu_0$ if the sample mean \bar{x} is too large. We should reject the null hypothesis $H_0: \mu \geq \mu_0$ and claim the alternative $H_a: \mu < \mu_0$ if the sample mean \bar{x} is too small. [\[Return to Figure 8.2\]](#)

Figure 8.3 Image Description: Two identical unimodal and right-skewed density curves are shown over a horizontal axis. The one on the left is labelled at critical value C and the area to its right under the density curve is shaded in grey. This shaded area is called the rejection region. The density curve on the right is labelled at x_0 and the area to its right under the density curve is shaded in grey. This shaded area is called the p-value. [\[Return to Figure 8.3\]](#)

Figure 8.4 Image Description: The figure illustrates that the rejection region of a two-tailed z test is either the z-score greater than $z_{\alpha/2}$ or less than negative $z_{\alpha/2}$; the rejection region of a right-tailed z test is the z-score greater than z_{α} ; the rejection region of a left-tailed z test is the z-score less than negative z_{α} . [\[Return to Figure 8.4\]](#)

Figure 8.5 Image Description: The figure summarises a table illustrating the calculation of a p-value. For a two-tailed z test with the null hypothesis $H_0: \mu = \mu_0$ and the alternative hypothesis $H_a: \mu \neq \mu_0$, the p-value equals twice of the area under a standard normal curve to the right of the absolute value of the observed z-score z_0 . For a right-tailed z test with the null hypothesis $H_0: \mu \leq \mu_0$ and the alternative hypothesis $H_a: \mu > \mu_0$, the p-value equals the area under a standard normal curve to the right of the observed z-score z_0 . For a left-tailed z test with the null hypothesis $H_0: \mu \geq \mu_0$ and the alternative $H_a: \mu < \mu_0$, the p-value is calculated as the area under a standard normal curve to the left of the observed z-score z_0 . [\[Return to Figure 8.5\]](#)

Example 8.1 Image Description: A standard normal density curve. Both the area under the curve to the right of 4 and to the left of negative 4 are shaded in grey. Each area has a label saying “area equals p-value divided by 2”. Part of Table II: area under the standard normal curve for negative z is also shown. It indicates that the area to the left of negative 3.99 under the standard normal curve is 0.0000. [[Return to Example 8.1](#)]

Example 8.2 Image Description: A standard normal density curve is shown. Both the area under the curve to the right of 1.96 (which is $z_{\alpha/2}$) and to the left of negative 1.96 (which is $-z_{\alpha/2}$) are shaded in grey. Each area has a label saying “area equals alpha divided by 2”. [[Return to Example 8.2](#)]

Example 8.3 Image Description: Two t-density curve with 35 degrees of freedom. In the graph on the left panel, the area to the left of negative 0.714 under the t-curve is shaded in grey. In the graph on the right panel, the area to the right of 0.714 under the t-curve is shaded in grey. The two areas are the same. Part of Table IV: Values of t_{α} of t distribution is also shown. From the table, we know that for a t distribution with degrees of freedom df equals 35, the area to the right of 0.714 is between 0.20 and 0.30. [[Return to Example 8.3](#)]

Example 8.4 Image Description: A t-density curve with degrees of freedom df equals 35 is drawn. The area under the curve to the left of negative 2.438 is shaded in grey with a label saying “area equals alpha equals 0.01”. A purple vertical line at negative 0.714 is drawn. We know that negative 0.714 is outside the shaded area, the rejection region of the test. [[Return to Example 8.4](#)]

Figure 8.6 Image Description: The bell-shaped curve shows the density of a t-distribution with 35 degrees of freedom (df equals 35). Several critical values according to Table IV are shown. The largest area equalling 0.1 is to the right of a pink line at 1.306 under the density curve. The second largest area equalling 0.05 is to the right of a dark blue line at 1.690. The middling area equalling 0.025 is to the right of a bright green line at 2.030. The second smallest area equalling 0.01 is to the right of a red line at 2.438. The smallest area equalling 0.005 is shown to the right of a black line at 2.724. [[Return to Figure 8.6](#)]

