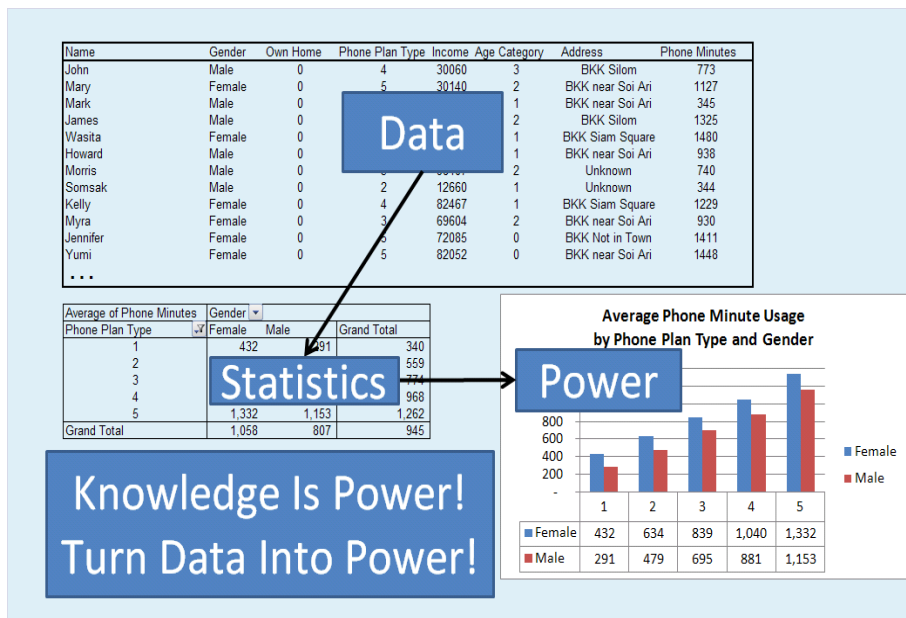


An Introduction to Business Statistics: 2nd Edition



By Arthur Dryver, Ph.D.

An Introduction to Business Statistics: 2nd Edition

By Arthur Dryver, Ph.D.

ISBN: 978-974-287-162-8 (hard copy version)

Editor: Bruce Leeds, Ph.D.

URL: www.BruceLeedsEditing.com

E-mail: bruceleeds@gmail.com

Published by:

Learnviaweb.com

URL: www.Learnviaweb.com

1st **Published** September 2008

Copyright by: Arthur Dryver, Ph.D. © 2008

Graduate School of Business Administration

National Institute of Development Administration

118 Seri Thai Road, Bangkok 10240 TH

E-mail: dryver@gmail.com

Contents

Preface

Significance of the book

Most introductory statistics books, even the applied books, do not cover preparing presentations for upper management. One of the unique features of this book is covering how to take statistical findings and turning them into presentations for upper management. Explaining statistical findings, especially advanced statistical findings in manner that people without a statistical background can understand can be difficult. Many managers may not calculate statistics at work, but will have to report the statistical findings of people they are managing. Thus this is essential for MBA's students. In addition, this book often writes in first person in order to better reach the reader (?). There are exercises for students which requires analyzing databases using software as opposed to the typical exercises that are to be done with a calculator.

Supplemental Material

Simulated datasets for homework or to play with can be found from one of the author's website www.learnviaweb.com/datasets/datasets.html In addition, other in-

structional material such as videos online to teach Microsoft Excel and SPSS can be found at the author's website www.learnviaweb.com.

There exists electronic versions, PDF, of the book comes in two formats, one is formatted for printing and the other to fit the computer screen and created for the instructor. The electronic version of this book contains in essence an infinite number of problems with solutions on such topics as binomial probability, normal probability, hypothesis testing, etc. by using JavaScript for random number generation. The click of a button can generate the question with a new set of numbers and another click can generate the answer, thus allowing the practice of problems until a deeper understanding is obtained, as opposed to traditional books with a limited number of problem and solution sets. Finally, there is a small PDF with select exercises that at present can be downloaded free where the material related to the author's book can be found.

Acknowledgments

In the process of completing this book, there are many people that have been great sources of inspiration and support. First, I would like to thank Steve Thompson, my adviser, for laying a strong foundation for me in the area of statistics, particularly for being the inspiration in the sampling realm. Then I would like to thank the Graduate School of Business Administration, the National Institute of Development Administration (NIDA) for the support for this textbook. I am greatly in debt to my parents, Morris and Myra Dryver, and my brother Howard Dryver for their love and support. I would also like to thank my father-in-law, Somsak Boonsathorn, for his support and encouragement on this endeavor. A special thanks goes to Bruce

Leeds for his expertise in editing the book and valuable comments. Finally,I would like to express heartfelt gratitude to my wife, Wasita Boonsathorn,who has always been there for me to provide me with unconditional love and support throughout the process of writing this book.

1

Introduction

1.1. The Importance Of Statistics

Good decisions often start from good information. This is true within the business world as well. Some examples include:

1. Warranties: Should you offer a warranty on your product? If so, how long should you make the warranty? What percent of your product will break before the warranty expires?
2. Production: What is the expected demand for your product for next quarter? How much should you produce given the expected demand?
3. Quality: Does the machine you use produce a high quality product? Are the specifications (e.g. length of ruler) close to the desired specifications and how

close?

4. Market Research: Who are your present customers? What are their attributes? What do your customers and the customers of competing products think of your product?
5. Churn: What is the churn rate of your customers, and how often are they switching from your company to another for the same service?
6. Risk Models: What is the risk of extending credit to an individual? What are the odds he/she will pay?
7. Response Models: To whom should you mail for a mail out campaign, and who is most likely to respond?
8. Etc.: Much, much, more.

1.2. Introduction to Descriptive Statistics

There are many descriptive statistics, but not all are used regularly. The most commonly used descriptive statistics include frequency, proportion, sample mean, median, minimum, maximum, range, mode, variance, standard deviation (std.), and standard error (std. error). Some of the most common statistics can be put into two categories:

1. measures of central tendency - the expected value of the data, examples:
 - mean
 - median

- geometric mean

2. measures of dispersion - how spread out is the data, examples:

- variance
- standard deviation
- range

For supplemental material the reader should consider the book by ? or ?.

Definitions

Frequency refers to the number of observations with a specified characteristic. The sample proportion is the percent of observations with a specified characteristic within the sample. The sample proportion is denoted

$$\hat{p} = \frac{X}{n},$$

where X is the number of observations with the specified characteristic and n is the number of observations. For example, you observe 100 people and 45 of the people are women. Thus, $n = 100$ and $X = 45$, and the sample proportion of women is $\frac{45}{100} = .45$. The sample mean is

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

When x_i is defined as either zero or one depending on whether or not the observation meets the specified characteristic, then $\bar{x} = \hat{p}$. Referring to the latter example, define $x_i = 0$ for men and $x_i = 1$ for women. Then X is the sum x_i ; that is $X = \sum_{i=1}^n x_i$ and thus $\bar{x} = \hat{p} = .45$. The median is the 50 percentile or the midpoint of the data ordered from smallest to largest. The minimum (min) and maximum (max) are the smallest and largest data point, respectively. The mode is the most common observation. There may be more than one mode or no mode in any particular data set. The sample variance is

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

the sample standard deviation is

$$s = \sqrt{s^2},$$

and the standard error is

$$\frac{s}{\sqrt{n}}.$$

The sample standard deviation is used for understanding the spread of the data and the standard error is used for understanding the spread of \bar{x} from a sample of the same size, n .

Rare but extreme values can have a large effect on some of the descriptive statistics, these rare extreme values are called outliers. For example, the value 21 in sample # 6 in Table ?? might be considered an outlier. The median is an example

Sample #	the data	Min	Max	Median	Mean	s^2	std.	std. error
1	1,1,1,1,1	1	1	1	1	0	0	0
2	5,5,5,5,5	5	5	5	5	0	0	0
3	3,4,5,6,7	3	7	5	5	2.5	1.58	0.707
4	7,4,5,6,3	3	7	5	5	2.5	1.58	0.707
5	1,3,5,7,9	1	9	5	5	10	3.16	1.414
6	1,1,1,1,21	1	21	1	5	80	8.94	4.000

Table 1.1: A few fictitious samples for understanding descriptive statistics.

of a descriptive statistic unaffected by outliers.

Less commonly used descriptive statistics include coefficient of variation, first quartile (Q_1), third quartile (Q_3), range, geometric mean, and geometric mean rate of return. The coefficient of variation is the standard deviation divided by the mean,

$$C.V. = \frac{s}{\bar{x}}.$$

The first quartile, Q_1 , has 25% of the observations less than it and 75% of the observations greater than it. The third quartile, Q_3 , has 75% of the observations less than it and 25% of the observations greater than it. The range is defined as the maximum minus the minimum. The formula for the geometric mean is

$$\bar{x}_g = [x_1 \times x_2 \times \cdots \times x_n]^{1/n}.$$

The formula for the geometric mean rate of return is

$$\bar{r}_g = [(1 + r_1) \times (1 + r_2) \times \cdots \times (1 + r_n)]^{1/n} - 1,$$

where r_i is the rate of return for period i . The geometric mean rate of return is best for understanding how well an investment is doing (??). For example, you invested \$1,000 in a stock and the value decreased to \$500 by the end of the first year and then increased by \$250 by the end of the second year. After two time periods, beginning with \$1,000 you ended up with \$750. The mean rate of return equals

$$\bar{x} = \frac{-50\% + 50\%}{2} = 0,$$

which is misleading, as there is a loss of 25%, whereas the geometric mean rate of return equals

$$\bar{r}_g = [(1 - .5) \times (1 + .5)]^{1/2} - 1 = (.75)^{1/2} - 1 = .866 - 1 = -0.134,$$

indicating a loss since it is less than 0. Note: \$1,000 \times .866 \times .866 = \$750

1.2.1 Introduction to Parameters

Almost always, if not always, simple statistics are used to obtain a better understanding of a *population* of interest. A population is the set of all units of interest. Parameters are quantities that represent the population. The parameters of interest are estimated using statistics. If you were interested in all the people in Bangkok, then the population would consist of all people in Bangkok. The sample is a subset of the population. Sampling will be discussed in further detail in Chapter ??.

The population quantities for the proportion, mean, finite population variance

and the standard deviation are

$$\pi = \frac{X}{N},$$

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i,$$

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

and

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

respectively. Where capital N refers the number of units in the population.

1.3. Descriptive Statistics With Microsoft Excel

Microsoft Excel is the standard software used for presenting data. Almost everyone has it on their computer and it is very user friendly. In addition, most people are already somewhat familiar with Excel. Many books cover how to do statistics with Microsoft Excel as well (??)

There might be a few things of interest in looking at this data. One thing might be the average salaries for women and men, which can be calculated, and are

$$\bar{x}_w = \frac{\$47,647 + \$48,489 + \$61,617 + \$79,286 + \$58,963}{5} = \$59,200,$$

$$\bar{x}_m = \frac{\$79,343 + \$74,787 + \$78,616 + \$67,458 + \$54,894}{5} = \$71,020,$$

respectively. Another thing of interest might be the frequency of each gender at each

level. The frequency of gender at each level is called a *contingency table* or *cross classification table*, often called simply a cross table, or crosstab. Table ?? contains an example of a cross table calculated from the data in Table ??.

Gender	Level	Salary
Female	1	\$47,647
Female	4	\$48,489
Female	3	\$61,617
Female	4	\$79,286
Female	2	\$58,963
Male	4	\$79,343
Male	5	\$74,787
Male	5	\$78,616
Male	5	\$67,458
Male	3	\$54,894

Table 1.2: Sample data to illustrate the power of Excel

The frequency of a single variable, for example level, is called a *frequency table*. A graphical representation of a frequency table is easy to make within Excel. A histogram is a graphical representation of the frequency of the data without spaces between the bars. Figure ?? is an example of a histogram of salary for Table ??. Honestly, since the amount of data is very small it is easy to calculate. Were we working with a larger data set, even basic calculations, such as addition, can be very time consuming without the help of computers. What was done by pen and paper

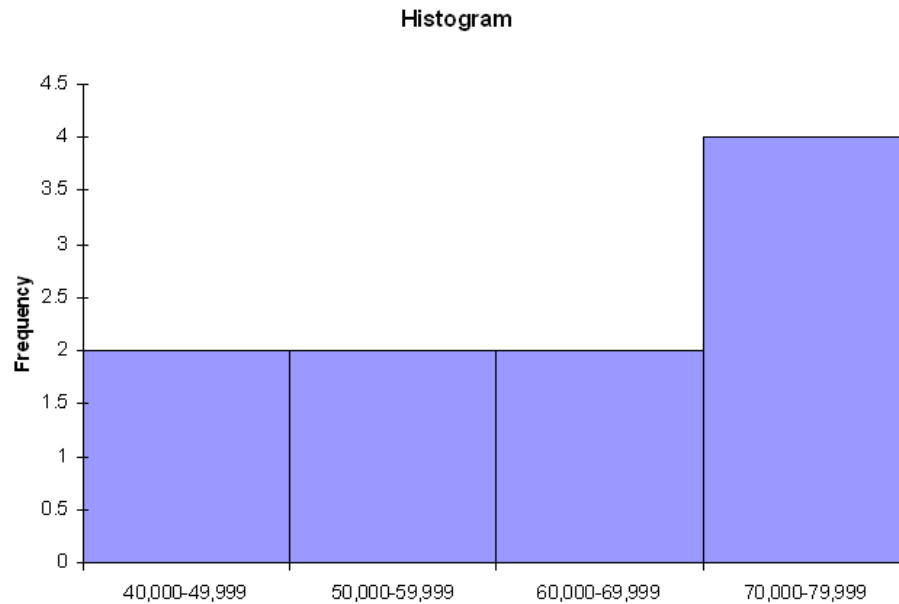


Figure 1.1: Example of a histogram for salary from Table ??.

can be done very quickly, easily and presented nicely using Excel. A function within Excel that is very useful for analyzing data is called "PivotTable and PivotChart Report..." under the tab "Data". The following pages illustrate how this can be done in Microsoft Office Excel 2003. Using other versions of Excel, the steps are very similar if not identical. For extensive training videos on Excel and other educational videos go to www.learnviaweb.com/videos.

Level	Female	Male
1	1	0
2	1	0
3	1	1
4	2	1
5	0	3

Table 1.3: Frequency at each level broken out by gender from Table ??.

1.4. Descriptive Statistics to Create a Marketing Presentation

In this section is an example of a marketing project utilizing **only descriptive statistics to gain valuable knowledge**. Figures ?? through ?? represent the resulting marketing presentation. A sample data set for this example marketing project can be found at www.learnviaweb.com/videos. Imagine working at a consulting firm and asked to do the following project:

Your client is the cell phone division within a telecommunications company (Example: DTAC). You want to give your client an understanding of its customers. Who are they, etc. In addition, you want to give your client an understanding of how valuable each customer is. Your client is the cell phone division within a telecommunications company. You want to give your client an understanding of their customers. Who are their customers and how valuable or not each specific customer is. To answer this you must first answer what is value? The client has given you the entire

1.4. DESCRIPTIVE STATISTICS TO CREATE A MARKETING PRESENTATION19

dataset, and you will work with the entire dataset. Such information will include payment history. Keep in mind that one of your clients' concerns is that some of its customers do not pay. Remember the question what is value? This is somewhat subjective. The client is concerned about people not paying though. Always listen to your client!!! Listening helps, even statistical projects. Note: Often a large part of success (assuming honesty) in consulting is a happy client. You can not ensure a happy client if you never listen to him or her.

The file you have been given includes the following customer information.

- Name
- Gender (male=1)
- Home owner or not (own=1)
- There are 5 mobile phone plan types, listed below, each with different minimum minutes required and pricing:
 - 1: no minutes charge 4.0 baht/minute (No minimum payment)
 - 2: 200 minutes charge 3.5 baht/minute (Pay at least 700 baht)
 - 3: 400 minutes charge 3.0 baht/minute (Pay at least 1200 baht)
 - 4: 600 minutes charge 2.5 baht/minute (Pay at least 1500 baht)
 - 5: 800 minutes charge 2.0 baht/minute (Pay at least 1600 baht)
- Customer income
- Government job or not (if working in government=1)

- Age category
- Location
- Minutes (the total number of minutes used for the most recent month)
- Payment history. You are given information on the past 24 months for each customer as to whether the customer paid on time or was late and how late.
 - 0=current, not late on payment
 - 1=30 days late
 - 2=60 days late
 - 3=90 days late
 - 4=120+ days in default, and are not expected to pay, very bad

1.4.1 The Marketing Presentation

The following pages consist of a sample marketing presentation for the example project in this section

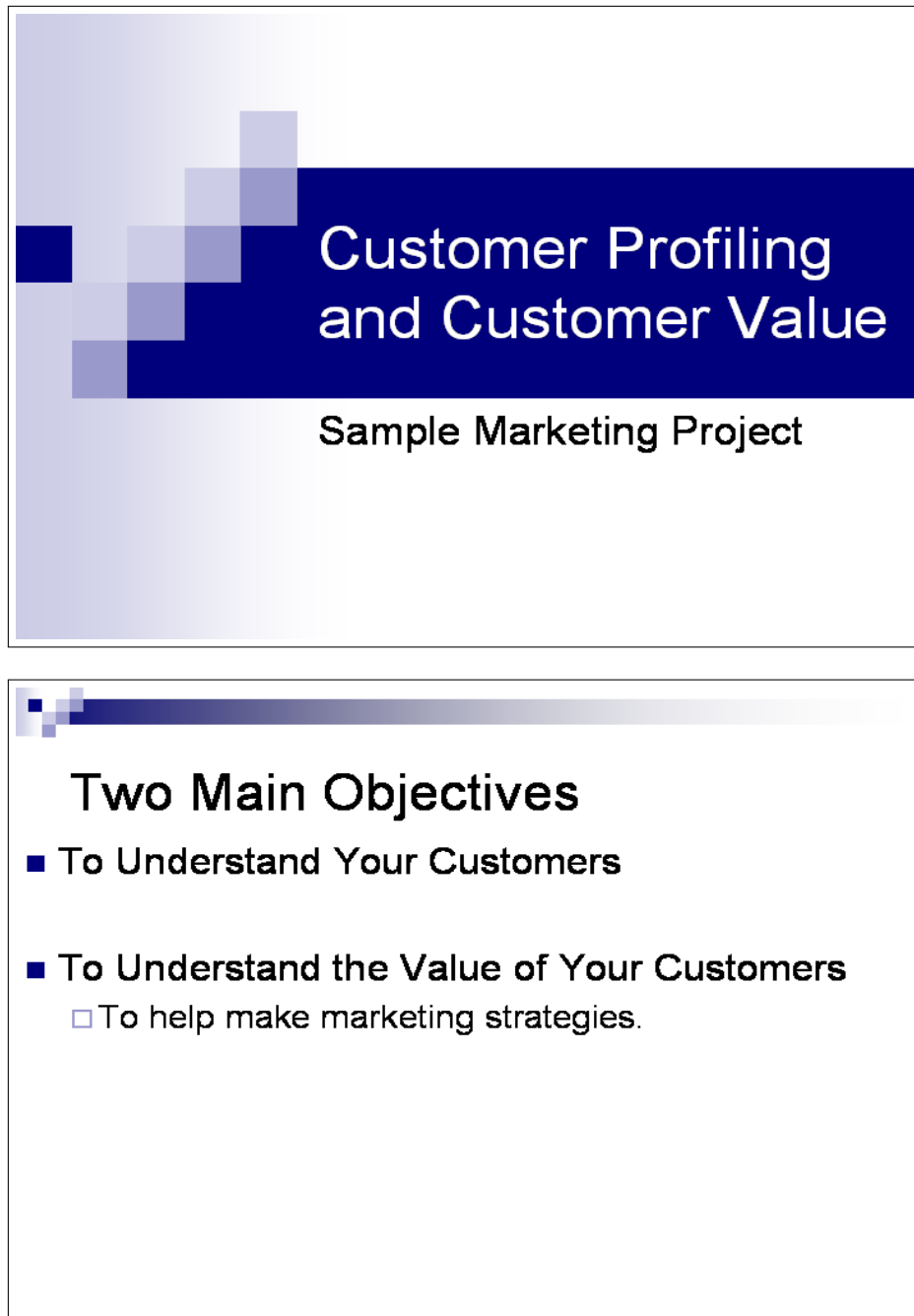


Figure 1.2: Slides one and two of the marketing presentation



First, Who Are Your Customers

We Looked At All 15,045 Customers To Understand Who They Are

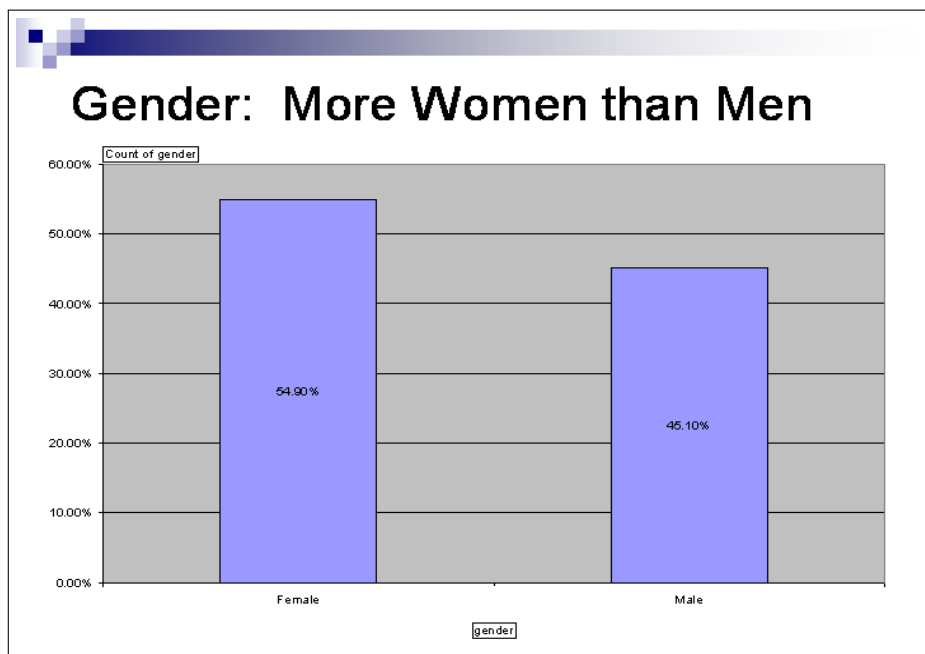


Figure 1.3: Next two slides of the marketing presentation

1.4. DESCRIPTIVE STATISTICS TO CREATE A MARKETING PRESENTATION23

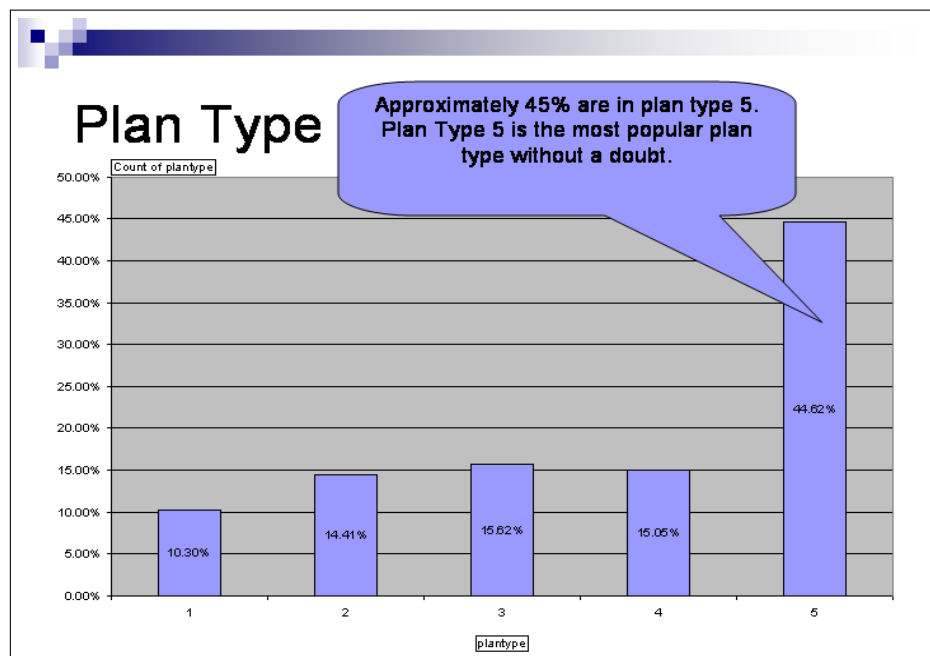
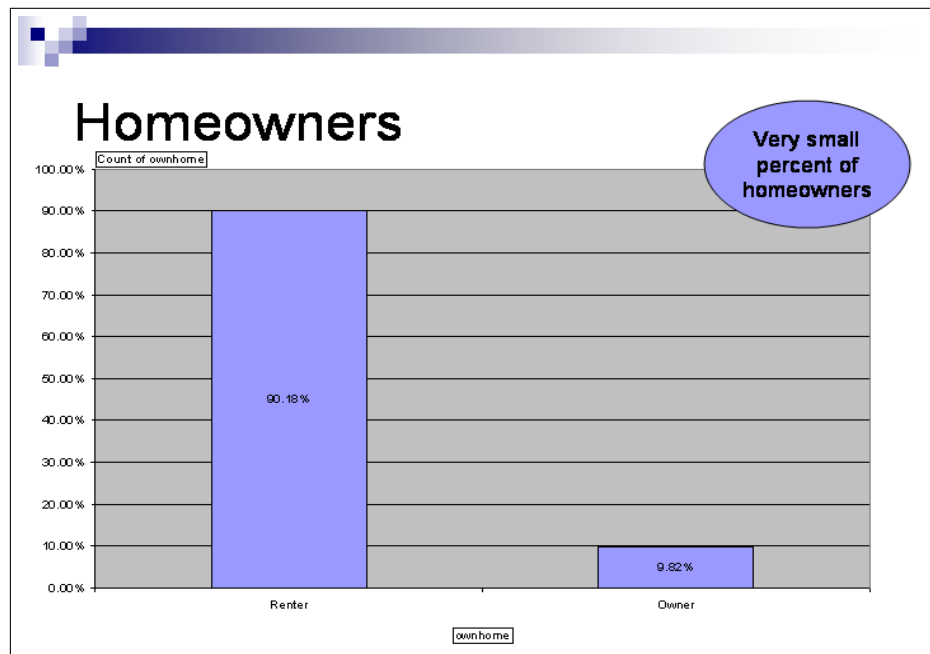
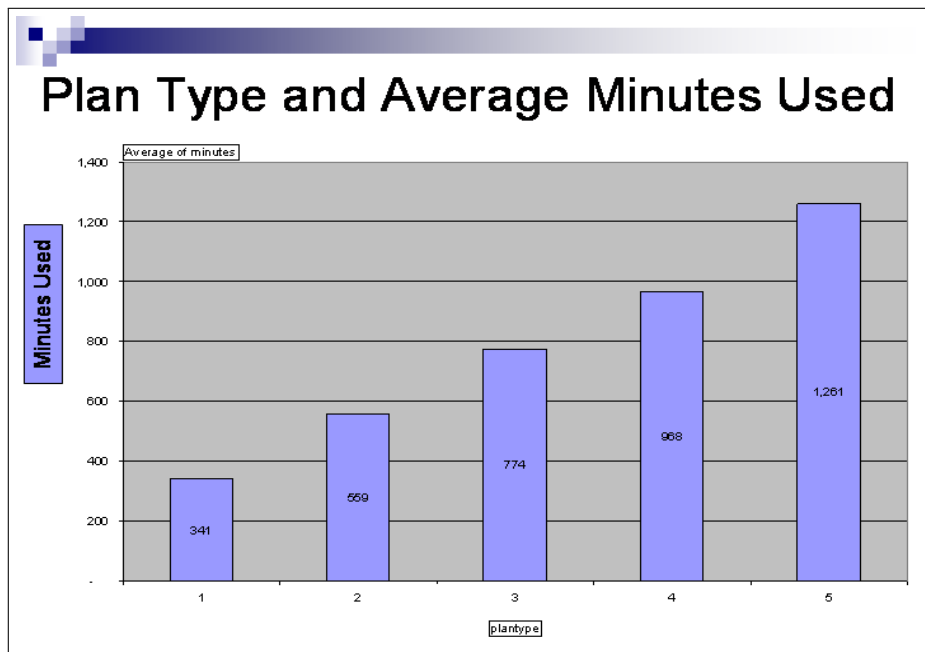


Figure 1.4: Next two slides of the marketing presentation



Plan Type and Mean Minutes Used

Plan Type	N	Mean	Minimum	Maximum
Type 1: 0	1550	341	77	648
Type 2: 200	2168	559	285	846
Type 3: 400	2350	774	481	1052
Type 4: 600	2264	968	679	1248
Type 5: 800	6713	1261	478	2634
Overall	15045	945	77	2634

Only in plan type 5 do the customers use less than the minimum. In this plan type, customers give you "free" money. The total "free" money is \$62,942 for December 2005, 0.2% of revenue, total=\$34,282,000.