Figure 9.1 Image Description: Two big identical ovals are presented side by side. The oval on the left-hand side is labeled as “Population 1” and the one on the right-hand side is labeled as “Population 2”. Below the two ovals it reads “Greek letter mu sub 1” and “Greek letter mu sub 1” respectively. It reads “Two independent samples” between the two bigger ovals. There is a smaller oval inside each of the bigger oval representing a simple random sample from each population. Inside the smaller oval of population 1, it reads vertically “n sub 1, x-bar sub 1 and s sub 1”. Inside the smaller oval of population 2, it reads vertically “n sub 2, x-bar sub 2 and s sub 2”. [[Return to Figure 9.1](#)]

Figure 9.2 Image Description: A t-curve with degrees of freedom df equals 35 is drawn. The area under the curve to the right of 2.403 (t-score with area 0.01 to its right) is shaded in grey with a label saying “rejection region” in the first line and “area equals alpha equals 0.01” in the second line. A purple vertical line at 5.332 is drawn. The graph indicates that the observed t-score $t_{sub o}$ is outside the shaded area, the rejection region of the test. [[Return to Figure 9.2](#)]

Figure 9.3 Image Description: A one-sided confidence interval starting from 9.887 pointing towards positive infinity is drawn. A short green vertical line is shown at 0 and a short blue vertical line is shown at 5. A short red vertical line is drawn somewhere within the one-sided interval with a label saying “Greek letter μ sub 1 minus Greek letter μ sub 2 greater than 9.887”. [[Return to Figure 9.3](#)]

Figure 9.4 Image Description: This graph titled “Normal Probability Plot on Differences” is a normal Q-Q plot on the paired differences given in the third column of Table 9.3. The x-axis labelled “norm quantiles” is in increments of 0.5 from negative 1.5 to 1.5. The y-axis labelled “Difference” is in increments of 10 from negative 10 to 50. There are 11 points plotted, all points are roughly on a red straight line and within the 95% simultaneous confidence band. The smallest and the large difference have a number “8” and “10” respectively next to the points. [[Return to Figure 9.4](#)]

Figure 10.1 Image Description: Four histograms of the sample proportion for sample size n equals 50, 100, 200 and 1000 are presented in a row. The first graph is titled “ n equals 50”. The y-axis labelled “Frequency” is in increments of 200 from 0 to 1400. The x-axis labelled “Sample proportion” is incremented in intervals of 0.05 from 0 to 0.2. The first graph consists of eight bars with a width of 0.03 each from 0 to 0.16. The heights of the bars decrease from above 1400 at the first bar to close 0 at the last bar showing an extremely right-skewed distribution. A red solid vertical line representing the population proportion and a blue dashed vertical line representing the mean of the sample proportions are drawn. The red and blue lines coincide at sample proportion equals 0.05.

This is the second graph with a title “ n equals 100”. The y-axis labelled “Frequency” is in increments of 200 from 0 to 1400. The x-axis labelled “Sample proportion” is incremented in intervals of 0.05 from 0 to 0.2. The graph consists of 13 bars of width 0.01 from 0 to 0.13. The heights of the bars increase from around 200 at the first bar to the peak around 900 at sample proportion equals 0.05 then decrease to around 0 at the last bar. The distribution is still right-skewed, but less so. A red solid vertical line representing the population proportion and a blue dashed vertical line representing the mean of the sample proportions are drawn. The red and blue lines coincide at sample proportion equals 0.05.

This is the third graph with a title “ n equals 200”. The y-axis labelled “Frequency” is in

increments of 200 from 0 to 1400. The x-axis labelled “Sample proportion” is incremented in intervals of 0.05 from 0 to 0.2. The graph consists of 11 bars of width 0.01 from 0 to 0.11. The heights of the bars increase from around 20 at the first bar to the peak around 1200 at sample proportion equals 0.05 then decrease to around 0 at the last bar. The distribution is only slightly right-skewed. A red solid vertical line representing the population proportion and a blue dashed vertical line representing the mean of the sample proportions are drawn. The red and blue lines coincide at sample proportion equals 0.05.

This is the fourth graph with a title “n equals 1000”. The y-axis labelled “Frequency” is in increments of 200 from 0 to 1400. The x-axis labelled “Sample proportion” is incremented in intervals of 0.05 from 0 to 0.2. The graph consists of 12 bars of width 0.005 from 0.025 to 0.085. The heights of the bars increase from 0 at the first bar to the peak around 1400 at sample proportion equals 0.05 then decrease to around 0 at the last bar. The distribution is now roughly symmetric. A red solid vertical line representing the population proportion and a blue dashed vertical line representing the mean of the sample proportions are drawn. The red and blue lines coincide at sample proportion equals 0.05. [\[Return to Figure 10.1\]](#)

Figure 10.2 Image Description: A curve from 0 to 1 is shown with a red dot at the peak. The y-axis of the graph labelled as “p-hat times (1-p-hat)” is in increments of 0.05 from 0 to 0.25. The x-axis labelled as “p-hat” is in increments of 0.2 from 0 to 1. The curve is smooth and symmetric; it keeps going upward from the coordinates (0, 0) to peak with coordinates (0.5, 0.25) and then starts going downward to the coordinates (1, 0). [\[Return to Figure 10.2\]](#)

Figure 11.1 Image Description: The y-axis of the graph is “Density” in increments of 0.1 from 0 to 0.4 and the x-axis is “X” in increments of 5 from 0 to 30. Five chi-square curves are shown in this figure. The black curve goes to infinity at x equals 0 and 0 as x approaches 30. This is the chi-square curve with 1 degrees of freedom. The red curve increases from y approximating 0.2 at x equals 0 to y approximating 0.24 at x equals 3 then approaches 0 as x approaches 30. This is the chi-square curve with 3 degrees of freedom. The blue curve is right-skewed with a peak at coordinates (4, 0.15). This is the chi-square with 5 degrees of freedom. The purple curve is right-skewed with a peak at coordinates (7, 0.1). This is the chi-square with 9 degrees of freedom. The green curve is slightly right-skewed with a peak at coordinates (14, 0.08). This is the chi-square with 15 degrees of freedom. [\[Return to Figure 11.1\]](#)

Table 11.1 Image Description: This figure shows part of Table V (Values of Greek letter chi-square sub Greek letter alpha of chi-square distribution). The first column of the table (labelled “df”) gives the degrees of freedom for chi-square distributions. The first row of the table reads “Greek letter alpha: Area to the Right of Greek letter chi-square sub alpha”. The second row gives the values of alpha: 0.995, 0.990, 0.975, 0.950, 0.9, 0.1, 0.05, 0.025, 0.01 and 0.005. The elements of the main body of the table are chi-square scores (in three decimal

places) having an area of alpha (given as the column name) to its right for a given degrees of freedom (given as the row name). [[Return to Table 11.1](#)]

Figure 11.2 Image Description: Three bars are presented for “Smoker”, “Non-Smoker” and “Total” from left to right. Each bar has two segments: the green segment at the bottom represents individuals with cancer and the red segment on the top for individuals without cancer. The y-axis labeled “Relative Frequency” is in increments of 0.2 from 0 to 1. The heights of the green segment are 0.333 for Smoker, 0.176 for Non-Smoker, and 0.2 for Total. [[Return to Figure 11.2](#)]