Figure 1.5: Next two slides of the marketing presentation

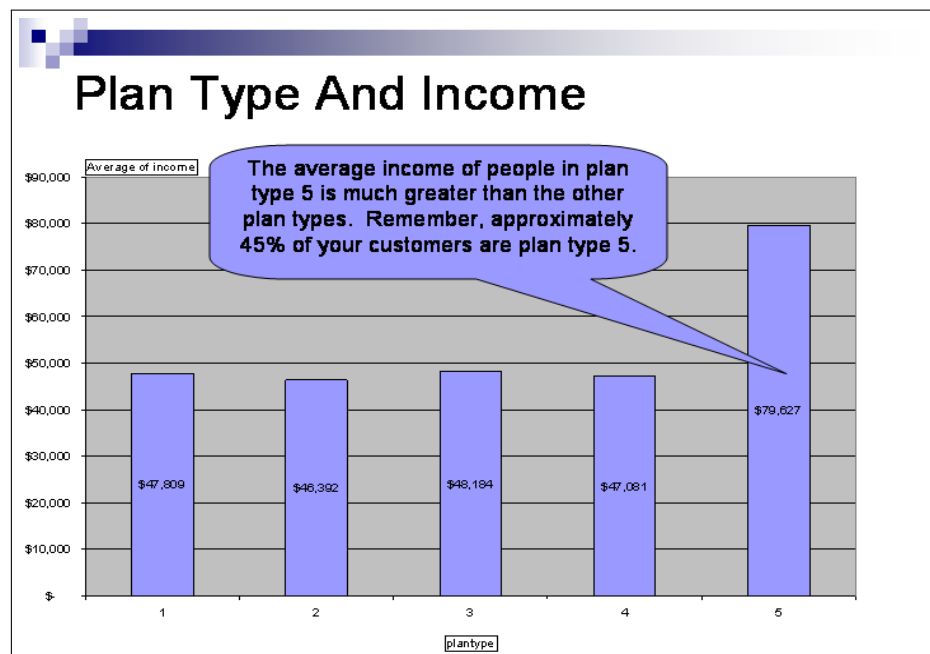
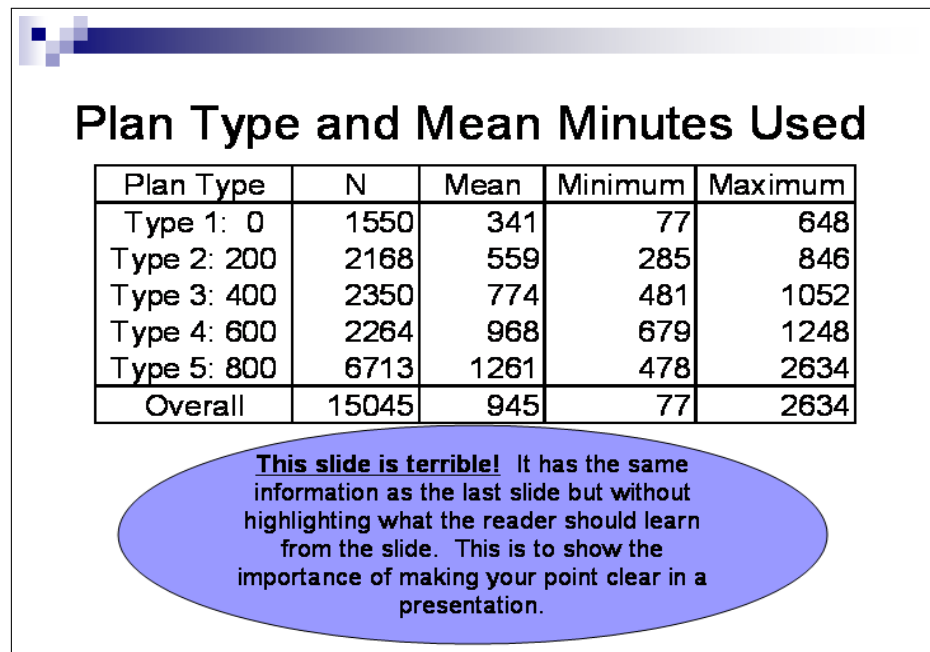


Figure 1.6: Next two slides of the marketing presentation

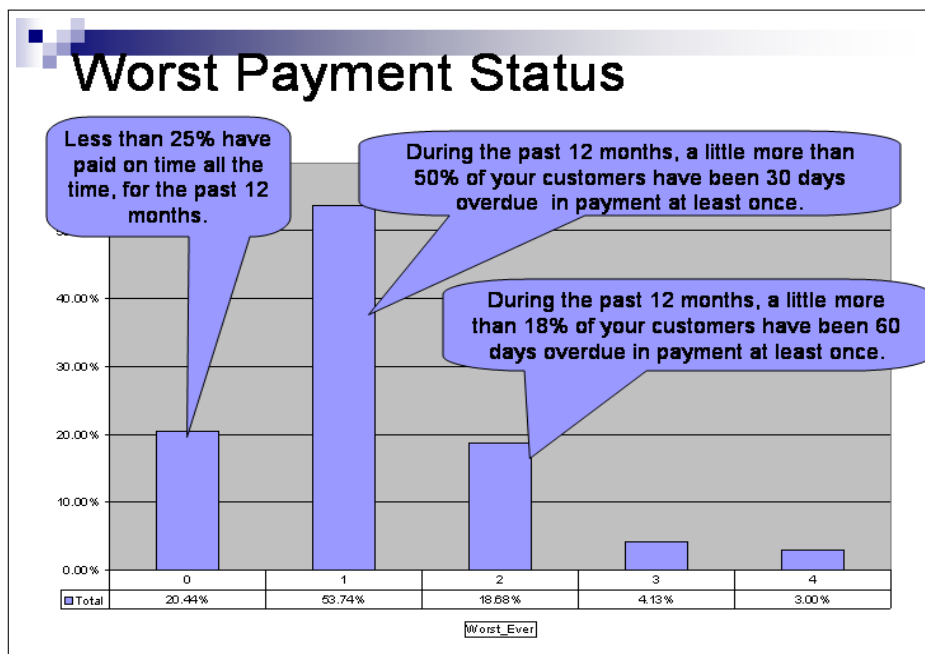
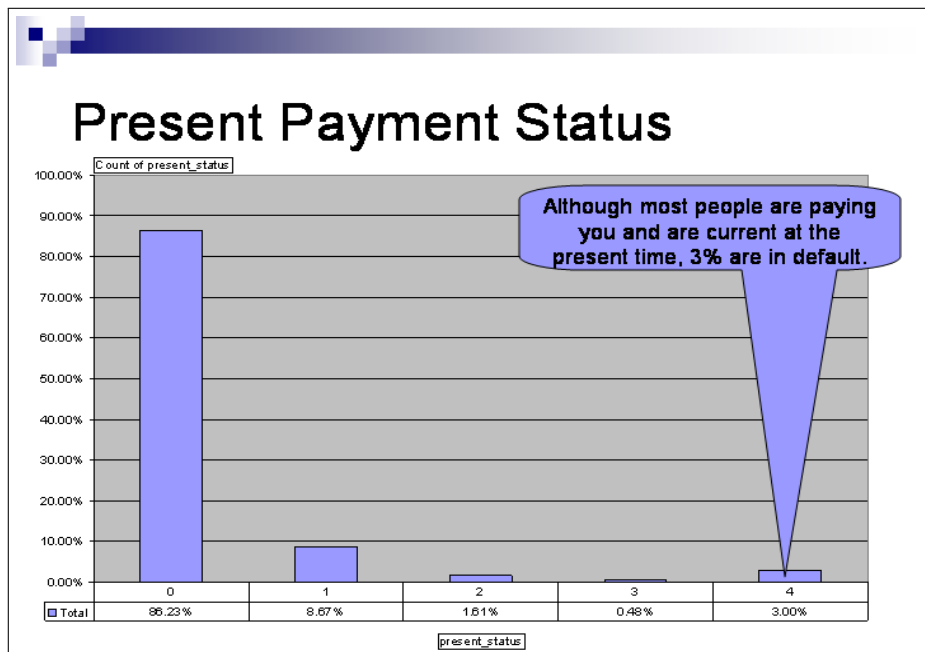
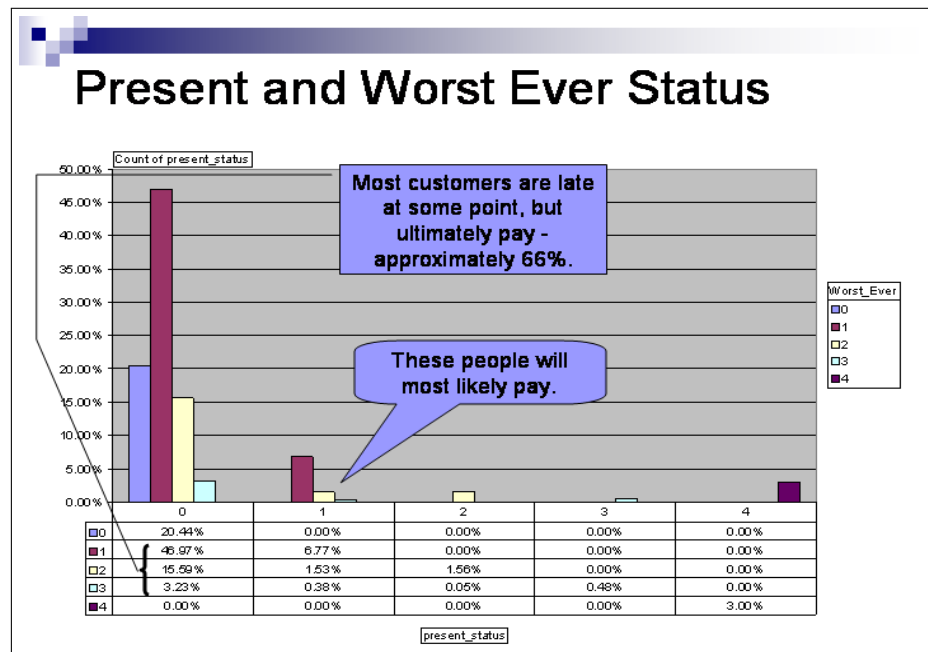


Figure 1.7: Next two slides of the marketing presentation



You Should Do Many More Graphs ...

- Also, the graphs should be made much prettier.

Figure 1.8: Next two slides of the marketing presentation

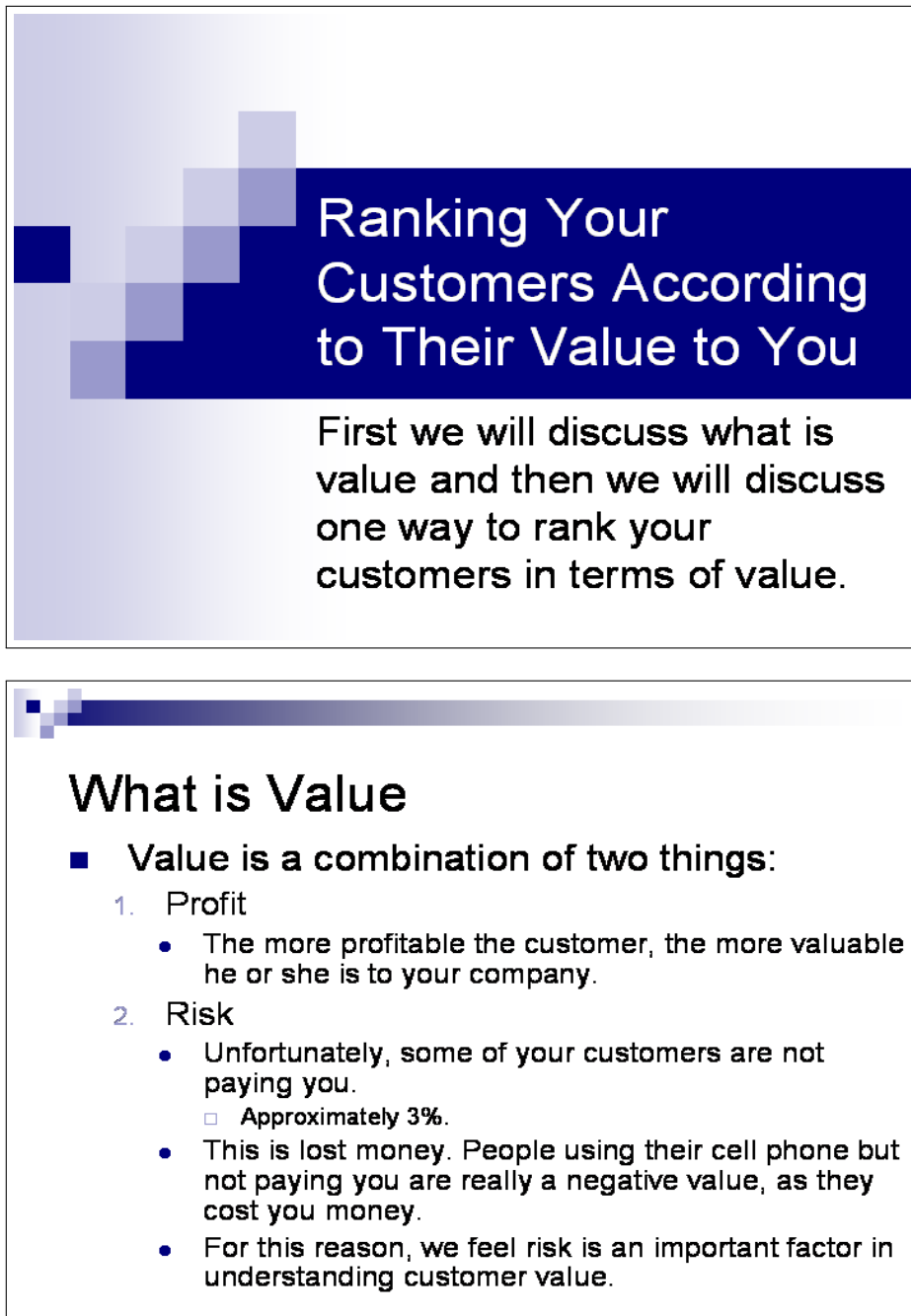


Figure 1.9: Next two slides of the marketing presentation

Profit

- Due to the sensitive nature of profit margins, we could not use actual profit.
- Thus, to understand your most profitable customers we used revenue.
 - Revenue was calculated using plan type and minutes used. (Should give more details).

Revenue: We created Five Categories For Revenue

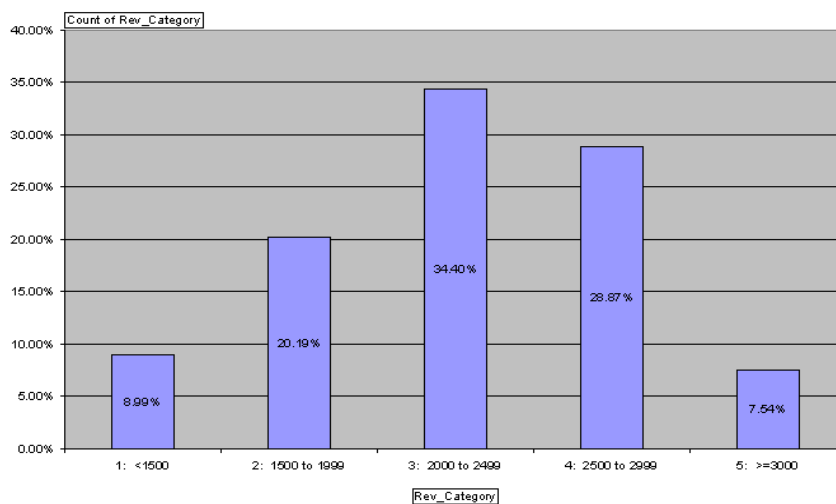


Figure 1.10: Next two slides of the marketing presentation

Risk

- To understand Risk we looked at all 12 months of the payment history data.
- Ultimately we decided to use the worst ever payment status for the past 12 months as a proxy for risk. The higher the value, the higher the risk.
- Thus Risk has 5 categories, with values ranging from 0-4; the value 0 being the least risky and 4 being the most risky. A value of 4 actually means the person is in default.

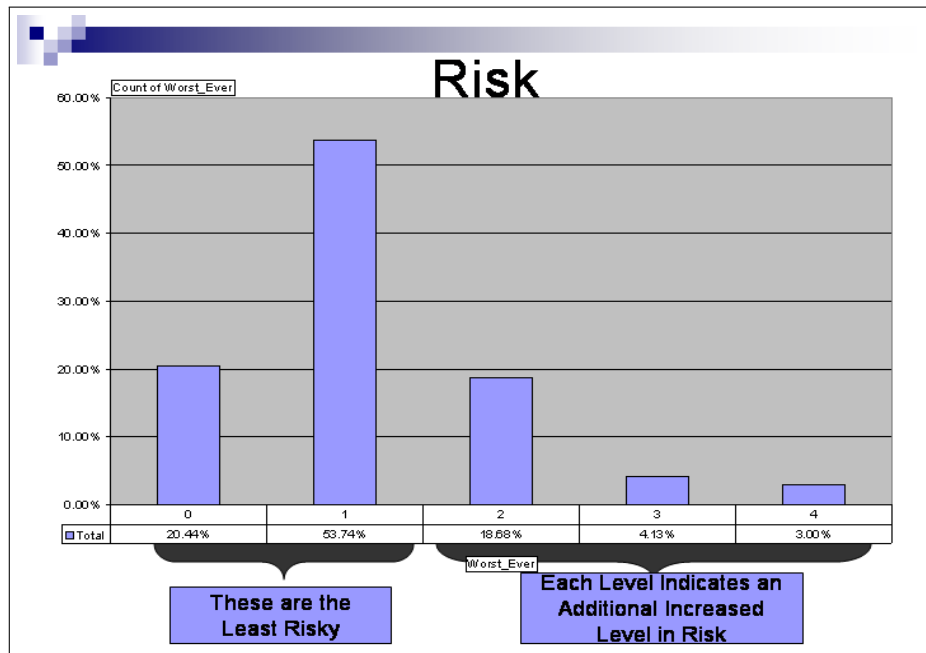


Figure 1.11: Next two slides of the marketing presentation

1.4. DESCRIPTIVE STATISTICS TO CREATE A MARKETING PRESENTATION31

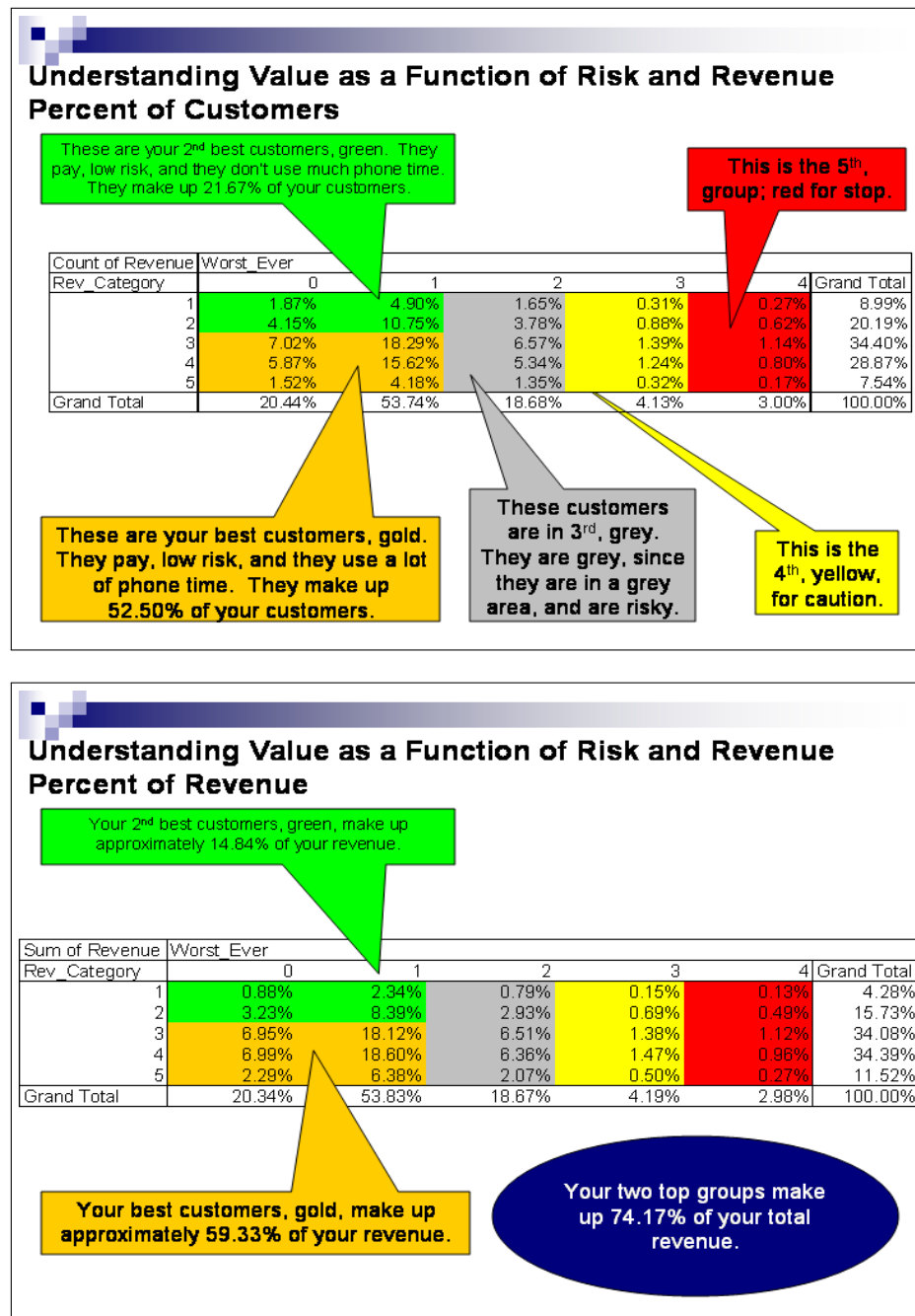


Figure 1.12: Next two slides of the marketing presentation

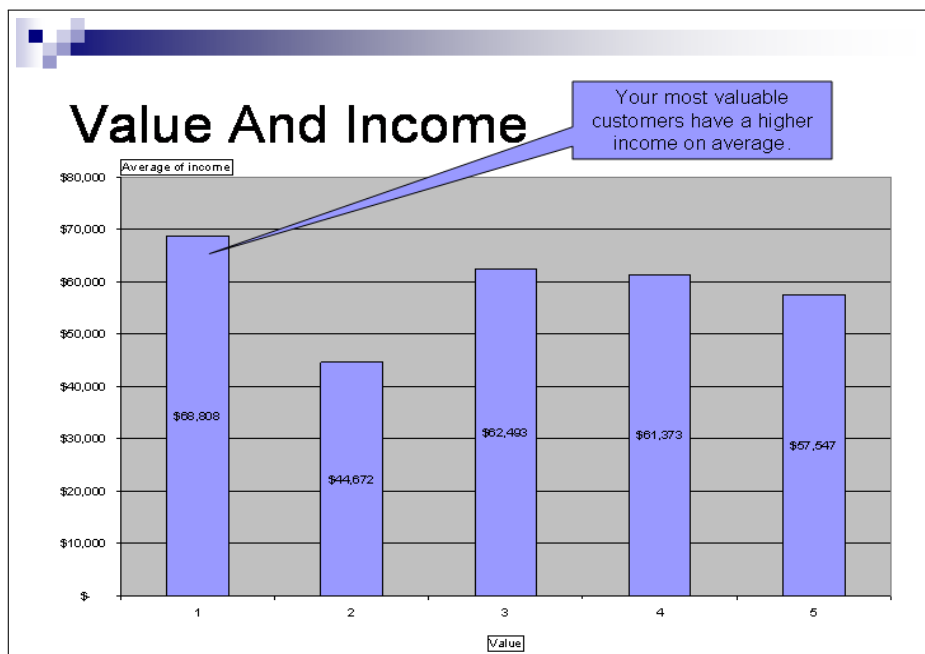
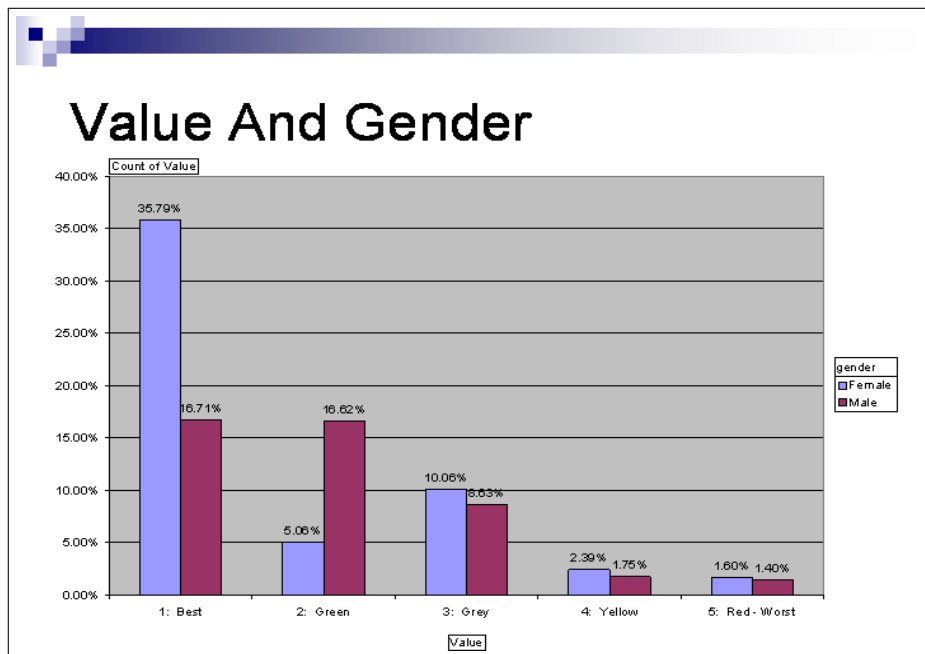


Figure 1.13: Next two slides of the marketing presentation

1.4. DESCRIPTIVE STATISTICS TO CREATE A MARKETING PRESENTATION33

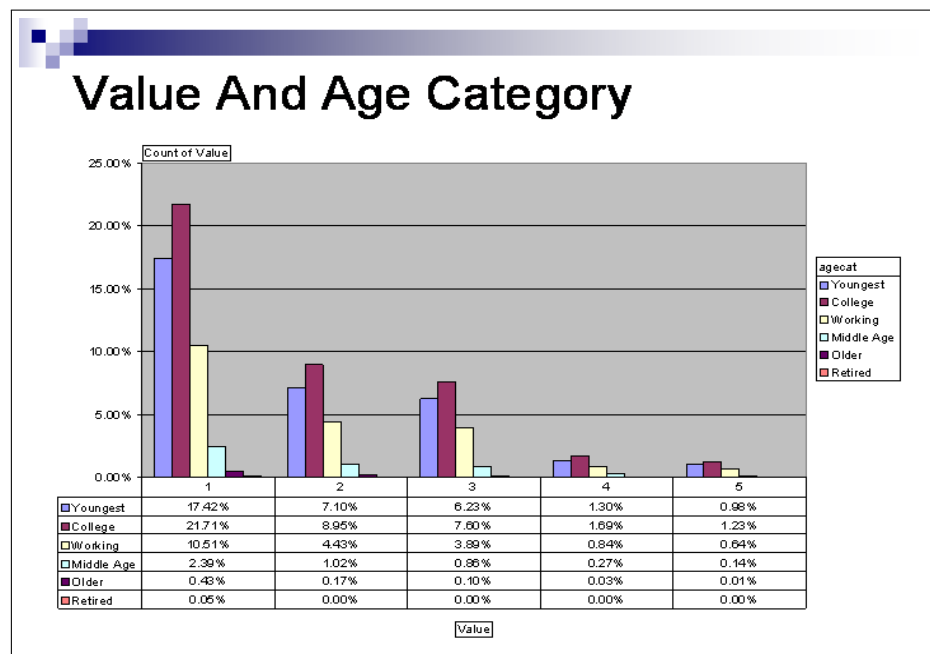
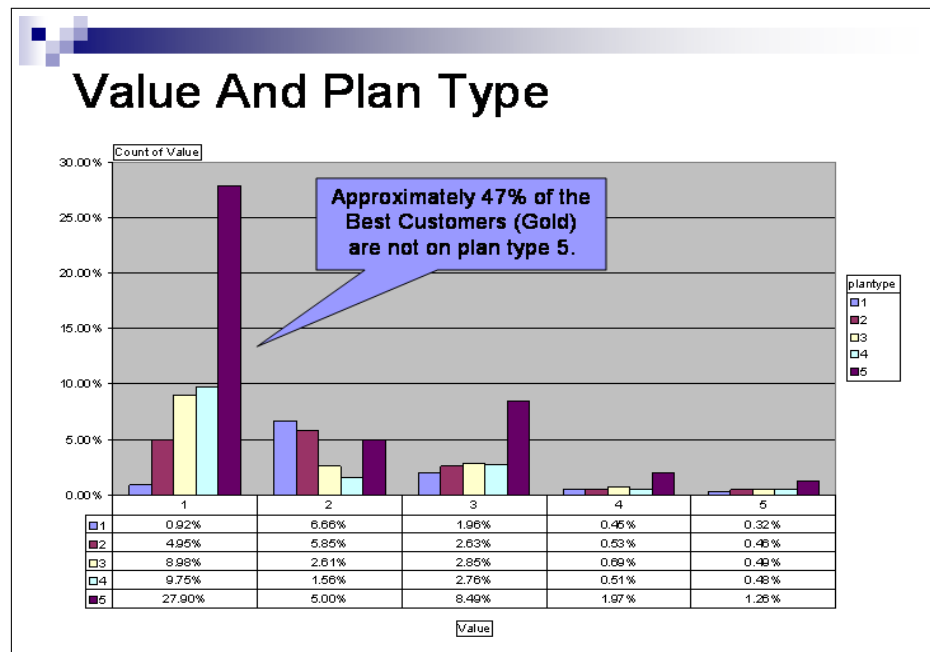




Figure 1.14: Next two slides of the marketing presentation



More Graphs For Understanding Value.

- More Pretty Graphs



Recommendations

- You want to keep your highest value customers happy.
- Consider creating an additional phone plan since more than 40% of your customers are in phone plan type 5.
 - This recommendation is in part a result of finding customers that do not even use the minimum minutes for plan type 5.
- Most of your customers pay late at least once in 12 months.
 - Consider charging late fees to increase revenue.
 - Adding all of the 12 months late for all of your customers, that are not in default equals approximately 27,478 late fee charges per year. If you charge a 10 baht per month late fee, you could make approximately an additional 274,780 baht per year.

Imagine a project costing 500,00 baht leading to an extra 274,780 baht/year in addition to doing the requested work. You have proven you're worth and more as a consultant. Statements such as the one above should be checked with the client before making them. Are they already using late fees?

Figure 1.15: Last two slides of the marketing presentation

1.4.2 High Level Take Away From This Sample Project

Do not underestimate the power of basic statistics. A lot can be learned from basic graphs and cross tables, and presentation is very important. The descriptive statistics are easy to calculate using computer software – organizing that computer output to learn something is the hard part. Be sure you leave time to make the presentation; do not spend too much time on the statistics and get overwhelmed.

1.4.3 General comments on a successful project

First, what is a successful project? Success is a happy client. Of course it is doing the work correct as well. Correct does not mean it has to be the most complicated, the best statistics, etc. The client wants insight into his or her problem, not a statistics lesson and confusion. I can not stress this concept enough. Finally, if you know that a client wants something, you need to do it. If you think the client is wrong, this needs to be discussed with the client politely before you present the deliverable to him or her and his or her boss or subordinates. There are three main possible outcomes from not presenting what was requested without discussing beforehand, and they are all bad:

1. You are right and make your client lose face. Very bad; you probably will not get repeat business.
2. You are wrong and the client is right, you lose face. Very bad as well; you probably will not get repeat business.

3. You are both wrong, you still lose face for presenting the wrong material and not doing what was requested. Very bad as well; you probably will not get repeat business.

This can be avoided with a discussion before the presentation of the deliverable. Often consultants discuss high level findings and what to expect with their client before the actual presentation.

1.4.4 The importance of the presentation

A major but common mistake is spending too much time on the statistics and not enough time on the presentation. The presentation is the deliverable; it is what you are graded on in real life. It can provide everyone with insight, or confusion. Honestly, this is something I had a difficult time with as well when doing projects. It was the most difficult thing for me to accept, but after being responsible, managing projects, and with enough client interaction, I learned and understood. "You can have brilliant ideas, but if you can't get them across, your brains won't get you anywhere." Lee Iacocca (?). Many times, if people do not understand your findings, they will not use them and they may not even know how to use them – thus, making all your hard work worthless! Yes, this really happens. Some people sometimes do great analytical work but cant explain it, making everyone unhappy. Many of you may work with statisticians at one point in the future. These statisticians may use advanced statistical techniques, but they can almost always be presented in a simple

manner. *How*, will be covered later in future chapters and in greater detail. Try to get the statistics into a simpler form for presentation. Do not present a bunch of stuff you do not understand. You are expected to know what you are presenting. You may think it is OK if you do not understand but the statistician does understand. It is often not OK. If the statistician can not explain it to you, there is a good chance he can not explain it to others. People do not buy/trust what they do not understand and they often do not even bother to listen after awhile. You as management will be responsible for avoiding this situation when delivering presentations.

1.5. Descriptive Statistics Examples

EXERCISE 1.5.1. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

18.20	26.80	13.00	13.10	11.70
-------	-------	-------	-------	-------

Table 1.4: The stock prices

EXERCISE 1.5.2. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

15.00	17.80	19.30	12.50	17.20
-------	-------	-------	-------	-------

Table 1.5: The stock prices

EXERCISE 1.5.3. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

26.20	28.80	19.10	14.70	20.30	11.10	16.90
-------	-------	-------	-------	-------	-------	-------

Table 1.6: The stock prices

EXERCISE 1.5.4. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

10.80	12.80	14.20	29.30	24.40
-------	-------	-------	-------	-------

Table 1.7: The stock prices

EXERCISE 1.5.5. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

12.20	10.40	28.90	14.30	26.20	22.70	14.10	16.60	22.30	15.60
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 1.8: The stock prices

1.6. Exercises

1.6.1 Descriptive Statistics Exercises

EXERCISE 1.6.1. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

17.70	30.70	10.70	35.80	12.20	19.20	14.10	11.80	30.90	24.60
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 1.9: The stock prices

EXERCISE 1.6.2. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

36.20	16.80	14.20	16.30
-------	-------	-------	-------

Table 1.10: The stock prices

EXERCISE 1.6.3. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

16.40	10.40	16.60	17.20	27.10	16.10	21.60	19.10
-------	-------	-------	-------	-------	-------	-------	-------

Table 1.11: The stock prices

EXERCISE 1.6.4. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

19.70	21.50	26.40	10.20	18.30	12.40	28.30	11.80
-------	-------	-------	-------	-------	-------	-------	-------

Table 1.12: The stock prices

EXERCISE 1.6.5. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

16.40	22.10	33.10	19.90	14.40	14.90
-------	-------	-------	-------	-------	-------

Table 1.13: The stock prices

EXERCISE 1.6.6. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

27.60	14.10	11.00	18.90	22.40	12.10	12.30	11.20
-------	-------	-------	-------	-------	-------	-------	-------

Table 1.14: The stock prices

EXERCISE 1.6.7. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

12.10	17.30	16.60	20.10	10.80	35.60	16.50	12.30
-------	-------	-------	-------	-------	-------	-------	-------

Table 1.15: The stock prices

EXERCISE 1.6.8. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

12.70	20.40	17.50	16.40	15.40	22.80	18.60	17.30	19.40
-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 1.16: The stock prices

EXERCISE 1.6.9. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

16.70	21.70	13.70	14.70	23.60	15.50	12.90	15.80	23.50
-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 1.17: The stock prices

EXERCISE 1.6.10. Solve for the sample min, max, mean, median, variance, and standard deviation of the following stock prices of the following companies:

21.60	10.80	14.30	17.00	16.10	17.00	10.90	17.60	12.10	15.20
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 1.18: The stock prices

1.6.2 Multiple Choice

Click "Begin" and when you are finished, click "End". Enjoy.

[Begin Multiple Choice Questions](#)

1. Is the mean, \bar{x} , affected by outliers?

(a) Yes

(b) No

2. Is the median affected by outliers?

(a) Yes

(b) No

3. Is the sample standard deviation, s , affected by outliers?

(a) Yes

(b) No

4. The mean income in Bangkok is 25,000 Baht/month. Thus 50% of all people in Bangkok make more than 25,000 Baht/month.

(a) True

(b) False

[End Multiple Choice Questions](#)

1.7. Assignment

1. Individual assignment. Find four stocks within the SET, two of which you think are good investments and two of which you think are bad investments. Create graphs within Microsoft Excel from data on the four companies and

make a presentation illustrating why you believe the companies are either good or bad investments. Create only a few graphs, a maximum of twenty. If you are selected you will have to present in class and everyone must hand in the presentation. The purpose of this assignment is not to test you on your knowledge of stocks and finance, but to make you more comfortable in looking at data, using Excel, and making presentations. Grading **will not** be on your knowledge of the stock market. The assignment is due next week.

2. Group assignment for 3-4 students. Customer Profiling and Determining Customer "Value". Your client is a start up cable company. It has been in business for approximately one year. They offer cable TV channels and Cable (High-Speed) Internet.

- You will pretend that you are from a consulting firm presenting to a client with very little or no statistical background.
- You want to give your client an understanding of their customers – who they are and how "valuable" or not they are. To answer this you must first answer, what is value?
- Each group will present the project on the date it is due, three weeks from when it is assigned.
- The presentation should be done in PowerPoint Graphs, etc., and must be done in Excel.
- The PowerPoint must be less than 5MB in size.

- Finally, you will hand in the PowerPoint presentation and all supporting materials.
- You will be graded almost entirely on the PowerPoint Presentation. The presentation should be made as a stand-alone document. It should be made in a manner that it can be read and understood without your presence. Finally, presentations that show nothing beyond the given example presentation will not be considered "A" material.
- The client has given you the entire dataset, and you will work with the entire dataset. You will be given the entire population of customers with various information. Such information will include payment history. One of your clients concerns is that some of their customers dont pay.
- The client intends to use your findings to offer incentives to the specific people you determine as high value.
 - Thus if you decide person "A" is of higher or lower value than person "B", you must be able to justify your claim/statement.
- Data can be found at a link from my website:
www.learnviaweb.com/datasets/datasets.html.
- Below is the data file layout:
 - gender (male=1)
 - ownhome (own=1)
 - Internet=1 then on the internet plan. Charge is 2,000 Baht/month for unlimited usage

- plantype there are 5 plan types
 - (a) 1=iron - the minimum number of channels offered 500 Baht/month
 - (b) 2=copper - 750 Baht/month
 - (c) 3=jade - 1,500 Baht/month
 - (d) 4=gold - 2,000 Baht/month
 - (e) 5=platinum - the maximum number of channels offered 2,500 Baht/month
- Income: persons income, in Baht/Month
- govtjob (if work in government=1)
- agecat:
 - (a) 0=Teenager,
 - (b) 1=College,
 - (c) 2=Newly Working,
 - (d) 3=Middle Age,
 - (e) 4=50's,
 - (f) 5=Retired
- Address (this is to represent the person's actual address)
- hours (the total number of hours used on the internet for the most recent month)
- payment history (12 months the first status, status1, is 12 months ago the last one is the present status, for the month of December 2004)
 - (a) 0=current, not late on payment

- (b) 1=30 days late
- (c) 2=60 days late
- (d) 3=90 days late
- (e) 4=in default, and are not expected to pay, they are no longer
considered your clients

2

Probability

2.1. Introduction to Probability

First, what is probability? Probability is a number that can take values ranging from zero to one representing the likelihood of an event. The more likely an event will occur, the higher its probability and the less likely an event will occur, the lower the probability.

The concept of probability is used everywhere. The perceived probability of a positive or negative outcome occurring and the associated benefit/loss respectively is often used in making decisions. Some examples:

1. Should a person see the latest "Harry Potter" movie? The person will question him or herself as to whether or not he or she expects to enjoy the movie. In other words, is there a high probability or low probability of enjoyment if the

person sees the movie.

2. Should we invest money in the stock market or earn interest in the bank? This decision is in part made from tolerance for risk and perceived probability of success (make money) in the stock market.
3. Should we bring an umbrella with us today? This is motivated in part by the perceived probability of rain today.
4. Should a person try to obtain a Bachelor's degree? Many estimate the probability of obtaining their career goals without a Bachelor's degree to be very low and with a Bachelor's degree considerably higher. For this reason many people attend college.
5. Etc.

Definitions

- **Sample space:** Set of all possible outcomes and denoted \mathcal{S} .
- **Event:** A set of possible outcomes. A simple event is a single possible outcome from the sample space.
- **Probability:** The chance or likelihood of a particular event occurring. Probability is between zero and one. An impossible event has a probability of zero. An event that will definitely occur has a probability of one. The probability of event A is denoted $P(A)$, where $0 \leq P(A) \leq 1$, and $P(\mathcal{S}) = 1$.
- **Complement:** The complement of A consists of all possible events within \mathcal{S} excluding A . The event A^C denotes the complement of event A and $P(A \cup A^C) = 1$, where \cup represents "or". The union of $A \cup A^C = \mathcal{S}$.
- **Mutually exclusive:** When event "A" and "B" can not occur together, at the same "time", they are said to be mutually exclusive events. $P(A \cap B) = 0$, where \cap represents "and". The intersection of mutually exclusive events A and B , $A \cap B$ is the null set, denoted \emptyset . $P(\emptyset) = 0$.
- **Conditional probability:** The probability of one event given the knowledge that another event(s) has already occurred is called conditional probability. The conditional probability of "A" given "B" is denoted $P(A|B)$, where $|$ represents "given".
- **Independent:** Two events are independent of each other if one event has occurred (or has not) does not alter the probability the other event will or has

occurred. Mathematically, $P(A|B) = P(A)$ and $P(B|A) = P(B)$ if event A is independent of event B, since if they are independent then event B adds no new knowledge or information about event A and vice versa.

- **Dependent:** Two events are dependent if one event has occurred (or has not) does alter the probability that the other event will or has occurred, $P(A|B) \neq P(A)$.

Important Probability Rules

1. Complement: $P(A^C) = 1 - P(A)$
2. A Or B (general): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. A Or B (mutually exclusive events): $P(A \cup B) = P(A) + P(B)$
4. A and B (general): $P(A \cap B) = P(A)P(B|A)$
5. A and B (independent events): if and only if $P(A \cap B) = P(A)P(B)$ otherwise they are dependent
 - Multiple independent events: $P(A_1 \cap A_2 \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n)$
6. B given A (conditional probability): $P(B|A) = \frac{P(A \cap B)}{P(A)}$
7. Bayes Rule: $P(A|B) = \frac{P(A \cap B)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$

Example: Flip a coin twice. H =heads and T =tails.

- **Sample space:** All possible outcomes: $\mathcal{S} = \{(H, H), (H, T), (T, H), (T, T)\}$.
Note: Each possible outcome is equally likely.

Event: Some possible events:

- A=Exactly 1 Head,
- B=Exactly 1 Tail,
- C=2 Heads,
- D=2 Tails,
- E=Head 1st toss and anything 2nd toss, and
- F=anything 1st and Head 2nd toss.

- **Probability:** Recall 4 possible outcomes and each equally likely.

- Event A consists of $\{H, T\}, \{T, H\} \Rightarrow P(A) = \frac{2}{4} = 0.5$,
- Event B consists of $\{H, T\}, \{T, H\} \Rightarrow P(B) = \frac{2}{4} = 0.5$,
- Event C consists of $\{H, H\} \Rightarrow P(C) = \frac{1}{4} = 0.25$,
- Event D consists of $\{T, T\} \Rightarrow P(D) = \frac{1}{4} = 0.25$,
- Event E consists of $\{H, H\}, \{H, T\} \Rightarrow P(E) = \frac{2}{4} = 0.5$, and
- Event F consists of $\{H, H\}, \{T, H\} \Rightarrow P(F) = \frac{2}{4} = 0.5$.

- **Complement:** The complement of 2 tails, event D , is less than two tails.

$P(D \cup D^C) = 1$. That is, the probability of two tails or less than two tails occurring when there are two tosses equals one.

- **Mutually exclusive:** Two heads and two tails, events C and D. $P(C \cap D) = 0$
- **Conditional probability:** For examples look at the next two items, "Independent" and "Dependent".
- **Independent:** Events E and F are independent. Knowledge about the 1st toss yields no information about the 2nd toss and vice versa. $P(E|F) = P(E)$ and $P(F|E) = P(F)$.
- **Dependent:** Events D and E are dependent. If you know two Tails were tossed, then event E could not have occurred. $P(E|D) = 0$.

2.2. Probability and Cross Classification Tables

Often a cross classification table is used for calculating *empirical probability*. Empirical probability is an estimate of the true probability of an event using observed data. When flipping a coin the probability of getting a head is a half. This is the true probability of getting a head and it is a known fact. The probability of randomly selecting a red jack in a single selection in a standard 52 deck of cards is $\frac{2}{52}$ because there are 2 red jacks within the 52 cards. The actual probability of the latter two examples does not need to be estimated; it can be determined, but in many cases the actual probability must be estimated. For estimating the probability of an event often a frequency table or crosstab (cross classification tables) is used. This technique is used for determining the true probability as well. This is illustrated by considering the probability of selecting a card within a standard deck of cards in Table ???. From Table ??? it is easy calculate the different probabilities related to selecting a single

	Red	Black	Total
Jacks	2	2	4
Not Jacks	24	24	48
Total	26	26	52

Table 2.1: Calculating Probability of Selecting a Card Using a Crosstab.

card with certain attributes. For example, the probability of selecting a:

- jack is $\frac{4}{52}$
- red card is $\frac{26}{52}$
- red jack is $\frac{2}{52}$
- red jack given the card selected is red is $\frac{P(\text{red} \cap \text{jack})}{P(\text{red})} = \frac{2}{52} / \frac{26}{52} = \frac{2}{26}$
- red jack given the card selected is black is $\frac{P(\text{red} \cap \text{jack} \cap \text{black})}{P(\text{black})} = \frac{0}{52} / \frac{26}{52} = \frac{0}{26}$
- red card given the card selected is a jack is $\frac{P(\text{red} \cap \text{jack})}{P(\text{jack})} = \frac{2}{52} / \frac{4}{52} = \frac{2}{4}$
- etc.

In this same manner probability can be estimated. Imagine if you wanted to understand employment and gender within Bangkok. You might randomly sample adults living in Bangkok and ask if they are working. Pretend that Table ?? is the crosstab generated from the data collected. From the survey results in Table ?? you might

	Female	Male	Total
Working	390	410	800
Not working	130	70	200
Total	520	480	1000

Table 2.2: A Crosstab of Gender and Working Status From a Survey.

estimate that the probability of selecting a:

- female is $\frac{520}{1000}$.
- male is $\frac{480}{1000}$.
- working person that is also male is $\frac{410}{1000}$.
- working male given the selected person is a male adult is $\frac{410}{1000} / \frac{480}{1000} = \frac{410}{480}$.
- etc.

2.3. Probability Examples

EXERCISE 2.3.1. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	669	673
Investing in Stocks	292	315

Table 2.3: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an

account with your bank.

- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.3.2. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	668	710
Investing in Stocks	300	326

Table 2.4: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.3.3. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	741	731

Investing in Stocks	313	286
---------------------	-----	-----

Table 2.5: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.3.4. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	683	707
Investing in Stocks	308	268

Table 2.6: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.3.5. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	652	661

Investing in Stocks	264	289
---------------------	-----	-----

Table 2.7: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

2.4. Exercises

2.4.1 Probability Exercises

EXERCISE 2.4.1. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	692	720
Investing in Stocks	289	322

Table 2.8: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.2. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	722	693
Investing in Stocks	308	322

Table 2.9: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.3. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	727	737
Investing in Stocks	322	298

Table 2.10: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.4. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	662	712
Investing in Stocks	304	304

Table 2.11: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.5. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	729	734
Investing in Stocks	333	319

Table 2.12: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.6. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	675	680
Investing in Stocks	358	313

Table 2.13: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.7. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	708	633
Investing in Stocks	313	297

Table 2.14: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.8. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	717	665
Investing in Stocks	300	296

Table 2.15: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.9. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	710	725
Investing in Stocks	298	307

Table 2.16: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

EXERCISE 2.4.10. Assume the survey results were obtained from a survey taken randomly from people inquiring about opening a bank account with your bank. Given the survey information in the table below estimate the following percentages:

	Male	Female
Not Investing in Stocks	699	731
Investing in Stocks	321	292

Table 2.17: The survey breakout for men and women on investing in stocks

- (a) Estimate the percent of inquiries for opening a bank account with your bank that are male.
- (b) Percent of women and investing in stocks of people interested in opening an account with your bank.
- (c) Percent of men and investing in stocks of people interested in opening an account with your bank.
- (d) Given the person is a male and interested in opening an account at your bank estimate the probability he is investing in stocks.

2.4.2 Multiple Choice

Imagine if two six-sided dice were rolled. What is the probability of observing ...?

Click "Begin" and when you are finished "End".

Begin Multiple Choice Questions

1. a 3 and a 4?

- (a) $\frac{6}{36}$ (b) $\frac{1}{36}$ (c) $\frac{3}{36}$ (d) $\frac{2}{36}$

2. a 3 on the 1st roll and then a 4 on the 2nd roll?

- (a) $\frac{6}{36}$ (b) $\frac{1}{36}$ (c) $\frac{3}{36}$ (d) $\frac{2}{36}$

3. a 5 on the 1st roll and then a 3 on the 2nd roll?

- (a) $\frac{6}{36}$ (b) $\frac{1}{36}$ (c) $\frac{3}{36}$ (d) $\frac{2}{36}$

4. Sixes on both rolls?

- (a) $\frac{5}{36}$ (b) $\frac{1}{36}$ (c) $\frac{3}{36}$ (d) $\frac{11}{36}$

5. A 6 on the 1st roll or on the 2nd roll?

- (a) $\frac{5}{36}$ (b) $\frac{1}{36}$ (c) $\frac{3}{36}$ (d) $\frac{11}{36}$

6. A 6 on the 2nd roll given the 1st roll was a 4?

- (a) $\frac{20}{36}$ (b) $\frac{1}{36}$ (c) $\frac{6}{36}$ (d) $\frac{11}{36}$

7. A 6 or a 4 on either the rolls?

- (a) $\frac{20}{36}$ (b) $\frac{1}{36}$ (c) $\frac{3}{36}$ (d) $\frac{11}{36}$

End Multiple Choice Questions

3

Probability Distributions and Random Variables

3.1. Random Variables

Understanding the concept of a random variable is important for a deeper understanding of statistics. Next some key terminology will be covered related to random variables.

Definitions

- **Random variable:** Is an outcome or observation whose value is determined by a process that is not predetermined and thus can't be predicted. Random variables are often denoted using capital letters,

and possible values that a random variable can take by a lower case letter.

1. **Categorical random variable:** Is a random variable that results in categorical response (non-numeric), such as gender (male or female), and opinion (strongly disagree, disagree, ..., or strongly agree).
 - **Dummy coding:** Dummy coding is turning a variable with two or more outcomes into a variable(s) with possible values of 0 and 1. Often categorical variables are dummy coded for analysis purposes. For example, the gender male might be assigned the value of 0 and females the value of 1. If there are several categories, several dummy variables are needed to capture all the information. The dummy coded data can now be treated as a numerical random variable.
2. **Numerical random variable:** Is a random variable that results in a numerical response. Examples include height, weight, age, income, etc. of a randomly selected individual.
 - (a) **Discrete random variable:** Resulting integer values, like the number of heads observed when flipping a coin four times, $x=0,1,2,3$ or 4. For an example, see Table ??.
 - (b) **Continuous random variable:** Resulting in continuous values, like income. For an example see Table ??.
- **Cumulative distribution function (c.d.f.):** Basically $P(X \leq x)$ where X is a random variable and x is a real number. The cdf is often denoted with a

capital F as $F(x)$, i.e. $F(x) = P(X \leq x)$.

• **Probability distribution function (p.d.f.):**

1. For a discrete random variable it is merely the probability of a certain value occurring, $P(X = x)$.

– The probability distribution function has the following properties:

(a) $f(x_i) \geq 0, \quad \forall i.$

(b) $\sum_{\forall i} f(x_i) = 1$

2. For a continuous random variable the $P(X = x) = 0$ and thus the definition is not the same. The p.d.f. for a continuous random variable is a curve described by the function, $f(x)$. The area under the curve within a given interval yields the probability of the continuous random variable falling within that given interval.

– The probability distribution function has the following properties:

(a) $f(x) \geq 0$

(b) $\int_{-\infty}^{\infty} f(x)dx = 1$

(c) $F(b) - F(a) = P(a \leq X \leq b) = \int_a^b f(x)dx$, which is the area under the curve $f(x)$ from a to b , $a \leq b$.

– Note: $P(X = b) = F(b) - F(b) = \int_b^b f(x)dx = 0$, that is the probability of a continuous random variable equaling a specific constant, say b , is zero.

• **Expectation** of a random variable is the mean value (a weighted mean) of the

Discrete	Continuous
0	736.1918273
1	759.5668806
2	812.7593044
3	562.2359305
4	798.2952718

Table 3.1: Example of Discrete and Continuous Data

variable X in the sample space, or population, of possible outcomes. *Expected value* can also be interpreted as the mean value that would be obtained from an infinite number of observations of the random variable.

**Examples of Categorical, Continuous and Discrete
Data.**

- Categorical:
 1. Gender
 2. Blood Type
 3. Marital Status
 4. Eye Color
 5. Political Party
- Discrete:
 1. Number of people using the ATM at a certain location within the past hour.
 2. Number of brothers or sisters a person has.

3. Number of times a person won at roulette within the past 20 spins.

- Continuous:

1. Income

2. Age

3. Height

4. Weight

3.2. General Formulas

This section covers the formulas for calculating the expectation of a random variable, variance of a random variable, and the covariance between two random variables. The expectation of the random variable, if it exists, is the mean for that random variable. The formulas in this section are especially helpful to understand for people interested in courses concerning finance and investing.

General case

$$E(X) = \mu_x$$

$$Var(X) = E[(X - \mu_x)^2] = \sigma_x^2$$

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = \sigma_{xy}$$

Discrete case

$$E(X) = \sum \pi_i x_i = \mu_x \quad (3.1)$$

$$Var(X) = \sum \pi_i [(x_i - \mu_x)^2] = \sum \pi_i x_i^2 - \left(\sum \pi_i x_i \right)^2 = \sigma_x^2$$

$$Cov(X, Y) = \sum \pi_i [(x_i - \mu_x)(y_i - \mu_y)] = \sum \pi_i x_i y_i - \left(\sum \pi_i x_i \sum \pi_i y_i \right) = \sigma_{xy}$$

where π_i is the probability of outcome i and x_i is the outcome. Example i , could refer to recession, stable economy, expanding economy. The π_i could be the given probabilities of each and x_i could be the profit/loss of a mutual fund given the situation. Note: $\sum_i \pi_i = 1$.

Continuous case

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu$$

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \sigma^2$$

$$Cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy = \sigma_{xy}$$

The following subsections will cover only the discrete case for a deeper understanding of the latter formulas presented. For discrete random variables summation is required to solve for expectation, variance and covariance. For continuous random variables integration is required, which is beyond the scope of this text.

3.2.1 Expectation and Mean

Let c be a constant, for example let $c = 5$.

$$E[X] = \sum \pi_i(x_i) = \mu_x$$

$$E[cX] = \sum \pi_i(cx_i) = c \sum \pi_i(x_i) = c\mu_x$$

$$E[5X] = \sum \pi_i(5(x_i)) = 5 \sum \pi_i(x_i) = 5\mu_x$$

$$E[X + c] = \sum \pi_i(x_i + c) = \sum \pi_i(x_i) + \sum \pi_i(c) = \mu_x + c \sum \pi_i = \mu_x + c$$

$$E[X + 5] = \sum \pi_i(x_i + 5) = \sum \pi_i(x_i) + \sum \pi_i(5) = \mu_x + 5 \sum \pi_i = \mu_x + 5$$

If you have a summation of something over i that doesn't change for any of the i , for example a constant c , or 5 it can be moved outside of the summation, as was done above.

3.2.2 Expectation and Variance

Let c be a constant, for example let $c = 5$.

$$\text{Var}(X) = \sum \pi_i [(x_i - \mu_x)^2] = \sigma_x^2$$

$$\text{Var}(cX) = \sum \pi_i [c(x_i) - c\mu_x]^2 = \sum \pi_i [c((x_i) - \mu_x)]^2 = c^2 \sum \pi_i [(x_i) - \mu_x]^2 = c^2 \sigma_x^2$$

$$\begin{aligned} \text{Var}(5X) &= \sum \pi_i [(5(x_i) - 5\mu_x)^2] = \sum \pi_i [5((x_i) - \mu_x)]^2 \\ &= 5^2 \sum \pi_i [(x_i) - \mu_x]^2 = 5^2 \sigma_x^2 = 25\sigma_x^2 \end{aligned}$$

$$\begin{aligned} \text{Var}(X + c) &= \sum \pi_i [(x_i + c) - (\mu_x + c)]^2 = \sum \pi_i [(x_i) + c - \mu_x - c]^2 \\ &= \sum \pi_i [(x_i) - \mu_x]^2 = \sigma_x^2 \end{aligned}$$

$$\begin{aligned} \text{Var}(X + 5) &= \sum \pi_i [(x_i + 5) - (\mu_x + 5)]^2 = \sum \pi_i [(x_i) + 5 - \mu_x - 5]^2 \\ &= \sum \pi_i [(x_i) - \mu_x]^2 = \sigma_x^2 \end{aligned}$$

3.2.3 Expectation and Covariance

Let c_1 be a constant, for example let $c_1 = 5$ and c_2 be another constant, for example let $c_2 = 7$.

$$Cov(X, Y) = \sum \pi_i [(x_i - \mu_x)(y_i - \mu_y)]$$

$$\begin{aligned} Cov(c_1X, c_2Y) &= \sum \pi_i [(c_1x_i - c_1\mu_x)(c_2y_i - c_2\mu_y)] \\ &= c_1c_2 \sum \pi_i [(x_i - \mu_x)(y_i - \mu_y)] = c_1c_2Cov(X, Y) \end{aligned}$$

$$\begin{aligned} Cov(5X, 7Y) &= \sum \pi_i [(5x_i - 5\mu_x)(7y_i - 7\mu_y)] \\ &= 5 * 7 \sum \pi_i [(x_i - \mu_x)(y_i - \mu_y)] = 35Cov(X, Y) \end{aligned}$$

$$\begin{aligned} Cov(X + c_1, Y + c_2) &= \sum \pi_i [((x_i + c_1) - (\mu_x + c_1)) ((y_i + c_2) - (\mu_y + c_2))] \\ &= \sum \pi_i [(x_i - \mu_x)(y_i - \mu_y)] = Cov(X, Y) \end{aligned}$$

$$\begin{aligned} Cov(X + 5, Y + 7) &= \sum \pi_i [((x_i + 5) - (\mu_x + 5)) ((y_i + 7) - (\mu_y + 7))] \\ &= \sum \pi_i [(x_i - \mu_x)(y_i - \mu_y)] = Cov(X, Y) \end{aligned}$$

The addition of constants cancel each other out as in the variance formula.

3.2.4 Expectation and Variance of a Weighted Sum

Let $Z = X + Y$ so:

$$E[Z] = \mu_z$$

by definition, but it can be broken into its components X and Y .

$$E[Z] = E[X + Y] = \sum \pi_i(x_i + y_i) = \sum \pi_i(x_i) + \sum \pi_i(y_i) = \mu_x + \mu_y$$

Recall:

$$(a + b)^2 = a(a + b) + b(a + b) = a^2 + 2ab + b^2$$

We will need this to understand $\text{Var}(c_1X + c_2Y)$

Again let $Z = X + Y$

$$\text{Var}(Z) = E[(Z - \mu_z)^2] = \sigma_z^2$$

but you may only be able to solve for Z from the X and Y .

$$\text{Var}(Z) = E[(Z - \mu_z)^2] = E[((X + Y) - (\mu_x + \mu_y))^2] =$$

Think of $(X + Y)$ as a and $(\mu_x + \mu_y)$ as b .

$$E[(X + Y)^2 - 2(X + Y)(\mu_x + \mu_y) + (\mu_x + \mu_y)^2] =$$

Again using $(a + b)^2 = a^2 + 2ab + b^2$

$$E[X^2 + 2XY + Y^2 - 2X\mu_x - 2X\mu_y - 2Y\mu_x - 2Y\mu_y + \mu_x^2 + 2\mu_x\mu_y + \mu_y^2] =$$

Yes, a big mess but the terms can be regrouped and come up with the following:

$$\begin{aligned} &= E[(X^2 - 2X\mu_x + \mu_x^2) + (Y^2 - 2Y\mu_y + \mu_y^2) + 2(XY - X\mu_y - Y\mu_x + \mu_x\mu_y)] \\ &= E[(X - \mu_x)^2 + (Y - \mu_y)^2 + 2(X - \mu_x)(Y - \mu_y)] \\ &= E[(X - \mu_x)^2] + E[(Y - \mu_y)^2] + 2E[(X - \mu_x)(Y - \mu_y)] \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) = \text{Var}(Z) \end{aligned}$$

From these concepts:

$$\begin{aligned} \text{Var}(c_1X + c_2Y) &= \text{Var}(c_1X) + \text{Var}(c_2Y) + 2\text{Cov}(c_1X, c_2Y) \\ &= c_1^2\text{Var}(X) + c_2^2\text{Var}(Y) + 2c_1c_2\text{Cov}(X, Y) \end{aligned}$$

If we added a constant, c_3 this not change the variance, see next formula:

$$\begin{aligned} Var(c_1X + c_2Y + c_3) &= Var(c_1X) + Var(c_2Y) + 2Cov(c_1X, c_2Y) \\ &= c_1^2Var(X) + c_2^2Var(Y) + 2c_1c_2Cov(X, Y) \end{aligned}$$

Note: $Cov(X, X) = Var(X)$ and correlation, $\rho = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$

3.2.5 Expectation and Variance of an Average of Independently Identically Distributed Random Variables

Let X_1, X_2, \dots, X_n be n independently identically distributed (i.i.d.) random variables with mean of μ and a variance of σ^2 . Let the average of the i.i.d. random variables be denoted

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then the expectation of

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} n\mu = \mu, \quad (3.2)$$

and the variance of

$$Var(\bar{X}) = Var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}. \quad (3.3)$$

Note: Since the random variables are independent the $Cov(X_i, X_j) = 0$ for $i \neq j$.

3.2.6 Useful Mathematical Formulas and Notation

The following are some general formulas that are needed for calculating expectation and variance of discrete and continuous random variables. These formulas are necessary for understanding the mathematics contained within the following sections of this chapter.

- $\binom{n}{x}$ number of ways, or combinations, to choose x different objects from n different objects.
- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
- $a! = a \times (a-1) \times (a-2) \times \cdots \times 1$, where a is an integer ≥ 0 , and $0! = 1$.
- $e = \lim_{n \rightarrow \infty} (1 + \frac{1}{n})^n \rightarrow 2.71828 \dots$. Note: e has special properties.
- natural log= \ln
- $\ln(1) = 0$, $\ln(e) = 1$, $\ln(e^2) = 2$, $\ln(e^c) = c$, where c is some constant.
- $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} e^{2x} = 2e^{2x}$, $\frac{\partial}{\partial x} e^{cx} = ce^{cx}$, where c is some constant.
- Mathematical notation used within some of the exercises and typical spreadsheet programs, such as Microsoft Excel:

1. The notation $x*y$ represents multiply x by y . Example:

$$- 2*3=6$$

2. The notation x^y , represents raise x to the power of y , or x^y . Examples:

$$- 5^2 = 5 * 5 = 25$$

$$- 5^3 = 5 * 5 * 5 = 125$$

$$- 4^3 = 4 * 4 * 4 = 64$$

Examples

- $3! = 3 \times 2 \times 1 = 6$
- $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
- $\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (2 \times 1)} = \frac{120}{6 \times 2} = 10$

The following sections cover various common random variables but not all random variables. For additional information on random variables the author recommends a book by ?.

3.3. Select Discrete Random Variables

3.3.1 Binomial Distribution

Binomial Distribution has the following properties:

1. There are a fixed number of trials or observations, n , determined in advance.
2. Each trial can take on one of two possible outcomes, labeled "success" and "failure".
3. Each trial's outcome is determined independently of all the other trials.