Figure 12.1 Image Description: Three big identical ovals are presented in a row. The first oval is labeled as “Population 1”, the second one as “Population 2”, and the third one as “Population k”. Below the ovals it reads “Greek letter mu sub 1”, “Greek letter mu sub 2” and “Greek letter mu sub k” respectively. It reads “Two independent samples” between the first two bigger ovals and the last two big ovals. There is a smaller oval inside each of the bigger oval representing a simple random sample from each population. Inside the smaller oval of population 1, it reads vertically “n sub 1, x-bar sub 1 and s sub 1”. Inside the smaller oval of population 2, it reads vertically “n sub 2, x-bar sub 2 and s sub 2”. Inside the smaller oval of population k, it reads vertically “n sub k, x-bar sub k and s sub k”. [[Return to Figure 12.1](#)]

Figure 12.2 Image Description: The x-axis is in increments of 2 from 2 to 10. Data set 1 has red circles at 1, 2, 3, 3, 4, and 5 and blue crosses at 6, 7, 8, 8, 9, and 10. Data set 2 has red circles at 1, 3, 4, 5, 8, and 9 and blue crosses at 2, 3, 6, 7, 8, and 10 [[Return to Figure 12.2](#)]

Figure 12.3 Image Description: The y-axis of the graph is “Density” in increments of 0.2 from 0 to 1.2 and the x-axis is “X” in increments of 1 from 0 to 5. Five F-density curves are shown in this figure. The black curve is right-skewed with a peak at coordinates (0.3, 0.7). This is the F distribution with (3, 30) degrees of freedom. The red curve is right-skewed with a peak at coordinates (0.6, 0.6). This is the F distribution with (30, 3) degrees of freedom. The green curve is right-skewed with a peak at coordinates (0.5, 0.6). This is the F distribution with (15, 3) degrees of freedom. The purple curve is right-skewed with a peak at coordinates (0.8, 0.98). This is the F distribution with (15, 30) degrees of freedom. The blue curve is right-skewed with a peak at coordinates (0.9, 1.18). This is the F distribution with (30, 30) degrees of freedom. [[Return to Figure 12.3](#)]

Table 12.3 Image Description: This figure shows part of Table VI (Values of F sub alpha of F distribution) (Table 6). The elements of the main body of the table are F scores having an area of alpha (given as the row name) to its right for a given numerator degrees of freedom df sub n (given as the column name). We see df sub n equals 1, 2, up to 10 in this image. The alpha values provided are 0.5, 0.1, 0.05, 0.025, 0.01, 0.005, and 0.001. The table is grouped by the denominator degrees of freedom df sub d. This image shows seven rows and ten

columns of F scores for df sub d equals 1, another seven rows for df sub d equals 2, and another seven rows for df sub d equals 3. [[Return to Table 12.3](#)]

Figure 12.4 Image Description: Six images are shown in a 2 by 3 matrix. The three graphs in the first rows are the side-by-side histograms of downloading time for 7 AM, 5 PM and 12 AM from left to right. Their corresponding side-by-side boxplots are presented in the second row. [[Return to Figure 12.4](#)]

Figure 12.5 Image Description: Six images are shown in a 2 by 3 matrix. The three graphs in the first rows are the side-by-side histograms of bone density for control, high jump and low jump groups from left to right. Their corresponding side-by-side boxplots are presented in the second row. [[Return to Figure 12.5](#)]

Figure 13.1 Image Description: The x-axis labeled “Age (Years)” is in increments of 2 from 2 to 12. The y-axis labeled “Price (\$1000)” is in increments of 2 from 4 to 14. Fifteen points are plotted at (1, 14), (1, 13), (3, 13), (4, 10), (4, 10), (5, 9), (5, 9), (6, 7), (7, 7), (7, 8), (8, 7), (8, 6), (10, 5), (10, 4) and (13, 3). The points roughly fall on a straight line going downward. [[Return to Figure 13.1](#)]

Figure 13.2 Image Description: A straight line given by “y equals b sub 0 plus b sub 1 times x” is drawn. The straight line goes downward and intersects the y-axis at coordinates (0, b sub 0). Two vertical lines at x and x plus 1 are drawn to show that the slope of the straight line b sub 1 is the change in the response Y when X increases by 1 unit. [[Return to Figure 13.2](#)]

Figure 13.3 Image Description: The fitted least-squares straight line is added to the scatter plot of 15 used cars with “Age (Years)” and “Price (\$1000)” as the x- and y-axis. Two points on the straight line are identified at age equals 3 and age equals 6, these two points are denoted as y-hat. Connect the two fitted value y-hat with their corresponding observed values, we see that the residual (defined as y minus y-hat) of the observation with age equals 3 is positive and the residual of the observation with age equals 6 is negative. [[Return to Figure 13.3](#)]

Figure 13.4 Image Description: On the scatter plot with the fitted least-squares straight line for those 15 used cars, three points on the least-squares straight line are chosen with age equals 2, 10 and 20. Their coordinates are (2, 12.2316), (10, 4.686), and (20, negative 4.746). The line slopes down through all three points. [[Return to Figure 13.4](#)]

Figure 13.5 Image Description: The x-axis is in increments of 5 from 0 to 20 and the y-axis is in increments of 10 from 0 to 40. Except for an outlier point at coordinates (20, 42), all data points are within the region [0, 5] by [0, 10]. The fitted least-squares straight lines with and without the outlier are almost identical. [[Return to Figure 13.5](#)]

Figure 13.6 Image Description: The x-axis is in increments of 5 from 0 to 20 and the y-axis is in increments of 2 from 0 to 8. Except for an outlier point at coordinates (20, 1), all data points are within the region [0, 4] by [3, 9]. The fitted least-squares straight lines with and without the outlier are almost orthogonal. The slope of the line with the outlier is negative, but the slope of the one without the outlier is positive. [[Return to Figure 13.6](#)]

Figure 13.7 Image Description: Four scatter plots are shown in a row. The first graph labelled a shows data points forming a “U” shape with a wide mouth. The second graph labelled b shows data points falling in a wide band going upward. The third graph labelled c shows the data points falling in a narrower but still wide band going downward. The fourth graph labelled d shows the data points falling in a very narrow band going upward. [[Return to Figure 13.7](#)]

Figure 13.8 Image Description: A vector Y starting from the origin represents the observed y values. The projection of Y onto a hyperplane is denoted as \hat{Y} . The projection of the vector Y minus \bar{Y} is denoted as $\hat{Y} - \bar{Y}$. The angle between the vector Y minus \bar{Y} and $\hat{Y} - \bar{Y}$ is θ . The residual vector $Y - \hat{Y}$ is orthogonal to the hyperplane. The vectors $Y - \bar{Y}$, $\hat{Y} - \bar{Y}$ and $Y - \hat{Y}$ form a right triangle. [[Return to Figure 13.8](#)]

Figure 13.9 Image Description: On the scatter plot with the fitted least-squares straight line for those 15 used cars, three identical vertical bell-shaped curves are shown at age equals 4, 8, and 10. The centers of the bell-shaped curves gather around the least-squares straight line. [[Return to Figure 13.9](#)]

Figure 13.10 Image Description: Six graphs are presented in a 2 by 3 matrix. The three graphs in the first row are residual plots with “Standardised Residuals” as the y-axis which is in increment 1 from -4 to 4. All three residual plots have a red horizontal line at y equals 0. The corresponding normal probability plots of the three sets of residuals are presented in the second row. The first residual plot labelled a shows the data points gathered randomly within a horizontal band from y equals negative 3 to y equals 3. The second residual plot labelled as b shows the data points in a “U” shape with a wide mouth. The third residual plot labelled c shows the data points having a wider range in y-axis as x increases.