4. The probability of a success and that of a failure remains the same from one trial to the next, and is denoted by π and $1 - \pi$, respectively.

Binomial Distribution:

$$P(X = x) = f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

- X is Binomially distributed
- x the number of successes, where $x = 0, 1, \dots, n$
- n the number of trials
- π the probability of success
- $1 - \pi$ the probability of failure
- $\mu = E(X) = n\pi$
- $\sigma^2 = V(X) = n\pi(1 - \pi)$

Examples

- The number of heads out of 2 coin tosses. The p.d.f. of this example can be seen in Figure ??
- Assume that the probability of SET, Securities Exchange of Thailand, will end the day on the positive side is constant and each day's outcome, positive or not, is independent of what occurred on prior days. The number of times the SET increases in 2 weeks, 10 working days.
- The number of women in a group of 20 randomly selected people.

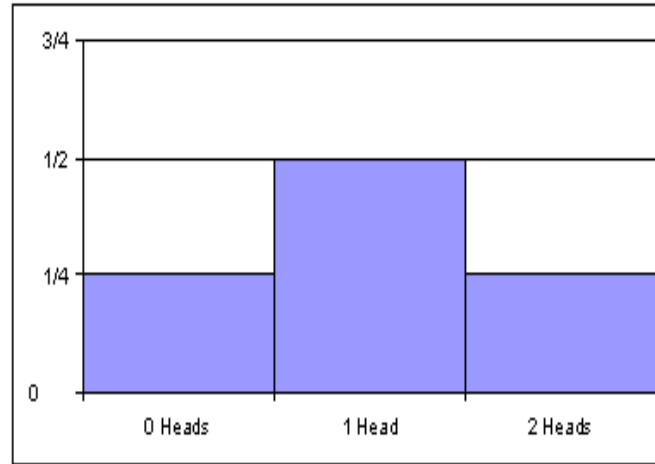


Figure 3.1: The p.d.f. of X =the number of Heads in two tosses of a coin. This follows a Binomial distribution, with $n = 2$, and $p = .5$.

3.3.2 Hypergeometric Distribution

Hypergeometric distribution has the following properties:

1. When units are selected from a finite population without replacement and the population consists of successes and failures.
2. The major difference between the Hypergeometric distribution and the Binomial distribution is that the probability of selecting a success is **not constant** and is **not independent** from each draw.

Hypergeometric Distribution

$$P(X = x) = f(x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}$$

- X has a Hypergeometric distribution.
- x is the number of successes in the sample
- n is the sample size
- A is the number of successes in the population
- N is the population size
- $N - A$ is the number of failures in the population
- $n - x$ is the number of failures in the sample
- $\mu = E(X) = \frac{nA}{N}$
- $\sigma^2 = V(X) = \left(\frac{N-n}{N-1}\right) \frac{nA(N-A)}{N^2}$
- $\frac{N-n}{N-1}$ is called the finite population correction factor.

Examples

- The number of spades selected when 5 cards are drawn from a standard 52 deck of cards.
- There are 20 Sony CD players in stock at the Sony store at the Mall Bangkapi; 5 are defective. A customer buys 6 of the 20 CD players. The number of defective CD players bought of the 6 CD players.
- The computer lab has 20 computers and 15 of the 20 computers have illegal software on them. The number of computers selected with illegal software from a sample of size 5.

3.3.3 Poisson Distribution

A Poisson process has the following properties:

- Within a given continuous interval of time (or area, surface, etc.) it is possible to observe discrete events.
- It is possible to partition the interval into subintervals of small enough length such that
 1. the probability of more than one success within a single subinterval is zero
 2. the probability of a success in a subinterval is constant for all subintervals and the probability is proportional to the length of the subinterval
 3. the occurrence of a success within a subinterval is independent of what occurs in any of the other subintervals

Poisson Distribution:

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

- X is Poisson Distributed
- x equals the number of success in the interval
- $x = 0, 1, 2, \dots$
- $0 < \lambda$

- $\mu = E(X) = \lambda$
- $\sigma^2 = V(X) = \lambda$

Examples

- The number of car accidents at NIDA in a month
- The number of paint imperfections on a Toyota car produced
- The number of phone calls in a day

3.4. Select Continuous Random Variables

3.4.1 Exponential Distribution

Exponential Distribution has the following properties:

1. Equals the distance between successive occurrences or arrivals of a Poisson process with mean $\lambda > 0$
2. λ is the average number of occurrences or arrivals per unit of time (length, space, etc.)
3. $\frac{1}{\lambda}$ is the average time between occurrences or arrivals.

Exponential Distribution:

$$f(x) = \lambda e^{-\lambda x}$$

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$$

- $0 \leq x < \infty$
- $\mu = E(X) = \frac{1}{\lambda}$
- $\sigma^2 = V(X) = \frac{1}{\lambda^2}$

Examples

- The amount of time until the next customer at The Pizza Customer will arrive.
- The amount of time until the DVD player will break. The exponential distribution is very useful for determining length of warranty.
- The amount of time until the next person will arrive at a specific ATM

3.4.2 Normal Distribution

Normal Distribution has the following properties:

- Symmetrical and a bell shaped appearance.
- The population mean and median are equal.
- An infinite range, $-\infty < x < \infty$
- The approximate probability for certain ranges of X -values:

1. $P(\mu - 1\sigma < X < \mu + 1\sigma) \approx 68\%$
2. $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%$
3. $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 99.7\%$

Normal Distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $-\infty < x < \infty$
- $N(\mu, \sigma^2)$ is used to denote the distribution
- $E(X) = \mu$
- $V(X) = \sigma^2$
- If $\mu = 0$ and $\sigma^2 = 1$, this is called a Standard Normal and denoted Z
 - All Normally distributed random variables can be converted into a Standard Normal
- To determine probabilities related to the Normal distribution the Standard Normal distribution is used

3.4.2.1 Standard Normal Distribution:

The following formula is used to transform a Normally distributed random variable, X , into a Standard Normally distributed random variable,

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

Note: if z is known we can solve for x :

$$x = \mu + z\sigma$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- $-\infty < z < \infty$
- Has the same properties as the Normal distribution and
 - $N(0, 1)$ is used to denote the distribution
 - $E(Z) = \mu = 0$
 - $V(Z) = \sigma^2 = 1$
- The approximate probability for certain ranges of Z -values:

1. $P(-1 < Z < 1) \approx 68\%$
2. $P(-2 < Z < 2) \approx 95\%$
3. $P(-3 < Z < 3) \approx 99.7\%$

- To find probabilities of any z -value refer to the Table ?? or computer software such as Microsoft Excel may be used

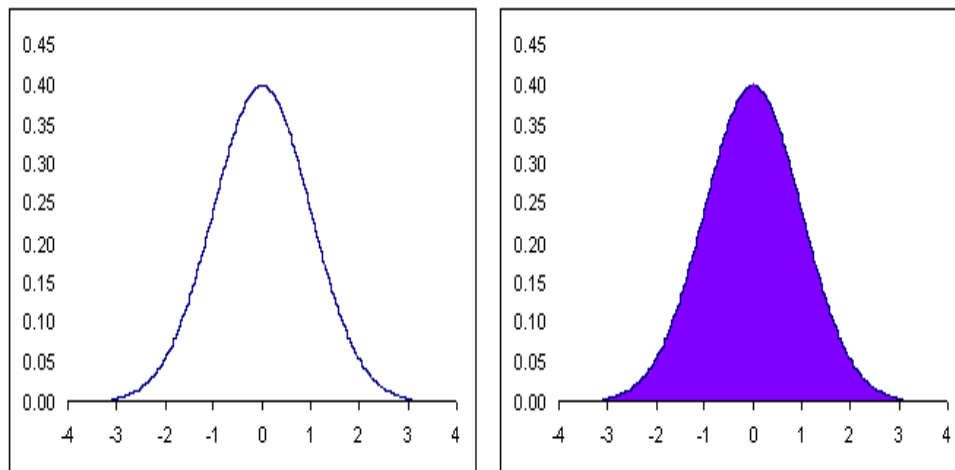


Figure 3.2: The Normal curve, $f(z)$, and again with the area underneath shaded, representing $P(-\infty < Z < \infty)$.

1. $P(Z = c) = 0$, where c is a constant.
2. $P(Z < -c) = P(Z > c)$
3. $P(Z > c) = 1 - P(Z < c)$
4. $P(Z < -c) = 1 - P(Z < c)$

The Figures ??,??,??,??,?? illustrate the important probabilities associated with the normal distribution and calculating approximate probabilities using the information on the Normal distribution given in this subsection.

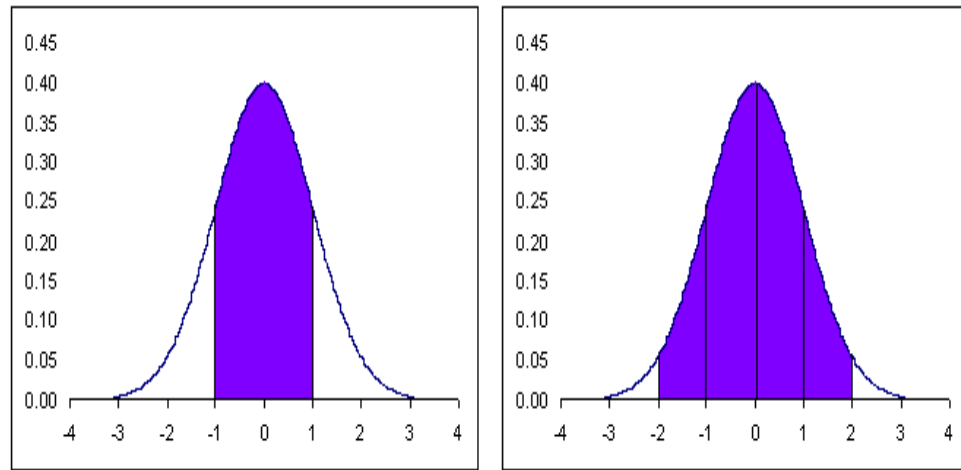


Figure 3.3: The area underneath the normal curve shaded from -1 to 1, and -2 to 2, illustrating $P(-1 < Z < 1) \approx 68\%$ and $P(-2 < Z < 2) \approx 95\%$ respectively.

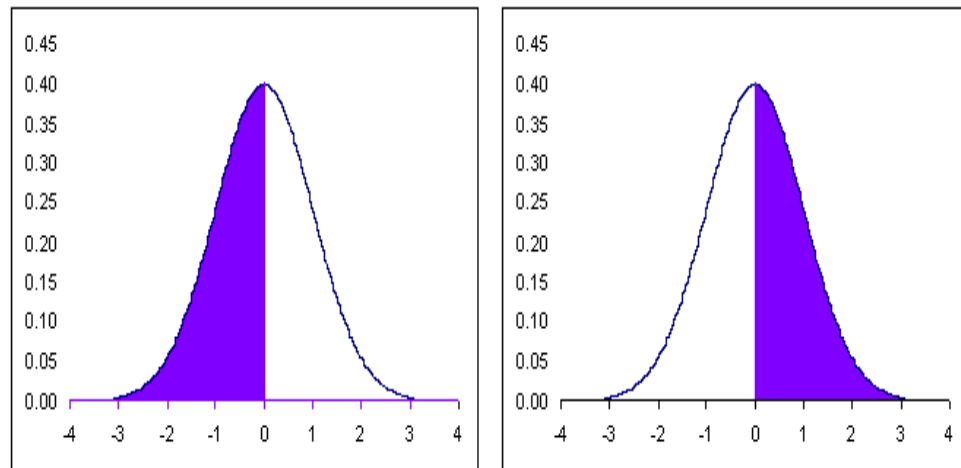


Figure 3.4: The area underneath the normal curve shaded from $-\infty$ to 0, and 0 to ∞ , illustrating $P(-\infty < Z < 0) = 50\%$ and $P(0 < Z < \infty) = 50\%$ respectively.

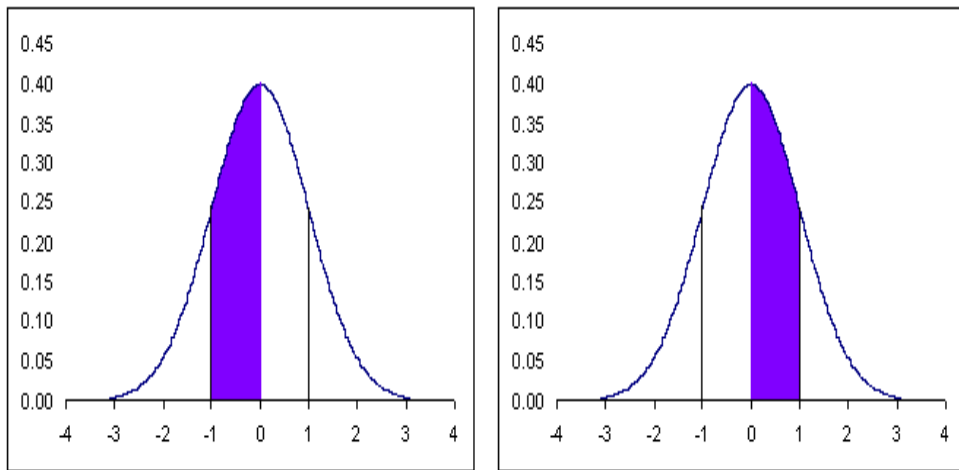


Figure 3.5: The area underneath the normal curve shaded from -1 to 0, and 0 to 1, illustrating $P(-1 < Z < 0) \approx 34\%$ and $P(0 < Z < 1) \approx 34\%$ respectively. Resulting from symmetry.

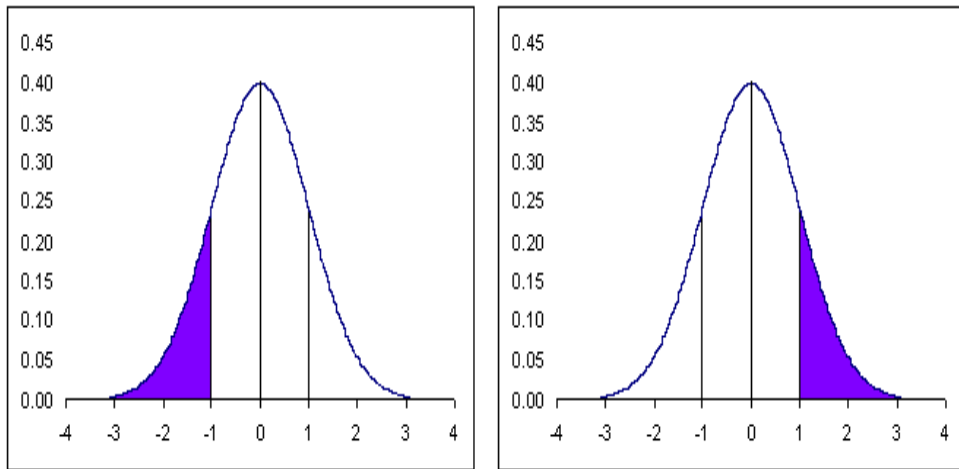


Figure 3.6: The area underneath the normal curve shaded from $-\infty$ to -1, and 1 to ∞ , illustrating $P(-\infty < Z < -1) \approx 16\%$ and $P(1 < Z < \infty) \approx 16\%$ respectively. Resulting from symmetry.

3.4.2.2 The Distribution of the Average of I.I.D. Normally Distributed Random Variables

Let X_1, X_2, \dots, X_n be n i.i.d. $N(\mu, \sigma^2)$ random variables. The expectation of the sample mean, \bar{X} , of n i.i.d. Normally distributed random variables is μ and the variance is $\frac{\sigma^2}{n}$. Recall equations ?? and ??. The situation where the random variables are Normally distributed is a special case in that \bar{X} is also Normally distributed and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In addition,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

where $Z \sim N(0, 1)$.

Examples

- The height of a randomly selected person is often assumed to be Normally distributed
- The weight of a randomly selected person is often assumed to be Normally distributed
- The pulse rate of a randomly selected person is often assumed to be Normally distributed

3.5. Examples

3.5.1 Expectation and Covariance of Random Variables: Examples

EXERCISE 3.5.1.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.21	20.35	18.15
2	0.30	16.35	16.65
3	0.18	19.61	18.04
4	0.31	17.40	22.82

Table 3.2: The data

EXERCISE 3.5.2.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.12	16.50	18.88
2	0.29	17.51	22.45
3	0.13	22.47	16.38
4	0.46	17.66	20.71

Table 3.3: The data

EXERCISE 3.5.3.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.24	13.78	19.79
2	0.23	22.66	20.69
3	0.11	22.80	15.54
4	0.42	18.24	20.69

Table 3.4: The data

EXERCISE 3.5.4.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.10	16.78	18.43
2	0.27	18.45	21.82
3	0.19	17.49	18.12
4	0.44	20.71	16.47

Table 3.5: The data

EXERCISE 3.5.5.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.15	17.24	21.24
2	0.11	20.66	20.21
3	0.12	20.00	17.80
4	0.62	20.56	20.72

Table 3.6: The data

3.5.2 Binomial Distribution Examples

EXERCISE 3.5.6. Assume the random variable(s) is from a binomial distribution with $n = 11$ and $\pi = 0.9$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 9$:
- (b) What is the probability $X > 9$:
- (c) What is the probability $X \geq 9$:
- (d) What is the probability $9 < X \leq 11$:

EXERCISE 3.5.7. Assume the random variable(s) is from a binomial distribution with $n = 8$ and $\pi = 0.7$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 5$:
- (b) What is the probability $X > 5$:
- (c) What is the probability $X \geq 5$:
- (d) What is the probability $5 < X \leq 6$:

EXERCISE 3.5.8. Assume the random variable(s) is from a binomial distribution with $n = 3$ and $\pi = 0.3$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 3$:

EXERCISE 3.5.9. Assume the random variable(s) is from a binomial distribution with $n = 7$ and $\pi = 0.7$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 5$:
- (b) What is the probability $X > 5$:
- (c) What is the probability $X \geq 5$:
- (d) What is the probability $5 < X \leq 6$:

EXERCISE 3.5.10. Assume the random variable(s) is from a binomial distribution with $n = 2$ and $\pi = 0.4$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 2$:

3.5.3 Hypergeometric Distribution Examples

EXERCISE 3.5.11. Assume the random variable(s) is from a hypergeometric distribution with $A = 10$ number of success and population size $N = 16$, and X is the number of successes you select. You select from the population $n = 15$ items without replacement. Answer the following:

(a) What is the probability $X \leq 9$:

(b) What is the probability $X > 9$:

(c) What is the probability $X \geq 9$:

(d) What is the probability $9 < X \leq 10$:

EXERCISE 3.5.12. Assume the random variable(s) is from a hypergeometric distribution with $A = 7$ number of success and population size $N = 13$, and X is the number of successes you select. You select from the population $n = 6$ items without replacement. Answer the following:

(a) What is the probability $X \leq 2$:

(b) What is the probability $X > 2$:

(c) What is the probability $X \geq 2$:

(d) What is the probability $2 < X \leq 4$:

EXERCISE 3.5.13. Assume the random variable(s) is from a hypergeometric distribution with $A = 7$ number of success and population size $N = 15$, and X is the number of successes you select. You select from the population $n = 10$ items without replacement. Answer the following:

(a) What is the probability $X \leq 5$:

(b) What is the probability $X > 5$:

(c) What is the probability $X \geq 5$:

(d) What is the probability $5 < X \leq 6$:

EXERCISE 3.5.14. Assume the random variable(s) is from a hypergeometric distribution with $A = 9$ number of success and population size $N = 17$, and X is the number of successes you select. You select from the population $n = 1$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 2$:

EXERCISE 3.5.15. Assume the random variable(s) is from a hypergeometric distribution with $A = 6$ number of success and population size $N = 13$, and X is the number of successes you select. You select from the population $n = 4$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 3$:

3.5.4 Poisson Distribution Examples

EXERCISE 3.5.16. Assume the random variable(s) is from a poisson distribution with $\lambda = 5.9$. Answer the following:

- (a) What is the probability $X \leq 5$:
- (b) What is the probability $X > 5$:
- (c) What is the probability $X \geq 5$:
- (d) What is the probability $5 < X \leq 7$:

EXERCISE 3.5.17. Assume the random variable(s) is from a poisson distribution with $\lambda = 1.9$. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 3$:

EXERCISE 3.5.18. Assume the random variable(s) is from a poisson distribution with $\lambda = 6$. Answer the following:

- (a) What is the probability $X \leq 6$:
- (b) What is the probability $X > 6$:
- (c) What is the probability $X \geq 6$:
- (d) What is the probability $6 < X \leq 9$:

EXERCISE 3.5.19. Assume the random variable(s) is from a poisson distribution with $\lambda = 5$. Answer the following:

- (a) What is the probability $X \leq 6$:
- (b) What is the probability $X > 6$:
- (c) What is the probability $X \geq 6$:
- (d) What is the probability $6 < X \leq 8$:

EXERCISE 3.5.20. Assume the random variable(s) is from a poisson distribution with $\lambda = 5.4$. Answer the following:

- (a) What is the probability $X \leq 3$:
- (b) What is the probability $X > 3$:
- (c) What is the probability $X \geq 3$:
- (d) What is the probability $3 < X \leq 4$:

3.5.5 Exponential Distribution Examples

EXERCISE 3.5.21. Assume the random variable(s) is from an exponential distribution with $\lambda = 8.3$. Answer the following:

- (a) What is the probability $X \leq 0$:
- (b) What is the probability $X \geq 0$:

(c) What is the probability $0 \leq X \leq 0.08$:

EXERCISE 3.5.22. Assume the random variable(s) is from an exponential distribution with $\lambda = 0.2$. Answer the following:

(a) What is the probability $X \leq 0.97$:

(b) What is the probability $X \geq 0.97$:

(c) What is the probability $0.97 \leq X \leq 11.5$:

EXERCISE 3.5.23. Assume the random variable(s) is from an exponential distribution with $\lambda = 3.6$. Answer the following:

(a) What is the probability $X \leq 0.12$:

(b) What is the probability $X \geq 0.12$:

(c) What is the probability $0.12 \leq X \leq 0.73$:

EXERCISE 3.5.24. Assume the random variable(s) is from an exponential distribution with $\lambda = 7.3$. Answer the following:

- (a) What is the probability $X \leq 0.03$:
- (b) What is the probability $X \geq 0.03$:
- (c) What is the probability $0.03 \leq X \leq 0.07$:

EXERCISE 3.5.25. Assume the random variable(s) is from an exponential distribution with $\lambda = 9.9$. Answer the following:

- (a) What is the probability $X \leq 0.13$:
- (b) What is the probability $X \geq 0.13$:
- (c) What is the probability $0.13 \leq X \leq 0.19$:

3.5.6 Normal Distribution Examples

EXERCISE 3.5.26. Assume the random variable(s) is from a normal distribution with $\mu = 6$ and $\sigma = 0.4$. Answer the following:

(a) What is the probability $X \leq 6.29$:

(b) What is the probability $X \geq 6.29$:

(c) What is the probability $6.29 \leq X \leq 6.49$:

EXERCISE 3.5.27. Assume the random variable(s) is from a normal distribution with $\mu = 1.8$ and $\sigma = 1.2$. Answer the following:

(a) What is the probability $X \leq 0.97$:

(b) What is the probability $X \geq 0.97$:

(c) What is the probability $0.97 \leq X \leq 1.95$:

EXERCISE 3.5.28. Assume the random variable(s) is from a normal distribution with $\mu = 6.7$ and $\sigma = 2.7$. Answer the following:

(a) What is the probability $X \leq 3.55$:

(b) What is the probability $X \geq 3.55$:

(c) What is the probability $3.55 \leq X \leq 6.58$:

EXERCISE 3.5.29. Assume the random variable(s) is from a normal distribution with $\mu = 8.5$ and $\sigma = 2.7$. Answer the following:

(a) What is the probability $X \leq 7.59$:

(b) What is the probability $X \geq 7.59$:

(c) What is the probability $7.59 \leq X \leq 9.88$:

EXERCISE 3.5.30. Assume the random variable(s) is from a normal distribution with $\mu = 7.7$ and $\sigma = 0.9$. Answer the following:

(a) What is the probability $X \leq 7.11$:

(b) What is the probability $X \geq 7.11$:

(c) What is the probability $7.11 \leq X \leq 7.39$:

3.6. Exercises

3.6.1 Expectation and Covariance of Random Variables: Examples

EXERCISE 3.6.1.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.29	16.55	19.48
2	0.16	20.93	16.70
3	0.20	17.66	11.21
4	0.35	19.57	21.48

Table 3.7: The data

EXERCISE 3.6.2.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.11	18.59	19.72
2	0.22	22.70	21.86
3	0.28	19.88	17.81
4	0.39	16.51	23.26

 Table 3.8: The data
EXERCISE 3.6.3.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.11	19.21	16.26
2	0.18	17.90	22.83
3	0.25	23.73	17.72
4	0.46	22.40	15.84

 Table 3.9: The data
EXERCISE 3.6.4.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.17	23.38	20.50
2	0.21	19.70	18.73
3	0.11	20.72	16.88
4	0.51	20.10	16.14

 Table 3.10: The data

EXERCISE 3.6.5.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.15	24.65	19.34
2	0.16	21.32	19.84
3	0.14	20.56	19.08
4	0.55	23.10	18.13

Table 3.11: The data

EXERCISE 3.6.6.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.19	19.00	24.40
2	0.22	20.82	13.17
3	0.28	18.70	17.23
4	0.31	19.17	19.56

Table 3.12: The data

EXERCISE 3.6.7.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.25	15.12	19.92
2	0.14	13.43	19.25
3	0.18	19.65	20.85
4	0.43	18.79	20.72

Table 3.13: The data

EXERCISE 3.6.8.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.26	25.49	19.43
2	0.20	17.22	23.89
3	0.27	21.69	20.24
4	0.27	20.56	18.55

Table 3.14: The data

EXERCISE 3.6.9.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.16	19.08	22.59
2	0.12	17.63	19.05
3	0.12	18.78	20.50
4	0.60	17.31	16.35

Table 3.15: The data

EXERCISE 3.6.10.

Using the data below calculate the expectation of X , Y , $Var(X)$, $Var(Y)$, $Covar(X, Y)$.

	π_i	x_i	y_i
1	0.27	9.66	21.11
2	0.27	24.47	21.03
3	0.25	19.29	22.04
4	0.21	22.83	23.39

Table 3.16: The data

3.6.2 Binomial Distribution Examples

EXERCISE 3.6.11. Assume the random variable(s) is from a binomial distribution with $n = 7$ and $\pi = 0.3$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 3$:

EXERCISE 3.6.12. Assume the random variable(s) is from a binomial distribution with $n = 2$ and $\pi = 0.6$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 2$:

EXERCISE 3.6.13. Assume the random variable(s) is from a binomial distribution with $n = 2$ and $\pi = 0.2$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 0$:
- (b) What is the probability $X > 0$:
- (c) What is the probability $X \geq 0$:
- (d) What is the probability $0 < X \leq 1$:

EXERCISE 3.6.14. Assume the random variable(s) is from a binomial distribution with $n = 11$ and $\pi = 0.4$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 3$:
- (b) What is the probability $X > 3$:
- (c) What is the probability $X \geq 3$:
- (d) What is the probability $3 < X \leq 6$:

EXERCISE 3.6.15. Assume the random variable(s) is from a binomial distribution with $n = 10$ and $\pi = 0.2$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 0$:
- (b) What is the probability $X > 0$:
- (c) What is the probability $X \geq 0$:
- (d) What is the probability $0 < X \leq 2$:

EXERCISE 3.6.16. Assume the random variable(s) is from a binomial distribution with $n = 2$ and $\pi = 0.8$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 2$:

EXERCISE 3.6.17. Assume the random variable(s) is from a binomial distribution with $n = 12$ and $\pi = 0.6$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 8$:
- (b) What is the probability $X > 8$:
- (c) What is the probability $X \geq 8$:
- (d) What is the probability $8 < X \leq 9$:

EXERCISE 3.6.18. Assume the random variable(s) is from a binomial distribution with $n = 7$ and $\pi = 0.6$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 4$:
- (b) What is the probability $X > 4$:
- (c) What is the probability $X \geq 4$:
- (d) What is the probability $4 < X \leq 6$:

EXERCISE 3.6.19. Assume the random variable(s) is from a binomial distribution with $n = 8$ and $\pi = 0.9$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 6$:
- (b) What is the probability $X > 6$:
- (c) What is the probability $X \geq 6$:
- (d) What is the probability $6 < X \leq 8$:

EXERCISE 3.6.20. Assume the random variable(s) is from a binomial distribution with $n = 7$ and $\pi = 0.6$, and X is the number of successes. Answer the following:

- (a) What is the probability $X \leq 4$:
- (b) What is the probability $X > 4$:
- (c) What is the probability $X \geq 4$:
- (d) What is the probability $4 < X \leq 5$:

3.6.3 Hypergeometric Distribution Examples

EXERCISE 3.6.21. Assume the random variable(s) is from a hypergeometric distribution with $A = 7$ number of success and population size $N = 13$, and X is the number of successes you select. You select from the population $n = 4$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 2$:
- (b) What is the probability $X > 2$:
- (c) What is the probability $X \geq 2$:
- (d) What is the probability $2 < X \leq 3$:

EXERCISE 3.6.22. Assume the random variable(s) is from a hypergeometric distribution with $A = 6$ number of success and population size $N = 12$, and X is the number of successes you select. You select from the population $n = 12$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 6$:
- (b) What is the probability $X > 6$:

(c) What is the probability $X \geq 6$:

(d) What is the probability $6 < X \leq 7$:

EXERCISE 3.6.23. Assume the random variable(s) is from a hypergeometric distribution with $A = 10$ number of success and population size $N = 16$, and X is the number of successes you select. You select from the population $n = 4$ items without replacement. Answer the following:

(a) What is the probability $X \leq 1$:

(b) What is the probability $X > 1$:

(c) What is the probability $X \geq 1$:

(d) What is the probability $1 < X \leq 3$:

EXERCISE 3.6.24. Assume the random variable(s) is from a hypergeometric distribution with $A = 8$ number of success and population size $N = 13$, and X is the number of successes you select. You select from the population $n = 2$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 0$:
- (b) What is the probability $X > 0$:
- (c) What is the probability $X \geq 0$:
- (d) What is the probability $0 < X \leq 1$:

EXERCISE 3.6.25. Assume the random variable(s) is from a hypergeometric distribution with $A = 6$ number of success and population size $N = 16$, and X is the number of successes you select. You select from the population $n = 2$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 0$:
- (b) What is the probability $X > 0$:
- (c) What is the probability $X \geq 0$:
- (d) What is the probability $0 < X \leq 1$:

EXERCISE 3.6.26. Assume the random variable(s) is from a hypergeometric distribution with $A = 5$ number of success and population size $N = 12$, and X is the number of successes you select. You select from the population $n = 6$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 3$:
- (b) What is the probability $X > 3$:
- (c) What is the probability $X \geq 3$:
- (d) What is the probability $3 < X \leq 4$:

EXERCISE 3.6.27. Assume the random variable(s) is from a hypergeometric distribution with $A = 7$ number of success and population size $N = 17$, and X is the number of successes you select. You select from the population $n = 17$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 7$:
- (b) What is the probability $X > 7$:
- (c) What is the probability $X \geq 7$:

(d) What is the probability $7 < X \leq 8$:

EXERCISE 3.6.28. Assume the random variable(s) is from a hypergeometric distribution with $A = 9$ number of success and population size $N = 17$, and X is the number of successes you select. You select from the population $n = 14$ items without replacement. Answer the following:

(a) What is the probability $X \leq 7$:

(b) What is the probability $X > 7$:

(c) What is the probability $X \geq 7$:

(d) What is the probability $7 < X \leq 8$:

EXERCISE 3.6.29. Assume the random variable(s) is from a hypergeometric distribution with $A = 6$ number of success and population size $N = 14$, and X is the number of successes you select. You select from the population $n = 11$ items without replacement. Answer the following:

(a) What is the probability $X \leq 3$:

- (b) What is the probability $X > 3$:
- (c) What is the probability $X \geq 3$:
- (d) What is the probability $3 < X \leq 6$:

EXERCISE 3.6.30. Assume the random variable(s) is from a hypergeometric distribution with $A = 10$ number of success and population size $N = 18$, and X is the number of successes you select. You select from the population $n = 1$ items without replacement. Answer the following:

- (a) What is the probability $X \leq 0$:
- (b) What is the probability $X > 0$:
- (c) What is the probability $X \geq 0$:
- (d) What is the probability $0 < X \leq 1$:

3.6.4 Poisson Distribution Examples

EXERCISE 3.6.31. Assume the random variable(s) is from a poisson distribution with $\lambda = 9.2$. Answer the following:

- (a) What is the probability $X \leq 7$:
- (b) What is the probability $X > 7$:
- (c) What is the probability $X \geq 7$:
- (d) What is the probability $7 < X \leq 14$:

EXERCISE 3.6.32. Assume the random variable(s) is from a poisson distribution with $\lambda = 9.6$. Answer the following:

- (a) What is the probability $X \leq 3$:
- (b) What is the probability $X > 3$:
- (c) What is the probability $X \geq 3$:
- (d) What is the probability $3 < X \leq 12$:

EXERCISE 3.6.33. Assume the random variable(s) is from a poisson distribution with $\lambda = 9.8$. Answer the following:

- (a) What is the probability $X \leq 9$:
- (b) What is the probability $X > 9$:
- (c) What is the probability $X \geq 9$:
- (d) What is the probability $9 < X \leq 10$:

EXERCISE 3.6.34. Assume the random variable(s) is from a poisson distribution with $\lambda = 1.5$. Answer the following:

- (a) What is the probability $X \leq 1$:
- (b) What is the probability $X > 1$:
- (c) What is the probability $X \geq 1$:
- (d) What is the probability $1 < X \leq 2$:

EXERCISE 3.6.35. Assume the random variable(s) is from a poisson distribution with $\lambda = 5.1$. Answer the following:

- (a) What is the probability $X \leq 2$:
- (b) What is the probability $X > 2$:
- (c) What is the probability $X \geq 2$:
- (d) What is the probability $2 < X \leq 3$:

EXERCISE 3.6.36. Assume the random variable(s) is from a poisson distribution with $\lambda = 6.1$. Answer the following:

- (a) What is the probability $X \leq 6$:
- (b) What is the probability $X > 6$:
- (c) What is the probability $X \geq 6$:
- (d) What is the probability $6 < X \leq 9$:

EXERCISE 3.6.37. Assume the random variable(s) is from a poisson distribution with $\lambda = 5.8$. Answer the following:

- (a) What is the probability $X \leq 3$:
- (b) What is the probability $X > 3$:
- (c) What is the probability $X \geq 3$:
- (d) What is the probability $3 < X \leq 5$:

EXERCISE 3.6.38. Assume the random variable(s) is from a poisson distribution with $\lambda = 2.1$. Answer the following:

- (a) What is the probability $X \leq 2$:
- (b) What is the probability $X > 2$:
- (c) What is the probability $X \geq 2$:
- (d) What is the probability $2 < X \leq 3$:

EXERCISE 3.6.39. Assume the random variable(s) is from a poisson distribution with $\lambda = 2.9$. Answer the following:

- (a) What is the probability $X \leq 2$:
- (b) What is the probability $X > 2$:
- (c) What is the probability $X \geq 2$:
- (d) What is the probability $2 < X \leq 4$:

EXERCISE 3.6.40. Assume the random variable(s) is from a poisson distribution with $\lambda = 8.9$. Answer the following:

- (a) What is the probability $X \leq 5$:
- (b) What is the probability $X > 5$:
- (c) What is the probability $X \geq 5$:
- (d) What is the probability $5 < X \leq 9$:

3.6.5 Exponential Distribution Examples

EXERCISE 3.6.41. Assume the random variable(s) is from an exponential distribution with $\lambda = 3.3$. Answer the following:

- (a) What is the probability $X \leq 0.03$:
- (b) What is the probability $X \geq 0.03$:
- (c) What is the probability $0.03 \leq X \leq 0.07$:

EXERCISE 3.6.42. Assume the random variable(s) is from an exponential distribution with $\lambda = 1.3$. Answer the following:

- (a) What is the probability $X \leq 0.38$:
- (b) What is the probability $X \geq 0.38$:
- (c) What is the probability $0.38 \leq X \leq 0.78$:

EXERCISE 3.6.43. Assume the random variable(s) is from an exponential distribution with $\lambda = 7.7$. Answer the following:

- (a) What is the probability $X \leq 0.12$:
- (b) What is the probability $X \geq 0.12$:
- (c) What is the probability $0.12 \leq X \leq 0.44$:

EXERCISE 3.6.44. Assume the random variable(s) is from an exponential distribution with $\lambda = 7.2$. Answer the following:

- (a) What is the probability $X \leq 0.01$:
- (b) What is the probability $X \geq 0.01$:
- (c) What is the probability $0.01 \leq X \leq 0.06$:

EXERCISE 3.6.45. Assume the random variable(s) is from an exponential distribution with $\lambda = 4.5$. Answer the following:

- (a) What is the probability $X \leq 0.03$:
- (b) What is the probability $X \geq 0.03$:

(c) What is the probability $0.03 \leq X \leq 0.23$:

EXERCISE 3.6.46. Assume the random variable(s) is from an exponential distribution with $\lambda = 5.8$. Answer the following:

(a) What is the probability $X \leq 0.07$:

(b) What is the probability $X \geq 0.07$:

(c) What is the probability $0.07 \leq X \leq 0.09$:

EXERCISE 3.6.47. Assume the random variable(s) is from an exponential distribution with $\lambda = 1.6$. Answer the following:

(a) What is the probability $X \leq 0.03$:

(b) What is the probability $X \geq 0.03$:

(c) What is the probability $0.03 \leq X \leq 0.46$:

EXERCISE 3.6.48. Assume the random variable(s) is from an exponential distribu-

tion with $\lambda = 6$. Answer the following:

- (a) What is the probability $X \leq 0.02$:
- (b) What is the probability $X \geq 0.02$:
- (c) What is the probability $0.02 \leq X \leq 0.05$:

EXERCISE 3.6.49. Assume the random variable(s) is from an exponential distribution with $\lambda = 5.6$. Answer the following:

- (a) What is the probability $X \leq 0.12$:
- (b) What is the probability $X \geq 0.12$:
- (c) What is the probability $0.12 \leq X \leq 0.23$:

EXERCISE 3.6.50. Assume the random variable(s) is from an exponential distribution with $\lambda = 6$. Answer the following:

- (a) What is the probability $X \leq 0.03$:

(b) What is the probability $X \geq 0.03$:

(c) What is the probability $0.03 \leq X \leq 0.08$:

3.6.6 Normal Distribution Examples

EXERCISE 3.6.51. Assume the random variable(s) is from a normal distribution with $\mu = 5.5$ and $\sigma = 0.5$. Answer the following:

(a) What is the probability $X \leq 5.82$:

(b) What is the probability $X \geq 5.82$:

(c) What is the probability $5.82 \leq X \leq 5.97$:

EXERCISE 3.6.52. Assume the random variable(s) is from a normal distribution with $\mu = 0.8$ and $\sigma = 0.8$. Answer the following:

(a) What is the probability $X \leq 0.03$:

(b) What is the probability $X \geq 0.03$:

(c) What is the probability $0.03 \leq X \leq 0.32$:

EXERCISE 3.6.53. Assume the random variable(s) is from a normal distribution with $\mu = 4.4$ and $\sigma = 0.5$. Answer the following:

(a) What is the probability $X \leq 5$:

(b) What is the probability $X \geq 5$:

(c) What is the probability $5 \leq X \leq 5.42$:

EXERCISE 3.6.54. Assume the random variable(s) is from a normal distribution with $\mu = 0.1$ and $\sigma = 0.5$. Answer the following:

(a) What is the probability $X \leq 0.1$:

(b) What is the probability $X \geq 0.1$:

(c) What is the probability $0.1 \leq X \leq 0.24$:

EXERCISE 3.6.55. Assume the random variable(s) is from a normal distribution with $\mu = 5.9$ and $\sigma = 1.2$. Answer the following:

- (a) What is the probability $X \leq 6.15$:
- (b) What is the probability $X \geq 6.15$:
- (c) What is the probability $6.15 \leq X \leq 7.19$:

EXERCISE 3.6.56. Assume the random variable(s) is from a normal distribution with $\mu = 4.7$ and $\sigma = 0.5$. Answer the following:

- (a) What is the probability $X \leq 4.81$:
- (b) What is the probability $X \geq 4.81$:
- (c) What is the probability $4.81 \leq X \leq 5.43$:

EXERCISE 3.6.57. Assume the random variable(s) is from a normal distribution with $\mu = 0.9$ and $\sigma = 0.3$. Answer the following:

- (a) What is the probability $X \leq 0.83$:
- (b) What is the probability $X \geq 0.83$:

(c) What is the probability $0.83 \leq X \leq 0.88$:

EXERCISE 3.6.58. Assume the random variable(s) is from a normal distribution with $\mu = 4.8$ and $\sigma = 1.9$. Answer the following:

(a) What is the probability $X \leq 1.61$:

(b) What is the probability $X \geq 1.61$:

(c) What is the probability $1.61 \leq X \leq 3.66$:

EXERCISE 3.6.59. Assume the random variable(s) is from a normal distribution with $\mu = 2.1$ and $\sigma = 0.6$. Answer the following:

(a) What is the probability $X \leq 1.51$:

(b) What is the probability $X \geq 1.51$:

(c) What is the probability $1.51 \leq X \leq 2.87$:

EXERCISE 3.6.60. Assume the random variable(s) is from a normal distribution with

$\mu = 2.2$ and $\sigma = 0.3$. Answer the following:

- (a) What is the probability $X \leq 2.16$:
- (b) What is the probability $X \geq 2.16$:
- (c) What is the probability $2.16 \leq X \leq 2.31$:

3.6.7 Concepts

The following questions pertain to the Normal distribution, but the concepts go beyond merely understanding the Normal distribution. Use approximately probability (68, 95, 99.7) and what you have learned on basic probability to solve the following problems. Do not use tables or computers. This example is about pencils, but pens, computer parts, automobile parts, etc., could just as easily have been used

Problem 3.6.61.

The scenario: You are planning on buying a machine to produce pencils. You plan to sell pencils of approximately 12 inches in length. You realize that a machine does not exist that can produce 12 inch pencils exactly and all the time. If the pencils produced are between 11.999 and 12.001 inches, that is close enough to 12 inches for you to sell them. You are considering the purchase of two machines. Both machines produce pencils with a normal distribution and $\mu = 12$, but machine one has $\sigma_1 = .0005$ and machine 2 has a $\sigma_2 = .001$.

(a) What is the probability a pencil produced:

1. from machine 1 is greater than 12.001 inches?
2. from machine 1 is less than 12.001 inches?
3. from machine 1 is less than 11.999 inches?
4. from machine 1 is greater than 11.999 inches?
5. from machine 1 is between 11.999 and 12.000 inches?
6. from machine 1 is between 12.000 and 12.001 inches?

7. from machine 1 is between 11.999 and 12.001 inches?
8. from machine 2 is greater than 12.001 inches?
9. from machine 2 is less than 12.001 inches?
10. from machine 2 is less than 11.999 inches?
11. from machine 2 is greater than 11.999 inches?
12. from machine 2 is between 11.999 and 12.000 inches?
13. from machine 2 is between 12.000 and 12.001 inches?
14. from machine 2 is between 11.999 and 12.001 inches?

(b) Answer the following questions in terms of what the different machines mean to you in terms of your business.

1. What percent of pencils do you expect to be able to sell produced from machine 1.
2. What percent of pencils do you expect to be able to sell produced from machine 2.
3. If it costs 5 baht to produce a pencil and you can sell a pencil for 10 baht then
 - (a) what is your expected profit from machine 1 per pencil.
 - (b) what is your expected profit from machine 2 per pencil.
 - (c) what is your expected profit from machine 1 on 1000 pencils.
 - (d) what is your expected profit from machine 2 on 1000 pencils.

For an almost infinite number of problems on many of the probability questions in this chapter the reader should see ?. It is an interactive PDF file utilizing JavaScript to create an almost infinite number of problem-solution sets.

4

Introduction to Sampling and Sampling Distribution

4.1. Sampling: The Basics

This Chapter can be difficult to follow, but within this Chapter are some key concepts. Don't get lost in the details or formulas. Focus on the big picture. There are many books on sampling and this Chapter will not make you an expert within the field of survey sampling and data collection, but it will get you started. For more on sampling the author recommends ? and ?.

Focus on the following key concepts that will be covered:

1. Garbage In Garbage Out (G.I.G.O.)
2. Sampling Error
3. Non-Sampling Error
 - Selection Bias
 - Measurement Error
4. Unbiased
5. Central Limit Theorem

Often a subset of data is collected from a larger group in order to learn about the larger group. This is the basis of sampling. We wish to take a subset of data, *sample*, and use this sample to learn about what is called a *target population*. To learn about the target population, certain population quantities of interest are estimated from the sample.

Terminology and Notation

1. Target population
 - Collection of objects that we want to study and learn about.
 - Example might be all residents in Bangkok.
 - Population size is denoted by upper case N .

- The set of unit labels representing the population units is denoted

$$u = (1, 2, \dots, N)$$

- The vector of the associated y -values of the population variable of interest is denoted

$$\mathbf{y} = (y_1, y_2, \dots, y_N)'$$

- Often within the sampling context the y -values are considered fixed but unknown constants, not random variables. The "random" part is introduced within the sampling.
- In some context the y -values, $\mathbf{y} = (y_1, y_2, \dots, y_N)'$ are considered one realization/observation from $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)'$ random variables.

2. Sample

- A set of units from a larger or equal size collection of units.
- Example: 100 residents in Bangkok selected by any manner from all residents in Bangkok.
- Example: 100 residents in Bangkok selected by any manner from 1000 residents in Bangkok.
- Sample size is denoted by lower case n .
- The set of units selected, units in the sample is denoted

$$s = (i_1, \dots, i_n).$$

- The vector of observed y -values associated with the sample, s , is denoted

$$\mathbf{y}_s = (y_{i_1}, \dots, y_{i_n})'$$

- For simplicity often unit labels are reordered from 1 to n and the vector of observed y -values associated with the sample, s , is denoted

$$\mathbf{y}_s = (y_1, \dots, y_n)'$$

This notation is often used without explanation.

- The *data* in survey sampling is comprised of both s and y_s , denoted as d , where $d = (s, \mathbf{y}_s)$

3. Sampling frame

- A sampling frame is the set of unit labels from which the units to be sampled can come from. Often the sample frame does not include all units in the target population due to various possible reasons.

4. Sampling design

- The way in which a sample is taken. In other words, the sampling methodology used to collect the data.

5. Sampling Error

- The difference between the estimate from a sample versus the value one would obtain from looking at all units within the sampling frame.

6. Non-Sampling Error

- Selection Bias exists if units within the target population have zero probability of being observed. For example, a mobile phone survey to estimate the average income of people in Bangkok. This type of survey would omit all people without mobile phones.
- Measurement Bias exists when the instrument measuring the item of interest is consistently different from the truth in a certain direction. In other words measurement bias is a result of the instrument measuring the item of interest yielding measurement with error, i.e. measurement error. A simple but illustrative example is a scale that consistently reports the weight of individuals 1 kilogram higher than the truth.

Typical Population Quantities of Interest

Population Mean :

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i$$

Population Total :

$$\tau = \sum_{i=1}^N y_i = N \cdot \mu$$

Finite Population Variance :

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2$$

Population Standard Deviation :

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

Population Proportion :

$$\pi = \frac{1}{N} \sum_{i=1}^N z_i,$$

where

$$z_i = \begin{cases} 1, & \text{if units } i \text{ satisfies the specified condition of interest.} \\ 0, & \text{Otherwise.} \end{cases}$$

There are a various ways to take a sample and collect data. Some ways are better than others. Often a sample is taken by those with little knowledge on sampling leading to various data analysis issues. Garbage in garbage out (G.I.G.O.) is very important to think about before sampling. **If a sample is taken properly, often simple statistics can yield great insight. On the other hand, if the sample is not taken properly, simple statistics and advanced statistical techniques may not yield insight, and can actually yield misleading information.** Unfortunately it is easy to understand the cost of sampling but not the importance of sampling. Sampling can be very costly and as a result sampling is often driven by financial concerns, as mentioned the cost of sampling is the easiest to understand.

Many samples are collected by *convenience sampling*. A convenience sample is collected by convenience. Convenience samples are not very costly, but the informa-

tion obtained can be very unreliable and can lead to very biased results. An extreme example of a convenience sample is collecting data on one's friends to learn about the target population, say people living in Bangkok. A more realistic example of a convenience sample, is collecting data from people at various malls in Bangkok on weekends to learn about people in Bangkok. Many statistical techniques you have learned and will learn **do not apply to a convenience sample**. It is easy to understand the number of observations one can obtain through convenience sampling versus more complicated sampling techniques. As with many things, in sampling quality is more important than quantity, again think G.I.G.O. Better to collect fewer observations through proper sampling techniques than convenience sampling. One technique used to aid in increasing the number of units of interest which can yield more reliable estimates of the population quantities of interest is adaptive sampling (e.g. ??).

Most statistical techniques **do apply to samples taken by simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR) from a sufficiently large population**. In SRSWR it is possible to observe the same unit (e.g. person) more than once in the sample. In SRSWOR, once a unit is selected to be in the sample it is removed from the list of units for subsequent selections. Thus in SRSWOR, it is not possible to sample the same unit more than once. SRSWR and SRSWOR are two sampling designs that involve probability sampling. A *probability sample* is a sample in which the units in the target population have a specified probability of being selected. In addition the probability of the sample, s , being observed is independent of the

y -values in the population, \mathbf{y} , that is, $P(s|\mathbf{y}) = P(s)$. For a SRSWR and SRSWOR the units are sampled with equal probability, they are types of equal probability sampling. Unequal probability sampling is when the units are selected with different probabilities. There are various probability sampling designs:

- Simple random sampling with replacement
- Simple random sampling without replacement
- Stratified sampling
- Cluster sampling
- Systematic sampling
- etc.

This chapter will not go into depth of the various sampling designs mentioned above, but will focus more on the fundamental concepts within sampling. The major benefit of taking a probability sample is that it is most reliable for extrapolating results to the target population of interest. In addition, of lesser benefit, often there exists an *unbiased estimator*, $\hat{\theta}$, for the population quantity of interest, θ . Most commonly, the population quantity of interest, θ , is the population mean, μ or the population total, τ . An unbiased estimator is an estimator that has an expected value equal to the parameter of interest. A sampling design unbiased estimator is such that

$$E[\hat{\theta}] = \sum_{s \in \mathcal{S}} P(s) \hat{\theta} = \theta, \quad (4.1)$$

where \mathcal{S} denotes the collection of all possible samples. Note: In this Chapter and when referring to sampling, expectation will be considered taken over all possible samples. The equation ?? can be viewed in a similar manner to equation ??, $E[X] = \sum p_i x_i = \mu_x$, where $\hat{\theta}$ is analogous to x_i and $P(s)$, the probability of observing a specific sample which yields $\hat{\theta}$, is analogous to p_i , the probability of observing x_i . Equation ?? is the expectation of a random variable, X , and equation ?? is the expectation of an estimator, $\hat{\theta}$.

For a biased estimator, the expectation of $\hat{\theta}$ does not equal θ , i.e. $E[\hat{\theta}] \neq \theta$. The bias of $\hat{\theta}$ is defined as the difference between the expectation of $\hat{\theta}$ and the population quantity of interest θ ,

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

One factor in determining which sampling design and which estimator to use to estimate θ , is the mean square error, MSE, of the sampling design and the estimator, (??). The MSE of $\hat{\theta}$ equals the expected squared error,

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E \left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta \right)^2 \\ &= E \left[\left(\hat{\theta} - E(\hat{\theta}) \right)^2 + \left(E(\hat{\theta}) - \theta \right)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) \right] \\ &= E \left[\hat{\theta} - E(\hat{\theta}) \right]^2 + E \left[E(\hat{\theta}) - \theta \right]^2 + 0 \\ &= \text{Var}(\hat{\theta}) + \left[\text{Bias}(\hat{\theta}) \right]^2 \end{aligned}$$

The MSE for an unbiased estimator equals the variance since the bias of the estimator equals zero.

4.2. Illustrative Examples

Imagine sampling 2 units, $n = 2$, from a population of size $N = 7$. Population size is denoted by capital N and sample size lower case n . Table ?? contains the unit labels, numbered 1 to 7, and their associated values. The population mean of this sample is $\mu_y = 57.00$. Table ?? represents all possible samples of size $n = 2$ from a simple random sample **with replacement**. In a SRSWR, each unit has a probability of $1 - (\frac{N-1}{N})^n = 1 - (\frac{6}{7})^2 = \frac{13}{49}$ of being in the sample. Each possible sample in Table ?? is equally likely with probability $\frac{1}{49}$, because there are 49 possible samples. Table ?? illustrates SRSWR with the sampling frame being all units in the target population, or simply the population. Often when sampling some units in the target population are not in the sample frame and thus have a zero probability of being included in the sample. Imagine if a six sided die was used to select the units to be sampled. Thus the units labeled 1 to 6 would have a probability of $\frac{11}{36}$ of being in the sample and unit number 7 has a zero probability. Table ?? illustrates the latter situation, with the sampling frame being units labeled 1 to 6 within the population.

Table ?? represents all possible samples of size $n = 2$ from a simple random sample **without replacement**. In a SRSWOR, each unit has a probability of $\frac{n}{N} = \frac{2}{7}$ of being in the sample, same probability as SRSWR. Each possible sample in Table ?? is equally likely with probability $\frac{1}{42}$, because there are 42 possible samples. For each

sample of size 2, there are two ways it could be obtained, considering order. Ignoring the order a unit is selected there are $\binom{7}{2} = 21$ possible different samples. Table ?? illustrates SRSWOR with the sampling frame being all units in the target population, or simply the population. Often when sampling, some units in the target population are not in the sample frame and thus have a zero probability of being included in the sample. Again imagine if a six sided die was used to select the units to be sampled. On the second roll of the dice if the number obtained equaled that of the first roll, the die would have to be rolled again, to obtain two distinct units for the sample (SRSWOR). Thus the units labeled 1 to 6 would have a probability of $\frac{2}{6}$ of being in the sample and unit number 7 has a zero probability. Table ?? illustrates the latter situation, with the sampling frame being units labeled 1 to 6 within the population.

Tables ??, ??, and ?? illustrate unequal probability sampling with and without replacement. Many samples are taken using unequal probability sampling. Again the majority of basic data analysis techniques assumes equal probability sampling was employed. In general, unequal probability sampling with replacement is much easier than without replacement to calculate estimates of the population quantity of interest. The author wishes to warn the reader to consider carefully before deciding on an unequal probability sample without replacement. Keep in mind that software is getting better and better at analyzing complicated sampling designs, in future the latter statement may not be valid. The main reason for the additional complication with unequal probability sampling without replacement is determining the probability of a specific unit will be in the sample. A general formula for calculating the

probability unit i is in the sample is:

$$P(i \in s) = \sum_{i \in s, s \in \mathcal{S}} P(s)$$

This formula could be used for SRSRWR or SRSWOR. For example, a SRSWR of size $n = 2$ from 6 out of the 7 units, Table ??, the probability of selecting unit 5 equals

$$\begin{aligned} P(i = 5 \in s) &= \sum_{5 \in s, s \in \mathcal{S}} P(s) \\ &= P(1, 5) \quad + P(2, 5) \quad + P(3, 5) \quad + P(4, 5) \quad + P(5, 5) \quad + P(6, 5) \quad + P(7, 5) \\ &= \frac{1}{18} \quad + \frac{1}{18} \quad + \frac{1}{18} \quad + \frac{1}{18} \quad + \frac{1}{36} \quad + \frac{1}{18} \quad + 0 \\ &= \frac{11}{36} \end{aligned}$$

An example with unequal probability sampling without replacement, Table ??, the probability of selecting unit 5 equals

$$\begin{aligned} P(i = 5 \in s) &= \sum_{5 \in s, s \in \mathcal{S}} P(s) \\ &= P(1, 5) \quad + P(2, 5) \quad + P(3, 5) \quad + P(4, 5) \quad + P(6, 5) \quad + P(7, 5) \\ &= 0.021 \quad + 0.057 \quad + 0.010 \quad + 0.064 \quad + 0.010 \quad + 0.092 \\ &= 0.254 \end{aligned}$$

Table 4.1: Imaginary Population of Size $N = 7$

unit label	1	2	3	4	5	6	7
y -value	56	60	45	65	55	46	72

Table 4.2: All Possible Samples of Size $n = 2$ with SRSWR

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)	(3,7)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)	(4,7)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)	(5,7)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)	(6,7)
(7,1)	(7,2)	(7,3)	(7,4)	(7,5)	(7,6)	(7,7)

4.3. Sampling Distribution

The sampling distribution of an estimator, $\hat{\theta}$, follows a discrete distribution within the finite population setting. Thus the expected value of $\hat{\theta}$ defined by equation ?? can also be written as

$$E(\hat{\theta}) = \sum_k k \cdot P(\hat{\theta} = k) \quad (4.2)$$

where

$$P(\hat{\theta} = k) = \sum_{\hat{\theta}=k, s \in \mathcal{S}} P(s).$$

As an example using Table ?? where \bar{y} is $\hat{\theta}$, the expectation of \bar{y} can also be determined using equation ?? as in Table ??.

Table 4.3: SRSWR and Sampling Frame All Units

Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$	Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$
(1,1)	1/49	56.0	1.14	(3,4)	2/49	55.0	2.24
(1,2)	2/49	58.0	2.37	(3,5)	2/49	50.0	2.04
(1,3)	2/49	50.5	2.06	(3,6)	2/49	45.5	1.86
(1,4)	2/49	60.5	2.47	(3,7)	2/49	58.5	2.39
(1,5)	2/49	55.5	2.27	(4,4)	1/49	65.0	1.33
(1,6)	2/49	51.0	2.08	(4,5)	2/49	60.0	2.45
(1,7)	2/49	64.0	2.61	(4,6)	2/49	55.5	2.27
(2,2)	1/49	60.0	1.22	(4,7)	2/49	68.5	2.80
(2,3)	2/49	52.5	2.14	(5,5)	1/49	55.0	1.12
(2,4)	2/49	62.5	2.55	(5,6)	2/49	50.5	2.06
(2,5)	2/49	57.5	2.35	(5,7)	2/49	63.5	2.59
(2,6)	2/49	53.0	2.16	(6,6)	1/49	46.0	0.94
(2,7)	2/49	66.0	2.69	(6,7)	2/49	59.0	2.41
(3,3)	1/49	45.0	0.92	(7,7)	1/49	72.0	1.47
				$E[\bar{y}] = 57.00$			
				$\text{Bias}(\bar{y}) = 0.00$			

Table 4.4: SRSWR and Sampling Frame Units 1-6

Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$	Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$
(1,1)	1/36	56.0	1.56	(3,4)	1/18	55.0	3.06
(1,2)	1/18	58.0	3.22	(3,5)	1/18	50.0	2.78
(1,3)	1/18	50.5	2.81	(3,6)	1/18	45.5	2.53
(1,4)	1/18	60.5	3.36	(3,7)	0	58.5	0.00
(1,5)	1/18	55.5	3.08	(4,4)	1/36	65.0	1.81
(1,6)	1/18	51.0	2.83	(4,5)	1/18	60.0	3.33
(1,7)	0	64.0	0.00	(4,6)	1/18	55.5	3.08
(2,2)	1/36	60.0	1.67	(4,7)	0	68.5	0.00
(2,3)	1/18	52.5	2.92	(5,5)	1/36	55.0	1.53
(2,4)	1/18	62.5	3.47	(5,6)	1/18	50.5	2.81
(2,5)	1/18	57.5	3.19	(5,7)	0	63.5	0.00
(2,6)	1/18	53.0	2.94	(6,6)	1/36	46.0	1.28
(2,7)	0	66.0	0.00	(6,7)	0	59.0	0.00
(3,3)	1/36	45.0	1.25	(7,7)	0	72.0	0.00
				$E[\bar{y}] = 54.50$			
				$\text{Bias}(\bar{y}) = -2.50$			

Table 4.5: All Possible Samples of Size $n = 2$ with SRSWOR

(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)
(2,1)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)
(3,1)	(3,2)	(3,4)	(3,5)	(3,6)	(3,7)
(4,1)	(4,2)	(4,3)	(4,5)	(4,6)	(4,7)
(5,1)	(5,2)	(5,3)	(5,4)	(5,6)	(5,7)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,7)
(7,1)	(7,2)	(7,3)	(7,4)	(7,5)	(7,6)

Table 4.6: SRSWOR and Sampling Frame All Units

Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$
(1,2)	1/21	58.0	2.76
(1,3)	1/21	50.5	2.40
(1,4)	1/21	60.5	2.88
(1,5)	1/21	55.5	2.64
(1,6)	1/21	51.0	2.43
(1,7)	1/21	64.0	3.05
(2,3)	1/21	52.5	2.50
(2,4)	1/21	62.5	2.98
(2,5)	1/21	57.5	2.74
(2,6)	1/21	53.0	2.52
(2,7)	1/21	66.0	3.14
(3,4)	1/21	55.0	2.62
(3,5)	1/21	50.0	2.38
(3,6)	1/21	45.5	2.17
(3,7)	1/21	58.5	2.79
(4,5)	1/21	60.0	2.86
(4,6)	1/21	55.5	2.64
(4,7)	1/21	68.5	3.26
(5,6)	1/21	50.5	2.40
(5,7)	1/21	63.5	3.02
(6,7)	1/21	59.0	2.81
$E[\bar{y}]$	= 57.00		
Bias(\bar{y})	= 0.00		

Table 4.7: SRSWOR and Sampling Frame Units 1 to 6

Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$
(1,2)	1/15	58.0	3.87
(1,3)	1/15	50.5	3.37
(1,4)	1/15	60.5	4.03
(1,5)	1/15	55.5	3.70
(1,6)	1/15	51.0	3.40
(1,7)	0	64.0	0.00
(2,3)	1/15	52.5	3.50
(2,4)	1/15	62.5	4.17
(2,5)	1/15	57.5	3.83
(2,6)	1/15	53.0	3.53
(2,7)	0	66.0	0.00
(3,4)	1/15	55.0	3.67
(3,5)	1/15	50.0	3.33
(3,6)	1/15	45.5	3.03
(3,7)	0	58.5	0.00
(4,5)	1/15	60.0	4.00
(4,6)	1/15	55.5	3.70
(4,7)	0	68.5	0.00
(5,6)	1/15	50.5	3.37
(5,7)	0	63.5	0.00
(6,7)	0	59.0	0.00
$E[\bar{y}]$	= 54.50		
Bias(\bar{y})	= -2.50		

Table 4.8: Imaginary population of size $N = 7$ from Table ??, for unequal probability sampling with probability of selection of unit i equaling p_i .

unit label	1	2	3	4	5	6	7
y -value	56	60	45	65	55	46	72
p_i	0.08	0.20	0.04	0.22	0.12	0.04	0.30

Table 4.9: Unequal Probability Sampling WR and \bar{y}

Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$	Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$
(1,1)	0.006	56.0	0.36	(3,4)	0.018	55.0	0.97
(1,2)	0.032	58.0	1.86	(3,5)	0.010	50.0	0.48
(1,3)	0.006	50.5	0.32	(3,6)	0.003	45.5	0.15
(1,4)	0.035	60.5	2.13	(3,7)	0.024	58.5	1.40
(1,5)	0.019	55.5	1.07	(4,4)	0.048	65.0	3.15
(1,6)	0.006	51.0	0.33	(4,5)	0.053	60.0	3.17
(1,7)	0.048	64.0	3.07	(4,6)	0.018	55.5	0.98
(2,2)	0.040	60.0	2.40	(4,7)	0.132	68.5	9.04
(2,3)	0.016	52.5	0.84	(5,5)	0.014	55.0	0.79
(2,4)	0.088	62.5	5.50	(5,6)	0.010	50.5	0.48
(2,5)	0.048	57.5	2.76	(5,7)	0.072	63.5	4.57
(2,6)	0.016	53.0	0.85	(6,6)	0.002	46.0	0.07
(2,7)	0.120	66.0	7.92	(6,7)	0.024	59.0	1.42
(3,3)	0.002	45.0	0.07	(7,7)	0.090	72.0	6.48
				$E[\bar{y}] = 62.62$			
				$\text{Bias}(\bar{y}) = 5.62$			

4.4. Central Limit Theorem

The central limit theorem (CLT) is one of the most powerful theorems within statistics. The central limit theorem: Let X_1, X_2, \dots, X_n be a random sample from i.i.d. random variables from any distribution with finite mean, μ , and finite variance, σ^2 . Then the limiting distribution of

$$\lim_{n \rightarrow \infty} \left[\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right] \sim N(0, 1), \quad (4.3)$$

where \bar{X}_n is the average of the n sampled observations. That is for a sufficiently large sample size, n , the sample mean \bar{x} from i.i.d. random variables with a finite mean and finite variance has an approximately Normal distribution $N(\mu, \sigma^2/n)$ and $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

Table 4.10: Unequal Probability Sampling WOR and \bar{y}

Sample	$P(s)$	\bar{y}	$P(s)\bar{y}$
(1,2)	0.037	58.0	2.17
(1,3)	0.007	50.5	0.34
(1,4)	0.042	60.5	2.52
(1,5)	0.021	55.5	1.18
(1,6)	0.007	51.0	0.35
(1,7)	0.060	64.0	3.86
(2,3)	0.018	52.5	0.96
(2,4)	0.111	62.5	6.96
(2,5)	0.057	57.5	3.29
(2,6)	0.018	53.0	0.97
(2,7)	0.161	66.0	10.61
(3,4)	0.020	55.0	1.12
(3,5)	0.010	50.0	0.52
(3,6)	0.003	45.5	0.15
(3,7)	0.030	58.5	1.73
(4,5)	0.064	60.0	3.83
(4,6)	0.020	55.5	1.13
(4,7)	0.179	68.5	12.25
(5,6)	0.010	50.5	0.53
(5,7)	0.092	63.5	5.86
(6,7)	0.030	59.0	1.75
$E[\bar{y}]$	$= 62.12$		
Bias(\bar{y})	$= 5.12$		

Table 4.11: Sampling Distribution of \bar{y} with an Unequal Probability Sampling WOR Design.