The second row shows the corresponding Q-Q plots. The first normal probability plot is of residual plot a. The y-axis labelled “Sample Quantiles” is in increment 1 from negative 2 to 2. The x-axis labelled “Theoretical Quantiles” is in increment 1 from negative 2 to 2. All points roughly fall on a straight line. The normal probability plot of residual plot b has a y-axis labelled “Sample Quantiles” and incremented in 1 from negative 1 to 4. The x-axis is labelled “Theoretical Quantiles” and is incremented in 1 from negative 2 to 2. The points show a flattened “J” shape. The normal probability plot of residual plot c has a y-axis

labelled “Sample Quantiles” and incremented 1 from negative 1 to 3. The x-axis is labelled “Theoretical Quantiles” and is increment 1 from negative 2 to 2. The points show an “s” shape. [\[Return to Figure 13.10\]](#)

Table 13.2 Image Description: A table of used car values. The given variables are: Age in years, Price in thousands of dollars, \hat{y} given by $\text{price-hat equals } 14.118 \text{ minus } 0.9432 \text{ times age}$, and residual given by $\text{error at } i \text{ equals } y \text{ minus } \hat{y}$. The first car has the following values: Age equals 1, Price equals 14, \hat{y} equals 13.1748, and Residual equals 0.8252. [\[Return to Table 13.2\]](#)

Figure 13.11 Image Description: The first graph is a normal Q-Q plot. The y-axis labelled “Sorted Residuals” is in increments of 0.5 from negative 1.5 to 1.5. The x-axis labelled “Normal Score” is in increment of 1 from negative 2 to 2. Fifteen points are plotted and they are roughly on a straight line.

The second image is a scatter plot of standardised residuals. The y-axis labelled “Standardised Residuals” is in increment of 1 from negative 3 to 3. The x-axis labelled “Age (Years)” is in increment of 1 from 1 to 13. Fifteen points are plotted and they randomly scatter within a horizontal band from negative 2 to 2. [\[Return to Figure 13.11\]](#)

Figure 13.12 Image Description: A histogram showing one possible distribution of the slope. The y-axis labelled “Density” is in increment of 1 from 0 to 5. The x-axis labelled “Slope b_1 (in \$1000 per year)” is in increment of 0.1 from negative 1.2 to negative 0.7. The graph consists of bars and a black bell-shaped curve ranging from negative 1.2 to negative 0.7 is drawn on the top of the bars. A red solid vertical line representing the population slope β_1 and a blue dashed vertical line representing the mean of the least-squares estimate b_1 are drawn. The red and blue lines almost coincide at round slope equals negative 0.945. [\[Return to Figure 13.12\]](#)

Figure 13.13 Image Description: Two histograms are presented side by side. The graph on the left panel is a histogram of the distribution for the conditional mean. The y-axis labelled “Density” is in increments of 0.5 from 0 to 2. The x-axis labelled “Conditional Mean (in \$1000)” is in increments of 1 from 5 to 10. The graph consists of bars and a black bell-shaped curve ranging from 6.5 to 8.5 is drawn on the top of the bars. A red solid vertical line representing the population conditional mean and a blue dashed vertical line representing the mean of the sample conditional mean are drawn. The red and blue lines almost coincide at round mean equals 7.5. The graph on the right panel is a histogram of the distribution of a single value response. The y-axis labelled “Density” is in increments of 0.1 from 0 to 0.5. The x-axis labelled “A single Response (in \$1000)” is in increments of 1 from 5 to 10. The graph consists of bars and a black bell-shaped curve ranging from 4 to 10 is drawn on the top of the bars. A red solid vertical line representing the population response

Y and a blue dashed vertical line representing the mean of the fitted value \hat{Y} are drawn. The red and blue lines almost coincide at single value equals 7.5. [[Return to Figure 13.13](#)]

Versioning History

This page provides a record of edits and changes made to this book since its initial publication in the MacEwan Open Books collection. Whenever the authors make edits or updates to the text, they provide a record and description of those changes here.

If the change is minor, the version number increases by 0.1. If the edits involve substantial updates, the version number goes up to the next full number. The work presented on our website always reflects the most recent version.

Version	Date	Change Details