\bar{y}	$P(\bar{y})$	$P(\bar{y}) \cdot \bar{y}$
45.5	0.003	0.15
50.0	0.010	0.52
50.5	0.017	0.87
51.0	0.007	0.35
52.5	0.018	0.96
53.0	0.018	0.97
55.0	0.020	1.12
55.5	0.041	2.31
57.5	0.057	3.29
58.0	0.037	2.17
58.5	0.030	1.73
59.0	0.030	1.75
60.0	0.064	3.83
60.5	0.042	2.52
62.5	0.111	6.96
63.5	0.092	5.86
64.0	0.060	3.86
66.0	0.161	10.61
68.5	0.179	12.25
$\sum P(\bar{y}) \cdot \bar{y}$	$= 62.12$	$= E[\bar{y}]$

is approximately $N(0, 1)$. In real life σ is almost always unknown, but we can calculate a sample variance, s^2 . If \bar{x} is from a random sample, X_1, X_2, \dots, X_n , from a normal distribution with mean μ , and finite variance, σ^2 then

$$t_{n-1} = \frac{\bar{x} - \mu}{\text{s.e.}(\bar{x})} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$$

has a t -distribution with $d.f. = n - 1$, where $d.f.$ stands for degrees of freedom. For a sufficiently large sample size, n , from any distribution with finite mean and variance

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Typically a minimum of $n > 30$ is desired before assuming a t -distribution when the data are known to come from a non-normal distribution. The t -distribution converges to the normal distribution as $n \rightarrow \infty$, i.e.

$$\lim_{n \rightarrow \infty} [t_{n-1}] \sim N(0, 1).$$

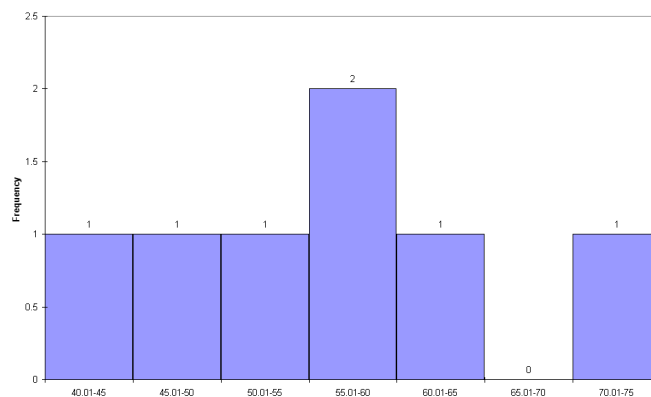


Figure 4.1: Histogram of \bar{y} from population Table ?? for SRSWOR of sample size $n = 1$.

4.5. Examples

4.5.1 Sampling Examples

EXERCISE 4.5.1. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

81	88	72	99	62
----	----	----	----	----

Table 4.12: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.5.2. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size

 Table 4.13: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.5.3. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

78	42	97	74	87
----	----	----	----	----

Table 4.14: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.5.4. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible

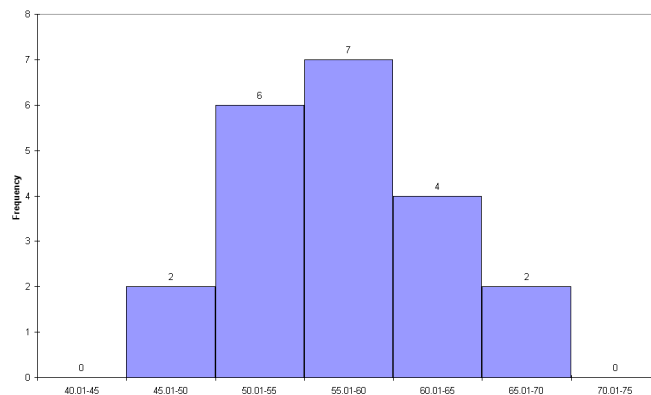


Figure 4.2: Histogram of \bar{y} from population Table ?? for SRSWOR of sample size $n = 2$.

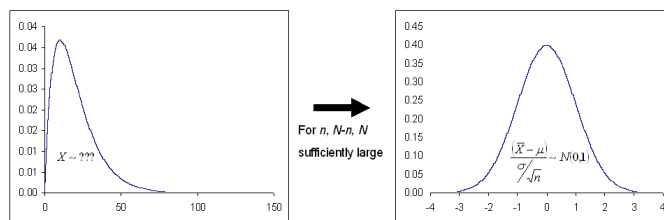


Figure 4.3: Illustrating the Power of the Central Limit Theorem.

sample outcomes assuming a simple random sample without replacement of size $n = 3$.

86 69 63 78 90

Table 4.15: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.5.5. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

4.6. Exercises

4.6.1 Multiple Choice Questions

Click "Begin" and when you are finished click "End". Enjoy.

[Begin Multiple Choice Questions](#)

1. The sample is used to learn about the target population.
(a) True (b) False
2. Advanced statistics are valuable for analyzing a convenience sample.
(a) True (b) False
3. From a convenience sample we can obtain unbiased estimates for the target population.
(a) True (b) False
4. The sample mean, \bar{y} , is always unbiased for μ regardless of how the sample was obtained.
(a) True (b) False
5. A survey with $n = 10,000$ observation is definitely more precise and accurate than a survey with $n = 1,000$ observations.
(a) True (b) False

End Multiple Choice Questions

4.6.2 Sampling Exercises

EXERCISE 4.6.1. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

77	50	49	68	71
----	----	----	----	----

Table 4.17: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.2. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

95	76	81	44	51
----	----	----	----	----

Table 4.18: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.3. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

86	66	61	72	57
----	----	----	----	----

Table 4.19: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.4. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible

sample outcomes assuming a simple random sample without replacement of size $n = 3$.

78	40	77	94	84
----	----	----	----	----

Table 4.20: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.5. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

48	66	94	87	69
<hr/>				

Table 4.21: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.6. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

61	95	97	49	83
<hr/>				

Table 4.22: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for

the mean and median and their biases in relation to μ .

EXERCISE 4.6.7. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

67	64	97	80	79
<hr/>				

Table 4.23: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.8. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

47	80	69	93	89
<hr/>				

 Table 4.24: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.9. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

62	42	98	72	50
----	----	----	----	----

Table 4.25: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

EXERCISE 4.6.10. Consider the table below the entire population data set.

- (a) Given the population data set calculate the mean and median for all possible sample outcomes assuming a simple random sample without replacement of size $n = 3$.

60	68	93	75	64
----	----	----	----	----

Table 4.26: The entire population data set

- (b) Calculate the parameter of interest the population mean, the expected values for the mean and median and their biases in relation to μ .

5

Introduction to Inferential Statistics

5.1. Concept Behind Inferential Statistics

To begin with there are certain fundamental concepts in this section that span the section and are used throughout statistics. Often a sample is taken and the sample is a subgroup of a larger group, the population. From the sample it is desired to learn about the population. For example, a survey is taken on 200 people living in Bangkok and their opinion about the underground train. Results are published and comments are made about it. Is anyone truly concerned about the specific 200 people in the survey? If 200 people do not like the underground train does it really matter? No. In fact there exist well over 200 people in Bangkok that have never taken the underground train. What people really want to learn from the survey is the general opinion within Bangkok about the underground and to do this a sample

of 200 people are surveyed and asked questions. If all 200 people surveyed did not like the underground train, this is of concern only because it leads us to believe that the general populace within Bangkok do not like the underground train and perhaps only a small minority like the train. The sample is almost immediately within our minds extrapolated to the population at large. In this case the population at large is people living in Bangkok. Inferential statistics are used to learn about the population from the sample.

Two common techniques to use a sample to learn about a population that go beyond descriptive statistics are hypothesis testing and confidence intervals. Hypothesis testing is used to test a theory. Confidence intervals are used to obtain a range of values for which you might consider the population mean, μ , to be within. Technically, from a frequentist viewpoint, the population mean is either within the interval or not.

5.1.1 Hypothesis testing

In general within hypothesis testing we wish to test a theory, belief or simply something of interest. It is desired to test if a quantity concerning the population, called a parameter, is either not equal to, greater than or less than some value. Typically, the population mean, μ , or proportion, π , is the parameter, but not always. In hypothesis testing the theory is turned into what is called a null hypothesis, denoted H_0 , and an alternative hypothesis, denoted H_1 or H_A . In general hypothesis testing one may want to compare one group/sample to a specific value, say μ_0 . Often within hypothesis testing one may want to compare two groups/samples to each other, such

as comparing the average salary of men, say μ_1 , to the average salary of women, say μ_2 .

The alternative hypothesis is what is desired to prove or show to be true and the null hypothesis the opposite. Examples: If it is desired to prove the ...

- average income in Bangkok is greater than 30,000 Baht/month:

$$- H_0 : \mu \leq 30,000 \text{ and } H_A : \mu > 30,000.$$

- average income in Bangkok of men is greater than that of women:

$$- H_0 : \mu_{men} \leq \mu_{women} \text{ and } H_A : \mu_{men} > \mu_{women}.$$

- percent of women in Hong Kong is less than 50%:

$$- H_0 : \pi \geq 50\% \text{ and } H_A : \pi < 50\%.$$

- etc.

Table ?? lists various null and alternative hypothesis combinations for one and two sample tests of population mean(s) and proportion(s) and how to calculate their associated p-values¹.

In hypothesis testing a decision is made by using what is known as a *p-value*. The p-value is the probability of observing what was observed or more extreme assuming the null hypothesis is true. If the probability of observing what was observed or more extreme assuming the null hypothesis is true is "very small" the researcher rejects the null hypothesis. The researcher rejects the null hypothesis when the p-value is

¹Note: The Table ?? for calculating p-value assumes variance is known for investigating the population mean, μ .

Investigate	Null Hypothesis	Alternative Hypothesis	Calculate p-value
μ from one group/ sample	$\mu = \mu_0$	$\mu \neq \mu_0$	$2 \times P(Z > z)$
	$\mu \geq \mu_0$	$\mu < \mu_0$	$P(Z < z)$
	$\mu \leq \mu_0$	$\mu > \mu_0$	$P(Z > z)$
π from one group/ sample	$\pi = \pi_0$	$\pi \neq \pi_0$	$2 \times P(Z > z)$
	$\pi \geq \pi_0$	$\pi < \pi_0$	$P(Z < z)$
	$\pi \leq \pi_0$	$\pi > \pi_0$	$P(Z > z)$
μ from two groups/ samples	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$2 \times P(Z > z)$
	$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$P(Z < z)$
	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$P(Z > z)$
π from two groups/ samples	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	$2 \times P(Z > z)$
	$\pi_1 \geq \pi_2$	$\pi_1 < \pi_2$	$P(Z < z)$
	$\pi_1 \leq \pi_2$	$\pi_1 > \pi_2$	$P(Z > z)$

Table 5.1: The more common H_0 and H_A combinations and how to calculate their associated p-values.

small because we trust the data over the null hypothesis. Typically p-values less than that of 0.1, 0.05, or 0.01 are considered too small to be random chance and the null hypothesis is rejected. The value which the null hypothesis will be rejected at is called the *level of significance* and denoted by α . Commonly for large data sets often a significance level of 0.01 is used. Typically in the class room setting an $\alpha = 0.05$ is used.

Important:

If p-value $< \alpha$ then reject H_0

If p-value $\geq \alpha$ then fail to reject H_0

For hypothesis testing regardless of the test chosen and the test-statistic used the steps are generally the same. This book will only cover the p-value approach to hypothesis testing. Other books cover a rejection region as well. The rejection region approach is useful for when a p-value can't be calculated. For example, when

the researcher does not have access to a computer, like on exams. When working, in this day in age the researcher will most likely have access to a computer and almost all, if not all statistical software calculates a p-value for hypothesis testing. For this reason only the p-value approach will be covered.

Steps Within Hypothesis Testing: P-value Approach

1. Determine the null hypothesis, H_0 , and the alternative hypothesis, H_A .
2. Decide on the appropriate level of significance, α .
3. Determine the sample size and sampling design to use.
 - The tests in this chapter are appropriate when the data comes from a simple random sample.
 - The tests in this chapter and other statistical tests are **not** appropriate when the data comes from a convenience or other type of non-probability sample.
4. Determine the appropriate test statistic given the data and sampling design.
5. Collect the data and calculate the appropriate test statistic.
6. Calculate the p-value for the H_0 and H_A combination.
7. Make a decision whether to fail to reject H_0 or reject the H_0 by comparing the p-value to α .

After making a decision there are two possible types of error, type I and type II. A Type I error is when when you reject the null hypothesis and the null hypothesis is actually true. A Type II error is when you fail to reject the null hypothesis and the null hypothesis is actually false, with probability β . The *power* of a test equals $1 - \beta$ which is the probability of rejecting the null hypothesis when the null hypothesis is

false. All possible error/no error results of a hypothesis test are given in Table ??.

	H_0 is true	H_0 is false
Fail to reject H_0	$P(\text{No error})=1 - \alpha$	$P(\text{Type II Error})=\beta$
Reject H_0	$P(\text{Type I error})=\alpha$	$P(\text{No error})=1 - \beta$

Table 5.2: No error, type I and type II error

Important

When a hypothesis test is performed, the result is either **fail to reject** the null hypothesis or **reject** the null hypothesis. Do not say "accept" the null hypothesis. There is a huge difference between not having enough evidence to disprove something and proving something. Reject H_0 is like disproving H_0 and fail to reject H_0 is like failing to disprove H_0 , but this is very different from saying accept H_0 or that H_0 has been proved. This is a very important concept and understanding it will help you avoid much confusion when performing hypothesis testing and working with data.

Scenario

You have a theory that on average men in Bangkok weigh more than 65 kilograms. $H_0 : \mu \leq 65$ and $H_A : \mu > 65$. Data are collected, a simple random sample of size $n = 100$ and the average weight from the sample is $\bar{x} = 67.3$ kilograms. The statistician performs a statistical test and the p-value is 0.23, so he **fails to reject** H_0 . Is he saying he believes the average male weight in Bangkok is less than or equal to 65 kilograms? No! Were he to say he accepts H_0 this would be implying he believes the average weight is less than or equal to 65 kilograms. What the statistician is saying

is that there is not enough evidence to show your theory beyond a reasonable doubt, so he can not reject H_0 . This subtle difference is very important. Imagine saying to someone that the sample average is $\bar{x} = 67.3$ kilograms so you believe the population average is less than or equal to 65 kilograms. That does not make any logical sense. Given the amount of data, the sample average, $\bar{x} = 67.3$, and the sample standard deviation, we are not confident in saying that the population average is greater than 65 kilograms. This is a what is being shown by the hypothesis test. With more information it could possibly be shown that $\mu > 65$. For this particular reason the author tends to prefer looking at confidence intervals for a deeper understanding of what the data collected are saying.

5.1.2 Confidence Intervals

In general when creating what is called a confidence interval, we wish to obtain a range of plausible values for a quantity concerning the population, a parameter. Typically the population mean, μ , or proportion, π , is the parameter, but not always. Also, it is often desired determine a plausible range between two groups/samples to each other, such as comparing the average salary of men, say μ_1 , to the average salary of women, say μ_2 . A $(1 - \alpha) \times 100\%$ confidence interval is the probability of obtaining the parameter of interest under what is known as a Bayesian approach and is often the way a confidence interval is explained. Bayesian's consider the parameter of interest a random variable. The author is a frequentist, and the author considers the parameter to be an unknown constant. Under the frequentist approach, a $(1 - \alpha) \times 100\%$ is

the percent of confidence intervals that are expected to contain the true value of the parameter of interest. This is assuming an infinite number of samples taken of the same size, under a simple random sample. Of course, in reality only a single sample is taken in practice. The confidence interval is thus often considered the range of plausible values the parameter might be, what it is, is unknown in reality though and may or may not be within the interval.

5.2. Hypothesis Tests and Confidence Intervals

5.2.1 One Sample z-test and Confidence Interval

Assumptions: Observations are independent of one another. The data come from a normal distribution with mean, μ , and variance, σ^2 , (i.e. $x_i \sim N(\mu, \sigma^2)$) with σ^2 known or large n with σ^2 known. Note: This test will not be stressed as the population variance is almost always unknown in real life. In addition, software such as SPSS does not even offer this option for the latter reason.

The one sample z-test and confidence interval are used when the true population variance, σ^2 , is known. This is almost never the case, but for large samples we sometimes treat the sample variance as the population variance. In addition, in order to calculate the true population variance you would need to know the true population mean, μ thus making this section more of academic value than practical value.

$$H_0 : \mu = \mu_0$$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

where the z -statistic follows a normal distribution. A $(1 - \alpha)100\%$ confidence interval for μ is,

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

5.2.2 Two Sample z-test and Confidence Interval

Assumptions: Observations are independent of one another. The data come from a normal distribution $x_{1i} \sim N(\mu_1, \sigma_1^2)$ and $x_{2i} \sim N(\mu_2, \sigma_2^2)$, with σ_1^2 and σ_2^2 known or large n_1 and large n_2 with σ_1^2 and σ_2^2 known. Note: This test will not be stressed as again the population variance is almost always unknown in real life. In addition, software such as SPSS, does not even offer this option for the latter reason.

Typically for two samples z-tests the null hypothesis has $\mu_1 = \mu_2$. The null hypothesis can be written as $\mu_1 = \mu_2 + \delta$, where δ is some constant representing the quantity difference between the population means. In this section and other sections difference between μ_1 and μ_2 , is assumed equal to zero, $\delta = 0$.

$$H_0 : \mu_1 = \mu_2$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

A $(1 - \alpha)100\%$ confidence interval for μ is,

$$\left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

5.2.3 One Sample z-test of Proportions and Confidence Interval

Assumptions are that data are from a Binomial distribution and that $n\pi > 15$ and $n(1 - \pi) > 15$. This is an approximation, for small sample size one should use the exact distribution, the Binomial distribution.

$$H_0 : \pi = \pi_0$$

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

A $(1 - \alpha)100\%$ confidence interval for π is,

$$\left(\hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}, \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n} \right).$$

5.2.4 Two Sample z-test of Proportions and Confidence Interval

Assumptions are that data are from two independent samples each from a Binomial distribution and $n_1\pi_1 > 15$, $n_1(1 - \pi_1) > 15$, $n_2\pi_2 > 15$, and $n_2(1 - \pi_2) > 15$. This test is an approximation, as with the one sample z-test of proportions.

$$H_0 : \pi_1 = \pi_2$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where

$$\hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}, \hat{p} = \frac{X}{n}, X = X_1 + X_2, \text{ and } n = n_1 + n_2.$$

A $(1 - \alpha)100\%$ confidence interval for $\pi_1 - \pi_2$ is,

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}$$

5.2.5 One Sample t-test and Confidence Interval

Assumptions: Observations are independent of one another and if the data come from a non-normal distribution, $n > 30$. If the data come from a normal distribution, n can be any size for each sample.

$$H_0 : \mu = \mu_0$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where the t -statistic has a t -distribution with $(n - 1)$ degrees of freedom (d.f.). A $(1 - \alpha)100\%$ confidence interval for μ is,

$$\left(\bar{x} - t_{\alpha/2, (n-1)} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, (n-1)} \frac{s}{\sqrt{n}} \right),$$

5.2.6 Two Sample t-test and Confidence Interval

Assumptions: Observations are independent of one another, if the data comes from a non-normal distribution, $n_1 > 30$ and $n_2 > 30$, and the data come from two independent samples. If the data come from a normal distribution, n_1 and n_2 can be any size for each sample.

$$H_0 : \mu_1 = \mu_2$$

In the case where the variances are not assumed to be equal,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where the t -statistic has a t-distribution with r degrees of freedom (d.f.), and r equals

$$r = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

The value of r is often not an integer, and in that case the nearest integer rounded down is often used. A $(1 - \alpha)100\%$ confidence interval for μ is,

$$\left(\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, r} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, r} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right),$$

In the case where the variances are assumed to be equal,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where the t -statistic has a t -distribution with $(n_1 + n_2 - 2)$ degrees of freedom (d.f.).

A $(1 - \alpha)100\%$ confidence interval for μ is,

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

where,

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

5.2.7 Paired t-test and Confidence Interval

The observations are paired. Let the difference between the pairs be denoted,

$$d_i = x_{1i} - x_{2i},$$

and the mean difference

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i,$$

with standard deviation

$$s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2}$$

For a small sample size it is assumed the differences come from a Normal distribution and for a sample size of $n > 30$ the test is robust to this assumption.

$$H_0 : \mu_d = 0$$

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}.$$

where the t -statistic has a t -distribution with $(n - 1)$ degrees of freedom (d.f.), A $(1 - \alpha)100\%$ confidence interval for the mean difference, μ_d , is,

$$\left(\bar{d} - t_{\alpha/2, (n-1)} \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2, (n-1)} \frac{s_d}{\sqrt{n}} \right)$$

5.3. Two Sided P-value Into A One Sided P-value

SPSS computer output and other software may only offer a two sided, or two tailed p-value. In the typical situation the researcher can simply divide the two sided p-value by two to get the one sided p-value, but not always. When the opposite of what is expected, that is the opposite of the alternative hypothesis occurs, then calculating the p-value does not fall into the typical situation. This section covers how to calculate the one sided p-value for a one sided test given a two sided p-value. For the following we will denote the two sided p-value by "*pval2s*".

1. One sample T-test or Z-test

$$(a) H_A : \mu < \mu_0 \text{ and } \bar{x} < \mu_0 \text{ then p-value} = \frac{pval2s}{2}$$

$$(b) H_A : \mu > \mu_0 \text{ and } \bar{x} > \mu_0 \text{ then p-value} = \frac{pval2s}{2}$$

$$(c) H_A : \mu < \mu_0 \text{ and } \bar{x} > \mu_0 \text{ then p-value} = 1 - \frac{pval2s}{2}$$

$$(d) H_A : \mu > \mu_0 \text{ and } \bar{x} < \mu_0 \text{ then p-value} = 1 - \frac{pval2s}{2}$$

2. Two sample T-test or Z-test

$$(a) H_A : \mu_1 < \mu_2 \text{ and } \bar{x}_1 < \bar{x}_2 \text{ then p-value} = \frac{pval2s}{2}$$

$$(b) H_A : \mu_1 > \mu_2 \text{ and } \bar{x}_1 > \bar{x}_2 \text{ then p-value} = \frac{pval2s}{2}$$

$$(c) H_A : \mu_1 < \mu_2 \text{ and } \bar{x}_1 > \bar{x}_2 \text{ then p-value} = 1 - \frac{pval2s}{2}$$

$$(d) H_A : \mu_1 > \mu_2 \text{ and } \bar{x}_1 < \bar{x}_2 \text{ then p-value} = 1 - \frac{pval2s}{2}$$

5.4. Examples

5.4.1 One-Sample Z-test Examples

EXERCISE 5.4.1. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 99.67$.

496.57	481.19	483.16	482.79	499.04	475.51	501.21
--------	--------	--------	--------	--------	--------	--------

Table 5.3: Randomly selected bank savings account data
in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 480.95$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.2. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 102.24$.

502.54	485.32	512.99	498.19	492.02	491.98
--------	--------	--------	--------	--------	--------

Table 5.4: Randomly selected bank savings account data
in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 499.07$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.3. Use the data below to answer the following questions. Assume the

data comes from a normal distribution and the variance is known. $\sigma^2 = 43.43$.

493.53	488.44	508.48	501.00	501.26	491.99	501.38
--------	--------	--------	--------	--------	--------	--------

Table 5.5: Randomly selected bank savings account data
in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 503.13$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.4. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 317.42$.

504.37	508.40	476.83
--------	--------	--------

Table 5.6: Randomly selected bank savings account data
in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 469.13$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.5. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 51.47$.

494.85	500.47	513.25	504.44	513.86	507.80
--------	--------	--------	--------	--------	--------

Table 5.7: Randomly selected bank savings account data
in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 505.2$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

5.4.2 Two Sample Z-test Examples

EXERCISE 5.4.6. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 8714.75$ and $\sigma_m^2 = 12138.23$.

7371.95	7576.63	7429.32	7574.12	7435.48
---------	---------	---------	---------	---------

Table 5.8: Randomly selected bank savings account data
in U.S. dollars of women.

7491.42	7675.13	7464.83	7409.07
---------	---------	---------	---------

Table 5.9: Randomly selected bank savings account data
in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.

- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.7. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 8850.16$ and $\sigma_m^2 = 48945.11$.

7291.19	7430.00	7524.52	7310.86	7344.95
---------	---------	---------	---------	---------

Table 5.10: Randomly selected bank savings account data in U.S. dollars of women.

7317.83	7097.75	7546.43
---------	---------	---------

Table 5.11: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.8. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 16111.12$ and $\sigma_m^2 = 5683.57$.

7519.47	7263.34	7357.58	7437.59	7424.97	7648.28
---------	---------	---------	---------	---------	---------

Table 5.12: Randomly selected bank savings account data in U.S. dollars of women.

7351.53	7336.27	7352.69	7330.08	7518.70	7294.46
---------	---------	---------	---------	---------	---------

Table 5.13: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.9. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 23738.77$ and $\sigma_m^2 = 60907.13$.

7422.96	7476.31	7717.92
---------	---------	---------

Table 5.14: Randomly selected bank savings account data in U.S. dollars of women.

6983.42	7554.29	7233.37	7110.67
---------	---------	---------	---------

Table 5.15: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.

- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.10. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 72973.52$ and $\sigma_m^2 = 52738.46$.

7049.46	7398.07	7419.66	7779.34	7393.97
---------	---------	---------	---------	---------

Table 5.16: Randomly selected bank savings account data in U.S. dollars of women.

7387.68	7403.09	6968.16	7255.45	6902.56
---------	---------	---------	---------	---------

Table 5.17: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.4.3 One-Sample Test of Proportions Examples

EXERCISE 5.4.11. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 103 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises

is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *greater than* 0.41. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.12. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 103 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *less than* 0.43. Use $\alpha = 0.05$. Do

not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.13. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 99 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *not equal to* 0.49. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.14. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 99 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *not equal to* 0.51. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.15. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 90 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *greater than* 0.48. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

5.4.4 Two-Sample Test of Proportions Examples

EXERCISE 5.4.16. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 102 days from the past 20 years for the SET and 99 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *greater than* the TWSE. Use $\alpha = 0.05$. Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.17. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 110 days from the past 20 years for the SET and 107 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *not equal to* the TWSE. Use $\alpha = 0.05$. Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.18. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 96 days from the past 20 years for the SET and 92 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *less than* the TWSE. Use $\alpha = 0.05$.
Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.19. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 102 days from the past 20 years for the SET and 92 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *greater than* the TWSE. Use $\alpha = 0.05$. Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.20. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 111 days from the past 20 years for the SET and 101 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises is *greater than* the TWSE. Use $\alpha = 0.05$. Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

5.4.5 One-Sample T-test Examples

EXERCISE 5.4.21. Use the data below to answer the following questions.

499.70	508.03	496.18	492.04	487.10	502.06	500.80	513.38	498.13	501.17	501.42
--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

Table 5.18: Randomly selected bank savings account data in U.S. dollars.

(a) Test if the population mean is *not equal to* $\mu = 479.35$. Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.22. Use the data below to answer the following questions.

515.78	481.30	509.18	505.28	508.00	504.66
--------	--------	--------	--------	--------	--------

Table 5.19: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 519.19$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.23. Use the data below to answer the following questions.

501.45	491.96	489.16	502.58	513.39
--------	--------	--------	--------	--------

Table 5.20: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 522.24$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.24. Use the data below to answer the following questions.

500.31	497.19	492.62	499.22
--------	--------	--------	--------

Table 5.21: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 498.26$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.4.25. Use the data below to answer the following questions.

501.88	512.61	497.46
--------	--------	--------

Table 5.22: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 526.98$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

5.4.6 Two Sample T-test - Equal variances assumed Examples

EXERCISE 5.4.26. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7777.06	7380.40	6992.67	7099.02	7476.95	7370.67	7413.97	7765.19	7645.61
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.23: Randomly selected bank savings account data in U.S. dollars of women.

7794.42	7967.22	7594.02	7429.02	7668.39
---------	---------	---------	---------	---------

Table 5.24: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.27. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7157.72	7866.16	7525.54	7565.08	7612.59	7908.91
---------	---------	---------	---------	---------	---------

Table 5.25: Randomly selected bank savings account data in U.S. dollars of women.

7474.35	7542.40	7616.46	7308.31	7497.00
---------	---------	---------	---------	---------

Table 5.26: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.28. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7583.70	7555.84	7430.35	7675.09
---------	---------	---------	---------

Table 5.27: Randomly selected bank savings account data in U.S. dollars of women.

7135.98	7286.23	7086.27	7454.68	7372.76	7239.88	7288.09
---------	---------	---------	---------	---------	---------	---------

Table 5.28: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.29. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7035.77	7455.09	7254.16	7175.42	7282.14	7970.19	7349.48	7350.33	7433.68
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.29: Randomly selected bank savings account
data in U.S. dollars of women.

7566.55	7816.60	7127.88
---------	---------	---------

Table 5.30: Randomly selected bank savings account
data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.30. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7393.70	7891.80	7636.65	7896.52
---------	---------	---------	---------

Table 5.31: Randomly selected bank savings account data in U.S. dollars of women.

7431.46	7765.21	7944.07	7576.29	7713.84
---------	---------	---------	---------	---------

Table 5.32: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.4.7 Two Sample T-test - Equal variances not assumed Examples

EXERCISE 5.4.31. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7668.41	7391.40	7455.85	7565.70	7443.52	7682.36
---------	---------	---------	---------	---------	---------

Table 5.33: Randomly selected bank savings account data in U.S. dollars of women.

7642.85	7788.47	7211.88	7349.28	7261.22	7514.62
---------	---------	---------	---------	---------	---------

Table 5.34: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.32. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7581.17	7535.35	7506.28	7718.34	7562.18	7751.86	7242.93	7403.41	7564.73
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.35: Randomly selected bank savings account data in U.S. dollars of women.

6775.07	7253.71	6912.73	7368.26	6938.54	7546.04	7309.36
---------	---------	---------	---------	---------	---------	---------

Table 5.36: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.33. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7305.89	7293.48	7727.57	7370.36	7438.92	7744.92	7236.42	7567.66	7535.48
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.37: Randomly selected bank savings account data in U.S. dollars of women.

7393.25	7519.27	7083.88	7621.47	7397.49	7508.09	7214.12
---------	---------	---------	---------	---------	---------	---------

Table 5.38: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.34. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7593.25	7435.47	7996.67	7168.08	7346.40
---------	---------	---------	---------	---------

Table 5.39: Randomly selected bank savings account data in U.S. dollars of women.

7516.31	7494.34	7513.33	7357.80
---------	---------	---------	---------

Table 5.40: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.35. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7529.86	7609.24	7327.26	7988.39
---------	---------	---------	---------

Table 5.41: Randomly selected bank savings account data in U.S. dollars of women.

8011.23	7290.20	7497.10	7798.68
---------	---------	---------	---------

Table 5.42: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.4.8 Paired T-test Examples

EXERCISE 5.4.36. Use the data below to answer the following questions.

Before	After
74.94	75.48
75.40	75.65
70.37	82.52
70.70	74.63
79.59	86.81

Table 5.43: Before and after training exam scores.

(a) Test if the population mean exam score after is *greater than* before training.

Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.37. Use the data below to answer the following questions.

Before	After
81.04	84.23
65.84	71.84
81.43	88.39
69.22	73.43

Table 5.44: Before and after training exam scores.

- (a) Test if the population mean exam score after is *greater than* before training.

Use $\alpha = 0.05$.

- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.38. Use the data below to answer the following questions.

Before	After
74.10	84.10
79.43	84.74
63.52	74.08

Table 5.45: Before and after training exam scores.

- (a) Test if the population mean exam score after is *greater than* before training.
Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.39. Use the data below to answer the following questions.

Before	After
76.92	78.55
81.85	82.07
72.22	74.47

Table 5.46: Before and after training exam scores.

- (a) Test if the population mean exam score after is *greater than* before training.
Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.4.40. Use the data below to answer the following questions.

Before	After
76.35	81.93
79.41	79.72
68.92	69.24
75.46	76.22
73.69	74.84
74.40	79.50

Table 5.47: Before and after training exam scores.

(a) Test if the population mean exam score after is *greater than* before training.

Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval for the difference.

5.5. Exercises

5.5.1 One-Sample Z-test Exercises

EXERCISE 5.5.1. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 14.19$.

497.14	492.04	499.15
--------	--------	--------

Table 5.48: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 494.01$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.2. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 130.46$.

496.86	504.34	484.64	509.55	512.74
--------	--------	--------	--------	--------

Table 5.49: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 511.33$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.3. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 74.46$.

500.76	498.94	485.80	505.38
--------	--------	--------	--------

Table 5.50: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 475.17$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.4. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 267.47$.

477.97	498.10	518.37	490.32
--------	--------	--------	--------

Table 5.51: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 509.87$. Use $\alpha = 0.05$.
 (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.5. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 85.34$.

493.53	484.64	506.63	497.98	501.68	518.49	494.67	495.29	510.95	504.96	509.32
--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

Table 5.52: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 502.21$. Use $\alpha = 0.05$.
 (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.6. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 23.98$.

489.49	500.40	502.32	494.02	502.62	501.92	500.54
--------	--------	--------	--------	--------	--------	--------

Table 5.53: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 499.31$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.7. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 111.18$.

495.53	497.42	481.85	507.51
--------	--------	--------	--------

Table 5.54: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 502.65$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.8. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 99.55$.

498	496.61	508.34	499.4	489.83	505.36	487.13	507.1	517.27	507.32	494.58	479.85	499.11
-----	--------	--------	-------	--------	--------	--------	-------	--------	--------	--------	--------	--------

Table 5.55: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 499.09$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.9. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 48.27$.

500.77	513.64	511.12	507.06	520.33
--------	--------	--------	--------	--------

Table 5.56: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 517.79$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.10. Use the data below to answer the following questions. Assume the data comes from a normal distribution and the variance is known. $\sigma^2 = 67.81$.

489.33	508.10	500.13	495.49	489.44
--------	--------	--------	--------	--------

Table 5.57: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 486.47$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

5.5.2 Two Sample Z-test Exercises

EXERCISE 5.5.11. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 21818.52$ and $\sigma_m^2 = 8163.91$.

7338.29	7660.45	7665.78	7476.08	7390.59
---------	---------	---------	---------	---------

Table 5.58: Randomly selected bank savings account data in U.S. dollars of women.

7503.88	7350.27	7577.44	7431.73	7557.93
---------	---------	---------	---------	---------

Table 5.59: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.12. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 21980.04$ and $\sigma_m^2 = 24153.92$.

7331.58	7574.26	7298.08
---------	---------	---------

Table 5.60: Randomly selected bank savings account data in U.S. dollars of women.

7820.29	7637.10	7584.08	7810.27	7489.31	7374.04	7782.02	7673.02
---------	---------	---------	---------	---------	---------	---------	---------

Table 5.61: Randomly selected bank savings account
data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
 (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.13. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 8417.67$ and $\sigma_m^2 = 5725.11$.

7336.30	7478.30	7485.31	7590.94	7523.39
---------	---------	---------	---------	---------

Table 5.62: Randomly selected bank savings account
data in U.S. dollars of women.

7591.37	7586.84	7618.97	7514.21	7572.17	7414.66
---------	---------	---------	---------	---------	---------

Table 5.63: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.14. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 69256.01$ and $\sigma_m^2 = 35013.32$.

7103.86	7593.68	7359.16	7726.00	7259.51
---------	---------	---------	---------	---------

Table 5.64: Randomly selected bank savings account data in U.S. dollars of women.

7197.08	7142.90	7167.69	7588.83	7621.50	7352.53	7330.15	7429.30
---------	---------	---------	---------	---------	---------	---------	---------

Table 5.65: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.15. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 15206.9$ and $\sigma_m^2 = 17521.15$.

7487.44	7489.67	7419.42	7788.16	7664.10	7545.74	7523.83
---------	---------	---------	---------	---------	---------	---------

Table 5.66: Randomly selected bank savings account data in U.S. dollars of women.

7704.06	7536.12	7799.93	7485.13	7715.60
---------	---------	---------	---------	---------

Table 5.67: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.16. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 9749.11$ and $\sigma_m^2 = 6171.83$.

7276.47	7345.62	7414.86	7528.94	7431.71
---------	---------	---------	---------	---------

Table 5.68: Randomly selected bank savings account data in U.S. dollars of women.

7353.63	7513.96	7361.97	7513.64	7488.89
---------	---------	---------	---------	---------

Table 5.69: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.17. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 55717.26$ and $\sigma_m^2 = 11828.6$.

7198.91	7537.53	7218.39	7393.53	7523.76	7037.41	7505.64	7714.63	7701.13
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.70: Randomly selected bank savings account data in U.S. dollars of women.

8030.16	7750.23	7889.84	7952.27	7780.62	7789.43	7776.16	7988.47
---------	---------	---------	---------	---------	---------	---------	---------

Table 5.71: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.18. Use the data below to answer the following questions. Assume

the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 22271.3$ and $\sigma_m^2 = 11944.51$.

7543.91	7753.98	7472.80
---------	---------	---------

Table 5.72: Randomly selected bank savings account data in U.S. dollars of women.

7721.85	7469.60	7717.24	7750.00	7724.97	7830.91	7661.28
---------	---------	---------	---------	---------	---------	---------

Table 5.73: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.19. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 56207.63$ and $\sigma_m^2 = 25181.28$.

7545.26	7821.63	7799.40	7207.25	7333.99	7592.93	7616.33	7627.18	7185.81
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.74: Randomly selected bank savings account data in U.S. dollars of women.

7416.69	7240.93	7544.75
---------	---------	---------

Table 5.75: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.20. Use the data below to answer the following questions. Assume the data comes from a normal distribution. The variances are known, $\sigma_f^2 = 5438.45$ and $\sigma_m^2 = 13288.07$.

7434.93	7371.22	7516.42
---------	---------	---------

Table 5.76: Randomly selected bank savings account data in U.S. dollars of women.

7310.85	7173.77	7345.94	7366.25	7519.57	7393.70	7241.82
---------	---------	---------	---------	---------	---------	---------

Table 5.77: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.5.3 One-Sample Test of Proportions Exercises

EXERCISE 5.5.21. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 99 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *not equal to* 0.59. Use $\alpha = 0.05$.
Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.22. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 110 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *greater than* 0.5. Use $\alpha = 0.05$.
Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.23. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand)

rises. You randomly select 102 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises π is *not equal to* 0.51. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.24. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 99 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises π is *less than* 0.56. Use $\alpha = 0.05$. Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.25. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 87 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *not equal to* 0.43. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.26. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 113 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *less than* 0.47. Use $\alpha = 0.05$. Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.27. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 104 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *less than* 0.49. Use $\alpha = 0.05$. Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.28. You are investigating the probability of the stock market having

a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 90 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises π is *not equal to* 0.48. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.29. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 94 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises π is *not equal to* 0.47. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.30. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) rises. You randomly select 101 days from the past 20 years. You assume each day is independent of one another and that the true unknown probability that the SET rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises π is *less than* 0.47. Use $\alpha = 0.05$. Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

5.5.4 Two-Sample Test of Proportions Exercises

EXERCISE 5.5.31. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 96 days from the past 20 years for the SET and 96 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *less than* the TWSE. Use $\alpha = 0.05$. Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.32. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 105 days from the past 20 years for the SET and 107 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges

movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *not equal to* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.33. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 98 days from the past 20 years for the SET and 98 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *less than* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.34. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 101 days from the past 20 years for the SET and 100 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges

movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *not equal to* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.35. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 102 days from the past 20 years for the SET and 104 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *not equal to* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.36. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 98 days from the past 20 years for the SET and 100 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges

movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises is *not equal to* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.37. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 101 days from the past 20 years for the SET and 88 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

(a) Test if the true probability the SET rises is *less than* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.38. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 102 days from the past 20 years for the SET and 102 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges

movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *less than* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.39. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 99 days from the past 20 years for the SET and 101 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *not equal to* the TWSE. Use $\alpha = 0.05$.

Do not use a continuity correction factor.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.40. You are investigating the probability of the stock market having a positive day. A positive day is defined as the SET (Stock Exchange of Thailand) for \hat{p}_1 and for \hat{p}_2 the TWSE (Taiwan Stock Exchange) rises. You randomly select 96 days from the past 20 years for the SET and 104 days from the past 20 years for the TWSE. You assume each day is independent of one another, the stock exchanges

movements are independent and that the true unknown probability that the SET rises and the TWSE rises is constant although unknown for the past 20 years.

- (a) Test if the true probability the SET rises is *less than* the TWSE. Use $\alpha = 0.05$.
Do not use a continuity correction factor.
- (b) Create a two-sided 95% confidence interval.

5.5.5 One-Sample T-test Exercises

EXERCISE 5.5.41. Use the data below to answer the following questions.

511.43	516.73	510.84	498.68	495.68
--------	--------	--------	--------	--------

Table 5.78: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 500.15$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.42. Use the data below to answer the following questions.

490.81	498.44	506.33	492.15	485.27	488.16	496.32
--------	--------	--------	--------	--------	--------	--------

Table 5.79: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 490.59$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.43. Use the data below to answer the following questions.

519.93	501.40	523.32	498.81
--------	--------	--------	--------

Table 5.80: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 546.98$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.44. Use the data below to answer the following questions.

494.62	484.64	482.23	499.97
--------	--------	--------	--------

Table 5.81: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 504.89$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.45. Use the data below to answer the following questions.

482.67	498.59	494.63	494.42	500.27	510.62	494.62	496.77	499.58	476.80
--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

Table 5.82: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 493.76$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.46. Use the data below to answer the following questions.

500.23	508.87	495.88	516.70	502.07	493.71	506.69
--------	--------	--------	--------	--------	--------	--------

Table 5.83: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 497.42$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.47. Use the data below to answer the following questions.

498.02	483.12	506.97	485.65	512.74	497.50	501.55	502.76
--------	--------	--------	--------	--------	--------	--------	--------

Table 5.84: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 496.82$. Use $\alpha = 0.05$.

- (b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.48. Use the data below to answer the following questions.

499.30	495.34	489.28	514.72
--------	--------	--------	--------

Table 5.85: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *less than* $\mu = 500.18$. Use $\alpha = 0.05$.
(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.49. Use the data below to answer the following questions.

504.44	515.16
--------	--------

Table 5.86: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *greater than* $\mu = 507.02$. Use $\alpha = 0.05$.
(b) Create a two-sided 95% confidence interval.

EXERCISE 5.5.50. Use the data below to answer the following questions.

507.91	501.79	486.84	514.70	489.75	501.56	509.28
--------	--------	--------	--------	--------	--------	--------

Table 5.87: Randomly selected bank savings account data in U.S. dollars.

- (a) Test if the population mean is *not equal to* $\mu = 510.21$. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval.

5.5.6 Two Sample T-test - Equal variances assumed Exercises

EXERCISE 5.5.51. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7225.84	7517.19	7466.25	7485.31	7218.93
---------	---------	---------	---------	---------

Table 5.88: Randomly selected bank savings account data in U.S. dollars of women.

7465.09	7346.55	7345.25	7005.45	7067.95	7294.74	7103.14
---------	---------	---------	---------	---------	---------	---------

Table 5.89: Randomly selected bank savings account data in U.S. dollars of men.

(a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.52. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7693.39	7177.44	7550.21	7387.35	7648.23
---------	---------	---------	---------	---------

Table 5.90: Randomly selected bank savings account data in U.S. dollars of women.

8170.84	7854.56	7827.07	7825.95	7959.07	7819.63	7986.02	8043.40
---------	---------	---------	---------	---------	---------	---------	---------

Table 5.91: Randomly selected bank savings account data in U.S. dollars of men.

(a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.53. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7335.28	7682.37	7611.14	7361.06
---------	---------	---------	---------

Table 5.92: Randomly selected bank savings account data in U.S. dollars of women.

7901.55	7570.49	7535.67	7492.01
---------	---------	---------	---------

Table 5.93: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.54. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7505.15	7772.00	7599.81	7596.26	7452.18	7472.88	7505.98
---------	---------	---------	---------	---------	---------	---------

Table 5.94: Randomly selected bank savings account data in U.S. dollars of women.

7232.18	7041.23	7166.57	7030.32	7359.85	7309.10
---------	---------	---------	---------	---------	---------

Table 5.95: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.55. Use the data below to answer the following questions. Assume

the data comes from a normal distribution. Also assume the variances are equal.

7598.42	7102.93	7726.92	7485.71	7379.37	7565.11	7588.52
---------	---------	---------	---------	---------	---------	---------

Table 5.96: Randomly selected bank savings account
data in U.S. dollars of women.

7379.92	7829.79	7386.90	7672.01	7097.16	7618.63	7075.15	7025.09	7530.69
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.97: Randomly selected bank savings account
data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.56. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7167.57	7319.28	7557.03	7567.24	7713.46	7519.02	7859.12	7341.21	7582.84
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.98: Randomly selected bank savings account
data in U.S. dollars of women.

7337.80	7432.61	7326.26
---------	---------	---------

Table 5.99: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.57. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7639.28	7877.66	7209.00	7511.97	7703.12	7819.04
---------	---------	---------	---------	---------	---------

Table 5.100: Randomly selected bank savings account data in U.S. dollars of women.

7587.07	7212.48	7739.01	7455.43	7300.69	7255.40	7475.13	7422.20	7654.95
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.101: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.58. Use the data below to answer the following questions. Assume

the data comes from a normal distribution. Also assume the variances are equal.

7553.87	7772.14	7752.89	8047.44	7763.90	7632.66	7547.67
---------	---------	---------	---------	---------	---------	---------

Table 5.102: Randomly selected bank savings account
data in U.S. dollars of women.

7915.25	7968.97	7717.82	7788.20	7655.13	7658.98	7285.52	7754.79	7605.84
---------	---------	---------	---------	---------	---------	---------	---------	---------

Table 5.103: Randomly selected bank savings account
data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.59. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7519.03	7707.48	7681.31	7482.55
---------	---------	---------	---------

Table 5.104: Randomly selected bank savings account
data in U.S. dollars of women.

7605.23	7813.90	7572.58	7412.41	7658.36
---------	---------	---------	---------	---------

Table 5.105: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.60. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are equal.

7635.18	7206.64	7590.27	7191.26
---------	---------	---------	---------

Table 5.106: Randomly selected bank savings account data in U.S. dollars of women.

7620.93	7734.67	7273.34	7383.45	7312.91	7069.93
---------	---------	---------	---------	---------	---------

Table 5.107: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.5.7 Two Sample T-test - Equal variances not assumed Exercises

EXERCISE 5.5.61. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7523.41	7437.08	7338.19	7415.50	7638.34
---------	---------	---------	---------	---------

Table 5.108: Randomly selected bank savings account data in U.S. dollars of women.

7273.77	7231.20	7140.85	7094.92	6903.09	7493.58
---------	---------	---------	---------	---------	---------

Table 5.109: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.62. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7299.15	7495.16	7674.21	7442.27	7581.87	7326.55
---------	---------	---------	---------	---------	---------

Table 5.110: Randomly selected bank savings account data in U.S. dollars of women.

6998.72	7466.48	7277.35	7333.27	7322.83	7337.85
---------	---------	---------	---------	---------	---------

Table 5.111: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.63. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7509.22	7319.94	7665.07	7707.58	7364.69	7377.69
---------	---------	---------	---------	---------	---------

Table 5.112: Randomly selected bank savings account data in U.S. dollars of women.

7380.62	7315.53	7113.54	7317.64	7549.99	7404.72
---------	---------	---------	---------	---------	---------

Table 5.113: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.64. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7592.51	7193.81	7569.57	7117.89	7924.66
---------	---------	---------	---------	---------

Table 5.114: Randomly selected bank savings account data in U.S. dollars of women.

7409.05	7504.03	7641.22	7182.75	7782.20
---------	---------	---------	---------	---------

Table 5.115: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.65. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7566.30	7471.68	7445.80
---------	---------	---------

Table 5.116: Randomly selected bank savings account data in U.S. dollars of women.

7246.49	7417.66	8094.39	7219.45	7588.39
---------	---------	---------	---------	---------

Table 5.117: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.66. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7678.85	7963.18	7164.40	7593.50	7668.11
---------	---------	---------	---------	---------

Table 5.118: Randomly selected bank savings account data in U.S. dollars of women.

7481.23	7347.14	7301.77	7709.10	7163.56
---------	---------	---------	---------	---------

Table 5.119: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.67. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7384.92	7376.13	7666.42	7483.36	7312.78	7678.23	7565.21	7540.48
---------	---------	---------	---------	---------	---------	---------	---------

Table 5.120: Randomly selected bank savings account data in U.S. dollars of women.

7732.66	7539.47	7496.76	7466.83	7436.19
---------	---------	---------	---------	---------

Table 5.121: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.68. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7629.81	7410.73	7394.48	7242.63
---------	---------	---------	---------

Table 5.122: Randomly selected bank savings account data in U.S. dollars of women.

7451.81	7310.54	7569.30	7524.42	7192.04	7545.43	7295.89
---------	---------	---------	---------	---------	---------	---------

Table 5.123: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *greater than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.69. Use the data below to answer the following questions. Assume the data comes from a normal distribution. Also assume the variances are unequal.

7533.74	7763.08	7476.79	7551.07	7408.11	7573.48	7511.34
---------	---------	---------	---------	---------	---------	---------

Table 5.124: Randomly selected bank savings account data in U.S. dollars of women.

7578.37	7678.07	7476.56	7757.14	7384.40	7779.96
---------	---------	---------	---------	---------	---------

Table 5.125: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *not equal to* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.70. Use the data below to answer the following questions. Assume

the data comes from a normal distribution. Also assume the variances are unequal.

7735.30	7673.13	7540.48	7567.95	7473.13
---------	---------	---------	---------	---------

Table 5.126: Randomly selected bank savings account data in U.S. dollars of women.

8010.86	7276.33	7483.19	7631.02
---------	---------	---------	---------

Table 5.127: Randomly selected bank savings account data in U.S. dollars of men.

- (a) Test if the population mean savings of women is *less than* men. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.5.8 Paired T-test Exercises

EXERCISE 5.5.71. Use the data below to answer the following questions.

Before	After
76.05	77.40
78.79	79.20
74.07	76.57
66.48	72.68
68.36	72.45

Table 5.128: Before and after training exam scores.

(a) Test if the population mean exam score after is *greater than* before training.

Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.72. Use the data below to answer the following questions.

Before	After
76.94	80.71
76.76	76.91
70.23	76.83
68.36	71.89
75.70	77.18
76.53	78.48
79.30	84.29
71.58	73.69
76.29	76.88

Table 5.129: Before and after training exam scores.

(a) Test if the population mean exam score after is *greater than* before training.

Use $\alpha = 0.05$.

(b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.73. Use the data below to answer the following questions.

Before	After
75.63	76.08
73.64	75.09
77.20	77.65
72.38	74.71
72.90	74.67

Table 5.130: Before and after training exam scores.

- (a) Test if the population mean exam score after is *not equal to* before training. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.74. Use the data below to answer the following questions.

Before	After
68.26	75.10
71.87	81.45
82.50	86.87
72.09	74.31
73.54	79.03

Table 5.131: Before and after training exam scores.

- (a) Test if the population mean exam score after is *greater than* before training.

Use $\alpha = 0.05$.

- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.75. Use the data below to answer the following questions.

Before	After
73.58	79.94
76.07	79.39
80.59	81.44

Table 5.132: Before and after training exam scores.

- (a) Test if the population mean exam score after is *not equal to* before training. Use

$\alpha = 0.05$.

- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.76. Use the data below to answer the following questions.

Before	After
78.56	79.11
73.39	92.57
82.09	82.55
68.24	76.01

Table 5.133: Before and after training exam scores.

- (a) Test if the population mean exam score after is *not equal to* before training. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.77. Use the data below to answer the following questions.

Before	After
72.00	73.42
80.25	84.29
74.79	76.98
73.58	73.80
81.38	88.60
73.66	74.38
73.38	77.46

Table 5.134: Before and after training exam scores.

- (a) Test if the population mean exam score after is *less than* before training. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.78. Use the data below to answer the following questions.

Before	After
78.12	83.17
72.15	73.92
76.94	82.67
73.93	81.20
74.04	82.81
71.09	74.43

Table 5.135: Before and after training exam scores.

- (a) Test if the population mean exam score after is *less than* before training. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.79. Use the data below to answer the following questions.

Before	After
77.66	80.80
78.58	81.84
81.08	82.18
76.16	79.80
74.91	75.61
76.73	77.39

Table 5.136: Before and after training exam scores.

- (a) Test if the population mean exam score after is *not equal to* before training. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

EXERCISE 5.5.80. Use the data below to answer the following questions.

Before	After
80.29	88.74
78.85	82.39
77.63	80.62
74.16	77.30
72.37	73.84
72.94	78.06
70.73	77.20

Table 5.137: Before and after training exam scores.

- (a) Test if the population mean exam score after is *less than* before training. Use $\alpha = 0.05$.
- (b) Create a two-sided 95% confidence interval for the difference.

5.5.9 Multiple Choice

Use $\alpha = 0.05$, where appropriate, to answer the following questions. Click "Begin" and when you are finished click "End."

Begin Multiple Choice Questions

1. Hypothesis testing is used to learn more about the sample and does not concern about the target population.
 - (a) True
 - (b) False

2. You believe that the average time spent in meetings for male managers is longer than that of female managers. Men and women will be denoted by a m and w , respectively. Select the appropriate null and alternative hypothesis.
 - (a) $H_0 : \mu_m \leq \mu_w$ and $H_1 : \mu_m > \mu_w$
 - (b) $H_0 : \bar{x}_m \leq \bar{x}_w$ and $H_1 : \bar{x}_m > \bar{x}_w$
 - (c) $H_0 : \mu_m \geq \mu_w$ and $H_1 : \mu_m < \mu_w$
 - (d) $H_0 : \bar{x}_m \geq \bar{x}_w$ and $H_1 : \bar{x}_m < \bar{x}_w$

3. You believe that the average income of people in Bangkok is less than 30,000 baht. You collect an SRSWOR from Bangkok of 100 people. What test should you use?
 - (a) One Sample Z-test of Proportions
 - (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test
 - (e) Paired T-test

4. You believe the average income of men in Bangkok is less than 35,000 baht. You collect an SRSWOR from Bangkok of 100 men. What test should you use?
- (a) One Sample Z-test of Proportions
 - (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test
 - (e) Paired T-test
5. You believe the average income of men in Bangkok is less than women. You collect an SRSWOR from Bangkok of 100 people. What test should you use?
- (a) One Sample Z-test of Proportions
 - (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test
 - (e) Paired T-test
6. You believe the percentage of people in Bangkok that are unemployed is less than 15%. You collect an SRSWOR from Bangkok of 100 people. What test should you use?
- (a) One Sample Z-test of Proportions
 - (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test

- (e) Paired T-test
7. You believe that the percentage of men in Bangkok that are unemployed is less than of women. You collect an SRSWOR from Bangkok of 100 people. What test should you use?
- (a) One Sample Z-test of Proportions
 - (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test
 - (e) Paired T-test
8. You believe that within married couples husbands are older than wives on average. You collect an SRSWOR from Bangkok of 100 married couples. What test should you use?
- (a) One Sample Z-test of Proportions
 - (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test
 - (e) Paired T-test
9. You believe the percentage of men in Bangkok is greater than that of women. You collect an SRSWOR from Bangkok of 100 people. What test should you use?
- (a) One Sample Z-test of Proportions

- (b) Two Sample Z-test of Proportions
 - (c) One Sample T-test
 - (d) Two Sample T-test
 - (e) Paired T-test
10. The government wants to determine if more than 30% of Thai people living in Bangkok use the sky train. The government takes a random sample of 1000 people from all people living in Bangkok. The statistician creates the null hypothesis and alternative hypothesis as, $H_0 : \pi \leq 30\%$ and $H_1 : \pi > 30\%$, then enters the data and gets a two-sided p-value of 0.0014. Question: Do you believe more than 30% of people in Bangkok use the sky train?
- (a) You think $\pi \leq 30\%$.
 - (b) You think $\pi > 30\%$.
 - (c) Given the information you are not sure.
11. You are producing pens. The mean length should be 12 cm, not less than nor greater than 12 cm. $H_0 : \mu = 12$ cm and $H_1 : \mu \neq 12$ cm. You collect data on 100 randomly selected pens produced and find that $\bar{x} = 11.95$ and $s = 2$ cm.
- (a) You think $\mu = 12$ cm, fail to reject H_0 .
 - (b) You think $\mu \neq 12$ cm, reject H_0 .
 - (c) Given the information you are not sure.
12. You are producing pens. The mean length should be 12 cm, not less than nor greater than 12 cm. $H_0 : \mu = 12$ cm and $H_1 : \mu \neq 12$ cm You collect data on 100

randomly selected pens produced and find $\bar{x} = 12.00$ and $s = 2$ cm.

- (a) You think the machine is working great.
- (b) You think the machine is working alright.
- (c) You think the machine is not working well and producing a lot of pens that can't be sold.

13. A marketing campaign using a letter to sell water guns is mailed in December. The same campaign using a postcard to sell water guns is mailed at the end of March. You believe postcards will yield a better response rate than letters. The statistician creates the null hypothesis and alternative hypothesis as, $H_0 : \pi_{letters} \geq \pi_{postcards}$ and $H_1 : \pi_{letters} < \pi_{postcards}$, then enters the data and calculates the appropriate one-sided p-value of 0.000024 and $\hat{p}_{letters} < \hat{p}_{postcards}$. Question: Are postcards better for selling water guns?

- (a) Letters are better than postcards to sell water guns.
- (b) Letters are worse than postcards to sell water guns.
- (c) Letters are equal to or better than postcards to sell water guns.
- (d) Given the information you are not sure which is better, letters or postcards to sell water guns.

14. You are a manager in a retail store. You are thinking of selling mens boxer briefs (type of underwear). You want to find out what percentage of people that visit your store are interested in purchasing this item of clothing. An employee decides to help and asks the first 200 people that arrive at your store on Wednesday and asks if they are interested. Then your employee creates a 95% C.I. for you, and

it is (7.5%,8.5%). If you believe that more than 10% of all visitors to your store are interested, you plan to sell boxers. From this information you decide:

- (a) to sell boxers.
- (b) not to sell boxers.
- (c) still undecided about selling boxers.

End Multiple Choice Questions

5.5.10 Assignment

This assignment involves analyzing data from a survey from randomly selected people within Bangkok. Data can be found at a link from my website:

www.learnviaweb.com/datasets/datasets.html. The survey responses were simulated. The data is saved in a comma delimited file. You must analyze the data to answer the questions below. There is a dataset for each student. Use the last two digits of your student id# to select the data file. The goal of this assignment as with most assignments is to learn. Although each person has his or her own dataset, you may ask people for help. In addition, you must make graphs using excel, supporting your conclusions. Use an $\alpha = 0.05$ for hypothesis testing. The deliverable is a PowerPoint slide of the graph from Excel with your conclusion for each question and place supporting computer output in the appendix. There are two possible answers for each question:

1. Yes we feel confidently from the sample that ...

2. No, there is insufficient evidence in the sample to confidently state that ...

Note: In the data file a "1" means the attribute has the property and a "0" represents the attribute does not have the property. For gender a "1" represents a male. The material covered in this chapter is not enough to answer all the questions below. Some questions may require ANOVA and Chi-Squared test, which will be covered in the following chapters.

1. Less than 70% of people in Bangkok have internet.
2. A higher percentage of men have Internet than women.
3. Women use the phone more than men.
4. On average people use the phone more than 390 minutes per month.
5. More than 85% of all people are working.
6. A higher percentage of men are working than women.
7. The average usage of the phone of working people is less than that of non-working people.
8. There exists a relationship between the cable company chosen (or no cable) and gender.
9. The average age of people is not the same for each cable company.
10. The average usage of the phone of people with Internet is less than those without Internet.

6

Various Chi-Square Tests and Analysis of Variance

6.1. Introduction to Chi-Square For Categorical Data

The Chi-Square test can be used to test three different hypotheses. Chi-Square tests can test:

1. Goodness of Fit
2. Independence
3. Homogeneity

Fortunately, to calculate the Chi-Square test statistic it is structurally the same for each of the three tests.

**One Variable With More Than Two Categories:
Test of Goodness of Fit**

This test can be performed on two categories, but technically one could use the binomial test using the one sample z-test of proportions.

The assumptions for using the Chi-Square test are:

1. The data come from a random sample. Note: this is standard for most statistical tests.
2. The data are count data, frequencies.
3. This test is an approximation, and for the test statistic to have an approximate χ^2 distribution, and the expected frequency in each cell should be at least 5.

6.1.1 Goodness of Fit Test

The test below is for a single categorical variable. The null hypothesis and alternative hypothesis for the goodness of fit test are

$$H_0 : \pi_1 = \pi_{1,0}, \pi_2 = \pi_{2,0}, \dots, \pi_k = \pi_{k,0}, \text{ and}$$

H_A : At least one of the probabilities does not equal its hypothesized value.

The $\pi_{1,0}, \pi_{2,0}, \dots, \pi_{k,0}$ represent the hypothesized probabilities of each category. The Chi-Squared test statistic is:

$$\chi^2 = \sum \frac{[O_i - E_i]^2}{E_i},$$

where O_i is the observed frequency and $E_i = n\pi_{i,0}$ is the expected frequency. The test statistic follows, approximately, a Chi-squared distribution with $k - 1$ degrees of freedom.

For example, imagine that it is believed that for mobile phones black is chosen 70% of the time, then pink at 20% and others at 10%. This would be the null hypothesis.

$$H_0 : \pi_{black} = .7, \pi_{pink} = .2 \text{ and } \pi_{others} = .1$$

The researcher decides to reject the null hypothesis at an alpha level of 0.05. A random sample of 100 people that have mobile phones are asked the color of their phone. From the survey, 75 people have black phones, 10 people pink, and 15 another color. Thus

$$\chi^2 = \frac{(75 - 70)^2}{70} + \frac{(10 - 20)^2}{20} + \frac{(15 - 10)^2}{10} = 7.857$$

which has approximately a Chi-squared distribution with (3-1) degrees of freedom. Thus since $7.857 > 5.991$ the researcher decides to reject the null hypothesis and believes that the percent of black, pink and other phones are not the percent previously believed.

6.1.2 Independence and Homogeneity Test

The test below is used when investigating the relationship between two categorical variables.

$$\chi^2 = \sum \frac{[O_{ij} - E_{ij}]^2}{E_{ij}},$$

where O_{ij} is the observed frequency and E_{ij} is the expected frequency. E_{ij} equals the product of the total of row i and total of column j divided by n , the overall total. Using mathematical notation

$$E_{ij} = \frac{O_{i+} \times O_{+j}}{n},$$

see Table ???. The Chi-squared statistic has an approximate Chi-squared distribution with $(R - 1)(C - 1)$ degrees of freedom ($d.f$), where R is the number of rows and C is the number of columns. For example say a researcher wants to investigate gender and religion. Gender (male/female) and religion will be broken into three categories: Buddhist, Christian, and other. The Chi-Squared test can be thought of testing either, independence of gender and religion or homogeneity of proportions between gender and religion. The reason is that if there is no association (independence) between gender and religion, then the proportion of men and women of each religion should be the same and vice versa.

6.2. Analysis of Variance

Analysis of variance, or ANOVA for short, is used to test if the population means of several groups are all equal or if at least one of the groups has a population mean

		Column				Row Totals
		1	2	...	C	
Row	1	O_{11}	O_{12}	\cdots	O_{1c}	O_{1+}
	2	O_{21}	O_{22}	\cdots	O_{2c}	O_{2+}
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	r	O_{r1}	O_{r2}	\cdots	O_{rc}	O_{r+}
Column Totals		O_{+1}	O_{+2}	\cdots	O_{+c}	n

Table 6.1: General R by C crosstab (Contingency Table).

that differs. Thus the null hypothesis and alternative hypothesis for testing k groups is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \text{ and}$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one } i \text{ and } j \text{ combination.}$$

ANOVA tests the equality of the populations means by comparing variances – thus the name analysis of variance. An important question to answer before covering ANOVA further is a "Why is ANOVA useful; why not multiple t-tests?" Each hypothesis test has a probability of α of rejecting the null hypothesis even if the null hypothesis is true, a type I error. The latter statement is true for a t-test as well. Thus

1. One t-test has a probability of type I error equal to $1 - (1 - \alpha)^1 = \alpha$.
2. Two t-tests have a probability of type I error equal to $1 - (1 - \alpha)^2$.
3. Three t-tests have a probability of type I error equal to $1 - (1 - \alpha)^3$.
4. Four t-tests have a probability of type I error equal to $1 - (1 - \alpha)^4$.

5. etc.

To test if one or more of the population means differs from any of the others for k groups, you have to do $\binom{k}{2}$ t-tests. If $k = 5$, then you have to test $\binom{5}{2} = 10$ t-tests and the probability of type one error for any of the ten tests with an $\alpha = 0.05$ equals $1 - (1 - .05)^{10} = 0.401$, larger than 40%. That is an extremely high probability of making a type one error or any error for that matter.

Using ANOVA to test if the population means of several groups are all equal or if at least one of the groups has a population mean that differs using α , yields the desired probability of a type I error of α .

6.2.1 One-way ANOVA

One-way ANOVA is used when you have a single categorical variable and a single continuous variable. It is called one-way ANOVA since there is only a single categorical variable. This chapter will also cover two-way ANOVA, which gets its name from the fact that it involves two categorical variables. This chapter does not go into depth about one-way nor two-way ANOVA as they can be viewed as general linear models with a single continuous dependent variable and one or two categorical independent variables. General linear models will be covered in the following chapter. The assumptions for using a one-way ANOVA are:

1. The samples for each group are from a simple random sample, and units are independent of one another.
2. Groups differ by only the factor being studied; everything is the same except

for the factor being studied.

3. Data come from a normal distribution.
4. Equal variance across the groups, this assumption is more robust to departure when the sample size for all groups is equal.

The overall variation, SST, can be broken into two parts:

1. The variation among the groups or sums of squares among (SSA), resulting from the differences among the group means.
2. The variation within the groups or sums of squared error (SSE), resulting from the unattributable randomness within the groups.

$$SST = SSA + SSE$$

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{..})^2 = \sum_{j=1}^k n_j (\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})^2$$

where $\bar{x}_{..}$ is the overall average, $\bar{x}_{.j}$ is the average of the j^{th} group, and x_{ij} is the i^{th} observation of the j^{th} group. The total sample size is denoted n which equals the sum of the samples from all groups, $n = \sum_{j=1}^k n_j$, where the number of groups is k .

The mean sum of squares for the among and error are $MSA = \frac{SSA}{k-1}$ and $MSE = \frac{SSE}{n-k}$, respectively. If the statistic $F = \frac{MSA}{MSE}$ is large then we reject the null hypothesis, where large can be determined by the p-value in the computer output. As with the previous chapters on hypothesis testing, if the p-value is less than α we reject the null

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F-value
Factor A	SSA	$k - 1$	$MSA = \frac{SSA}{k-1}$	$\frac{MSA}{MSE}$
Error	SSE	$n - k$	$MSE = \frac{SSE}{n-k}$	
Total	SST	$n - 1$		

Table 6.2: One-way ANOVA table, with a total of n observations.

hypothesis. The concept is that if the variation among the groups is large relative to within the groups then it is the result that at least one of the population means differs. Figure ?? illustrates this concept. A typical one-way ANOVA table looks like Table ??.

From ANOVA we can determine whether the population group means differ or not. We cannot determine which population means differ if they differ. There exist various post hoc tests to determine which population means differ among the groups. One common post hoc test used is called the Bonferoni test. The test is a multiple comparison test, comparing all possible combinations of the groups. To determine whether or not the population means of two groups differ using a Bonferoni test, a p-value less than α can also be used.

6.2.2 Two-way ANOVA

Two-way ANOVA handles two categorical variables. Two-way ANOVA can also investigate the interaction of the two categorical variables and its relation to the continuous dependent variable. For a two-way ANOVA the the sums of squares total can be broken down into four sums of squares, the sums of squares of factor A, sums

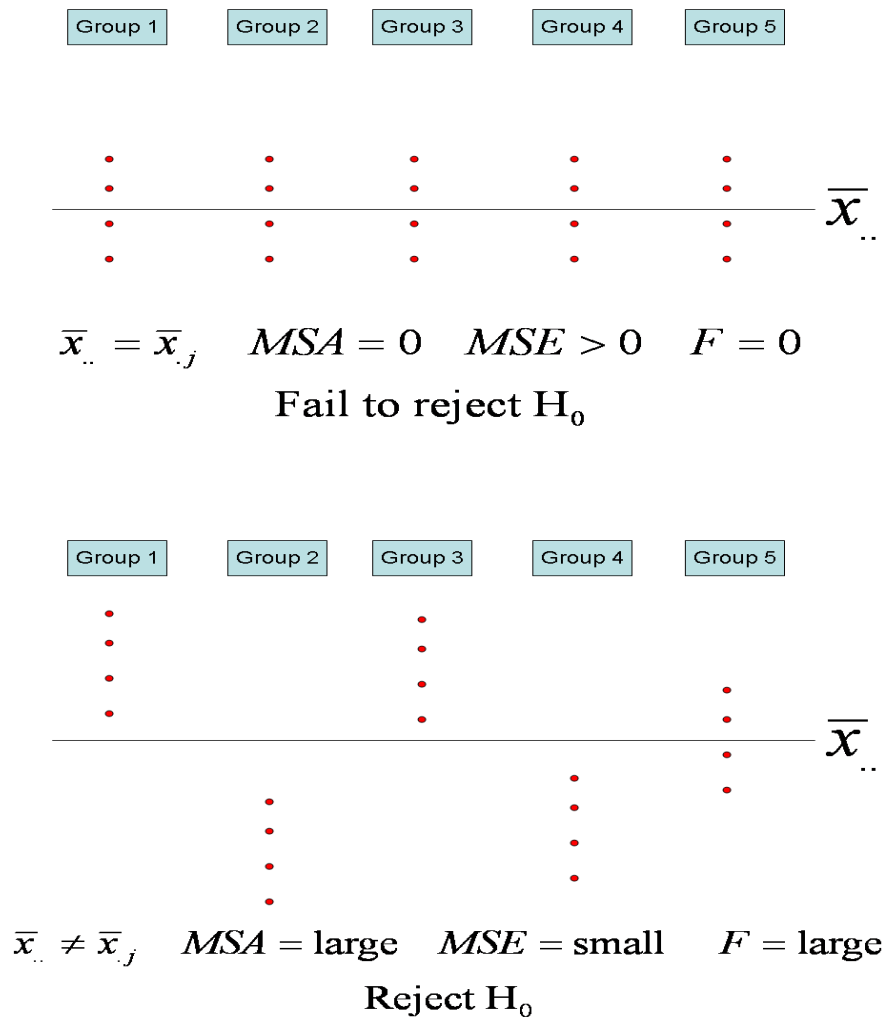


Figure 6.1: Illustrating the concept behind ANOVA.

of squares factor B, sums of squares of the interaction of A and B, plus the sums of squares error. Where the relevant sums of squares are denoted below:

$$\begin{aligned}
 SST &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (x_{ijk} - \bar{x}_{...})^2, \\
 SSA &= bc \sum_{i=1}^a (\bar{x}_{i..} - \bar{x}_{...})^2, \\
 SSB &= ac \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x}_{...})^2, \\
 SSAB &= c \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2, \text{ and} \\
 SSE &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (x_{ijk} - \bar{x}_{ij.})^2,
 \end{aligned}$$

where there are a levels for factor A and b levels for factor B and c observations at each factor combination. Thus there are $n = abc$ total observations. For two-way

ANOVA with interaction the sums of squares total are broken down as follows:

$$\begin{aligned}
 SST = & SSA + SSB + SSAB + SSE \\
 \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (x_{ijk} - \bar{x}_{...})^2 = & bc \sum_{i=1}^a (\bar{x}_{i..} - \bar{x}_{...})^2 + \\
 & ac \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x}_{...})^2 + \\
 & c \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2 + \\
 & \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (x_{ijk} - \bar{x}_{ij.})^2,
 \end{aligned}$$

where

$$\begin{aligned}
 \bar{x}_{ij.} &= \frac{1}{c} \sum_{k=1}^c x_{ijk}, \\
 \bar{x}_{i..} &= \frac{1}{bc} \sum_{j=1}^b \sum_{k=1}^c x_{ijk}, \\
 \bar{x}_{.j.} &= \frac{1}{ac} \sum_{i=1}^a \sum_{k=1}^c x_{ijk}, \text{ and} \\
 \bar{x}_{...} &= \frac{1}{abc} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c x_{ijk}.
 \end{aligned}$$

A typical two-way ANOVA table with interaction looks like Table ??.

Source	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F-value
Factor A	SSA	$a - 1$	$MSA = \frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Factor B	SSB	$b - 1$	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
Factor AB	SSAB	$(a - 1)(b - 1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{MSE}$
Error	SSE	$ab(c - 1)$	$MSE = \frac{SSE}{ab(c-1)}$	
Total	SST	$abc - 1$		

Table 6.3: Two-way ANOVA table with interaction and c observations at each factor combination.

6.3. Examples

6.3.1 Chi-Square Test Examples

EXERCISE 6.3.1. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	68	54	53
No	47	58	67
Not Sure	64	60	67

Table 6.4: The survey results

EXERCISE 6.3.2. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	68	72	73
No	53	75	72
Not Sure	80	53	81

Table 6.5: The survey results

EXERCISE 6.3.3. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	74	79	55
No	65	58	67
Not Sure	61	65	68

Table 6.6: The survey results

EXERCISE 6.3.4. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	77	73	64
No	57	58	59
Not Sure	70	55	81

Table 6.7: The survey results

EXERCISE 6.3.5. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	86	53	53
No	56	67	53
Not Sure	64	71	66

Table 6.8: The survey results

6.3.2 Analysis of Variance Examples

EXERCISE 6.3.6. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	-0.2
2	1.0	-3.9
3	2.0	-1.2
4	2.0	6.8
5	3.0	3.7
6	3.0	5.5

Table 6.9: The results from the experiment.

EXERCISE 6.3.7. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	4.4
2	1.0	2.9
3	2.0	1.4
4	2.0	2.6
5	3.0	1.9
6	3.0	-2.7

Table 6.10: The results from the experiment.

EXERCISE 6.3.8. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	1.3
2	1.0	3.1
3	2.0	5.1
4	2.0	3.6
5	3.0	5.1
6	3.0	8.8

Table 6.11: The results from the experiment.

EXERCISE 6.3.9. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	-0.9
2	1.0	2.4
3	2.0	7.8
4	2.0	3.0
5	3.0	3.3
6	3.0	5.0

Table 6.12: The results from the experiment.

EXERCISE 6.3.10. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	0.7
2	1.0	3.0
3	2.0	6.1
4	2.0	-2.0
5	3.0	-0.1
6	3.0	-1.2

Table 6.13: The results from the experiment.

6.4. Exercises

6.4.1 Chi-Square Test Examples

EXERCISE 6.4.1. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	63	69	62
No	58	64	60
Not Sure	57	63	78

Table 6.14: The survey results

EXERCISE 6.4.2. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	82	69	60
No	58	62	65
Not Sure	65	65	85

Table 6.15: The survey results

EXERCISE 6.4.3. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	68	61	60
No	57	76	63
Not Sure	69	70	79

Table 6.16: The survey results

EXERCISE 6.4.4. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	82	68	63
No	76	59	55
Not Sure	67	61	88

Table 6.17: The survey results

EXERCISE 6.4.5. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	73	66	50
No	66	51	52
Not Sure	55	57	77

Table 6.18: The survey results

EXERCISE 6.4.6. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	78	66	60
No	69	47	59
Not Sure	66	53	81

Table 6.19: The survey results

EXERCISE 6.4.7. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	70	67	48
No	56	55	58
Not Sure	58	56	69

Table 6.20: The survey results

EXERCISE 6.4.8. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	67	55	58
No	56	53	56
Not Sure	66	48	86

Table 6.21: The survey results

EXERCISE 6.4.9. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	73	58	64
No	57	55	65
Not Sure	55	67	66

Table 6.22: The survey results

EXERCISE 6.4.10. Perform a chi-square test on the following resulting survey data, set $\alpha = 0.05$.

	Staff	Manager	Senior Manager
Yes	69	60	64
No	78	63	58
Not Sure	46	60	81

Table 6.23: The survey results

6.4.2 Analysis of Variance Examples

EXERCISE 6.4.11. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	1.0
2	1.0	-1.0
3	2.0	-2.2
4	2.0	-0.3
5	3.0	3.4
6	3.0	-0.1

Table 6.24: The results from the experiment.

EXERCISE 6.4.12. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	-2.8
2	1.0	2.8
3	2.0	0.8
4	2.0	-0.1
5	3.0	1.7
6	3.0	3.8

Table 6.25: The results from the experiment.

EXERCISE 6.4.13. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	6.4
2	1.0	3.0
3	2.0	8.8
4	2.0	2.3
5	3.0	5.3
6	3.0	-1.5

Table 6.26: The results from the experiment.

EXERCISE 6.4.14. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	4.3
2	1.0	4.4
3	2.0	-0.6
4	2.0	1.0
5	3.0	1.0
6	3.0	8.8

Table 6.27: The results from the experiment.

EXERCISE 6.4.15. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	3.5
2	1.0	-0.5
3	2.0	3.7
4	2.0	-1.1
5	3.0	3.8
6	3.0	3.4

Table 6.28: The results from the experiment.

EXERCISE 6.4.16. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	0.9
2	1.0	1.3
3	2.0	4.8
4	2.0	0.8
5	3.0	3.5
6	3.0	2.2

Table 6.29: The results from the experiment.

EXERCISE 6.4.17. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	2.4
2	1.0	2.2
3	2.0	4.0
4	2.0	-0.4
5	3.0	0.9
6	3.0	-0.6

Table 6.30: The results from the experiment.

EXERCISE 6.4.18. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	-0.5
2	1.0	-2.1
3	2.0	5.0
4	2.0	1.2
5	3.0	1.5
6	3.0	5.9

Table 6.31: The results from the experiment.

EXERCISE 6.4.19. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	4.2
2	1.0	1.9
3	2.0	0.2
4	2.0	3.4
5	3.0	2.8
6	3.0	3.8

Table 6.32: The results from the experiment.

EXERCISE 6.4.20. Solve for all the parts in the ANOVA table.

	Factor	Response
1	1.0	-0.7
2	1.0	2.7
3	2.0	1.7
4	2.0	2.0
5	3.0	5.9
6	3.0	1.5

Table 6.33: The results from the experiment.

7

Introduction to General Linear Models

7.1. Simple Linear Regression

Simple linear regression involves two variables, a dependent (y) and an independent variable (x). The term "simple" in simple linear regression refers to the fact that there is only a single dependent and independent variable. Simple linear regression assumes a linear relationship between y and x .

**The Simple Linear Regression Model For n
Observations**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where

$E(y) = \beta_0 + \beta_1 x$, the expected value of y given a value of x .

- The ϵ_i = Random error component and the $E[\epsilon_i] = 0 \quad \forall i$
 - The error for each observation is assumed to be independent of one another. $cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$
 - the variance of the error is assumed to be the same regardless of the value of x , $var(\epsilon_i) = \sigma^2$
- β_0 is the y -intercept of the line, the expected value of y at $x = 0$.
- β_1 is the slope of the line, the amount of increase (or decrease) y for a single unit increase in x .

The correlation, r , between x and y is the strength of the linear relationship between x and y . The correlation can range between -1 and 1, and the sample correlation is calculated as follows:

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

A positive r represents a positive relationship between x and y , i.e. a higher x expect a higher y . A negative r represents a negative relationship between x and y , i.e. a higher x expect a lower y . See Figure ?? for visual examples of correlation. The R-squared is often used to determine how well x explains y . R-squared is the percent of variation in y that can be explained by x , and ranges between 0% and 100% and $R\text{-squared} = r^2$ for simple linear regression.

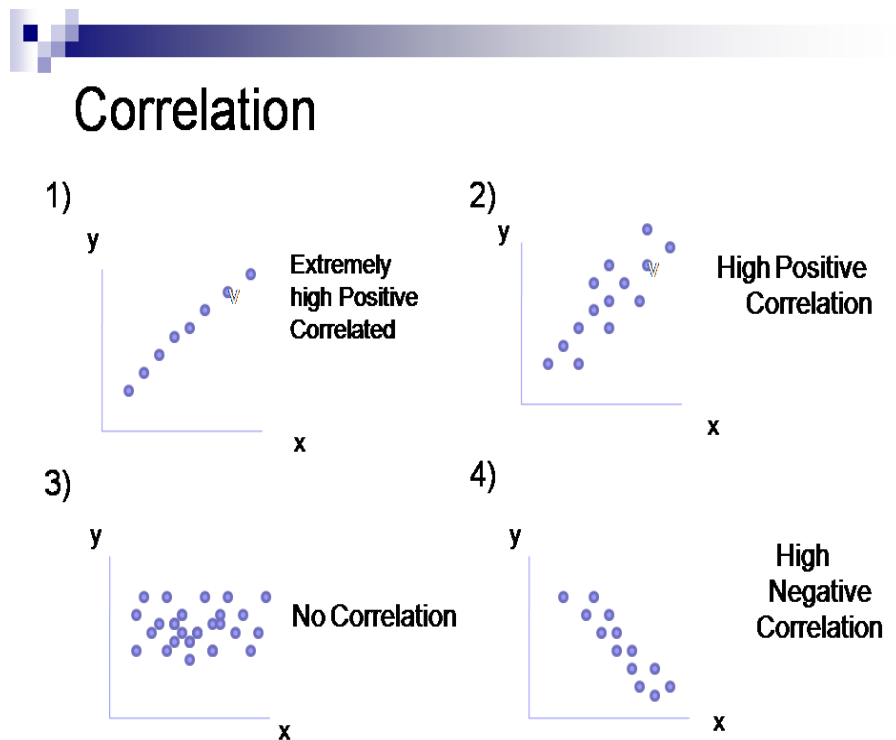


Figure 7.1: Examples of correlation

Formulas for the Least Squares Estimates

The parameters β_0 and β_1 are unknown and must be estimated using data collected. The following formula is used to predict y and varying levels of x ,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where

$$\text{slope: } \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

and

$$y\text{-intercept: } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The difference between y and \hat{y} is the error, $e_i = y_i - \hat{y}_i$. The estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, can be derived by minimizing the sums of squared error (?)

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2.$$

7.2. Multiple Linear Regression Model and General Linear Model

The multiple linear regression model and the general linear model can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

where y is still the dependent variable, but now there is more than one independent variable, x_1, x_2, \dots, x_k . In simple linear regression and multiple linear regression, the independent variables are considered continuous, in the general linear model the independent variables can be categorical, thus making GLM, more general. Simple linear regression is a subset of multiple linear regression, which is a subset of general linear models. Although the model has the term “linear” the function for the model does not need to be linear. For example, one of the independent variables denoted by x_k could actually represent x_k^2 , or $\frac{1}{x_k}$. The general linear model where the dependent variable is a quadratic function of a single independent variable could be written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad i = 1, 2, \dots, n.$$

7.3. Incorporating Categorical Data Into A GLM

How can we do computations on categorical data? How do we plug a categorical variable like gender (male/female) into a mathematical equation? Categorical variables must be converted into one or more indicator variables. An indicator variable is a variable that can take on two values: a zero or a one. A zero indicates the absence of the attribute and a one indicates the presence of the attribute. The number of indicator variables needed for a categorical variable with H levels is $H - 1$.

Examples

- Gender can either be male or female. An indicator variable for gender could be:

$$x_{gender} = 0 \text{ if female, and}$$

$$x_{gender} = 1 \text{ if male.}$$

- Highest level of education obtained, B.A., M.A., or Ph.D.

$$x_1 = 1 \text{ if M.A, and}$$

$$x_1 = 0 \text{ otherwise}$$

$$x_2 = 1 \text{ if Ph.D, and}$$

$$x_2 = 0 \text{ otherwise}$$

- If $x_1 = x_2 = 0$ then the highest level of education obtained must be B.A.

- etc.

The figures ?? and ?? illustrate the importance of utilizing indicator variables when analyzing categorical variables, as opposed to treating categorical variables as continuous variables.

7.4. Hypothesis Testing For GLM

1. Simple linear regression

- (a) If the p-value for the model is less than alpha, then it is believed the single independent variable is helpful in predicting/understanding the dependent variable.
- (b) If the p-value for the model is greater than alpha, then it is believed the single independent variable is not helpful in predicting/understanding the dependent variable and that $\beta_1 = 0$.



Ordinal Independent Variable in GLM

$$X_1 = \begin{cases} 1 & \text{if } M.A. \\ 0 & \text{otherwise} \end{cases}$$

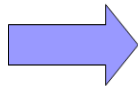
$$X_2 = \begin{cases} 1 & \text{if } Ph.D. \\ 0 & \text{otherwise} \end{cases}$$

If Average Salary

B.A. = 7,000

M.A. = 12,000

Ph.D. = 15,000



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\beta_0 = 7,000$$

→ B.A.

$$\beta_1 = 5,000$$

→ Increase from B.A. to M.A.

$$\beta_2 = 8,000$$

→ Increase from B.A. to Ph.D.

Figure 7.2: Proper handling of categorical data within a general linear model.

2. Multiple linear regression

- (a) If the p-value for the model is less than alpha then it is believed at least some of the independent variables are helpful in predicting/understanding the dependent variable.
- (b) If the p-value for the individual independent variable is less than alpha this implies that the particular independent variable is helpful in predicting/understanding the dependent variable.
- (c) if the p-value for the independent variable, x_i , is greater than or equal to

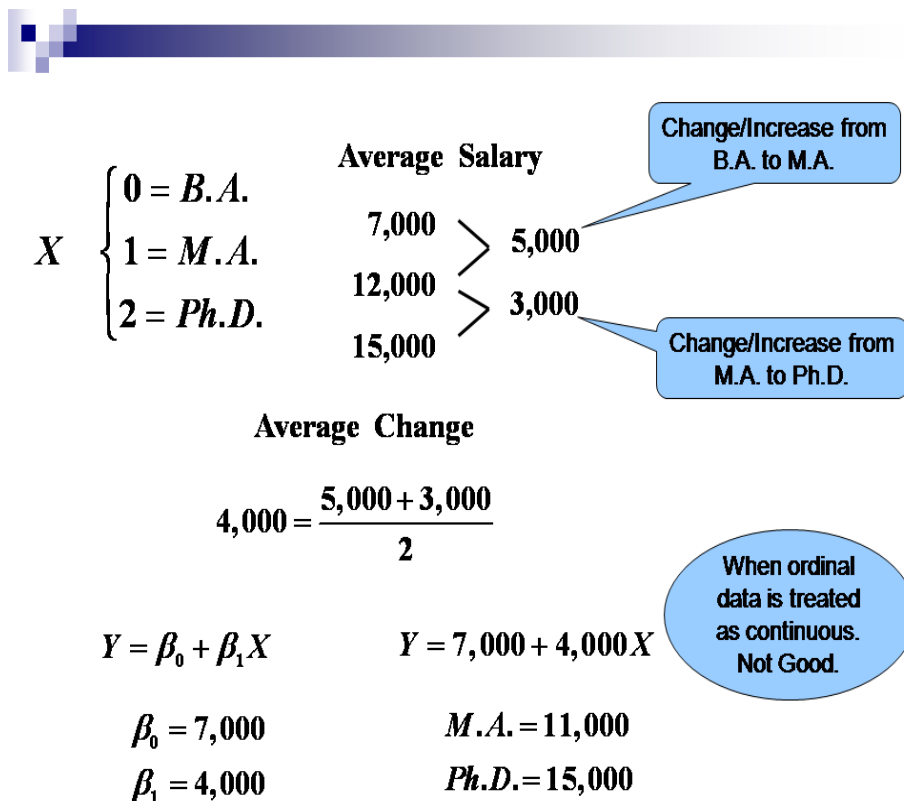


Figure 7.3: Improper handling of categorical data within a general linear model.

alpha it is believed that $\beta_i = 0$, and thus the independent variable can be removed from the model.

3. General linear model

- (a) If the p-value for the model is less than alpha then it is believed that at least some of the independent variables are helpful in predicting/understanding the dependent variable.
- (b) If the p-value for the individual independent continuous variable is less than alpha this implies that the particular independent variable is helpful

in predicting/understanding the dependent variable.

(c) If the p-value for the individual independent categorical variable is less than alpha this implies that the particular independent variable is helpful in predicting/understanding the dependent variable.

- If some categories of a categorical variable are significant (p-value less than alpha) but others are not, use the categorical variable as it is.

- Handling this situation in a better manner goes beyond the scope of this text.

(d) If the p-value for the independent variable, x_i , is greater than or equal to alpha it is believed that $\beta_i = 0$, and thus the independent variable can be removed from the model.

7.5. Introduction to Time Series Regression

7.5.1 Introduction to Time Series

Time series is appropriate when dealing with time series data. Time series data is data collected over time at equally spaced intervals. Stock market data is one type of times series data. Some other examples are:

- Yearly sales data
- Quarterly revenue data
- Yearly gross domestic product data

- etc.

Typically with time series data the researcher is trying to forecast the future using the knowledge of the past. Some examples:

- Predicting the next years sales
- Predicting the next quarters revenue
- Predicting the next hours stock price
- etc.

Time series data consists of 4 main components:

- trend
 - Long term direction the time series data is going in, for example of an upward trend see figure ??.
- seasonal
 - A pattern that occurs within the year, year over year. Usually, it is monthly or quarterly, see figure ??.
- cyclical
 - Like seasonal, cyclical is categorized by a pattern that occurs but over a much longer period of time and may vary in length, see figure ??.
- irregular

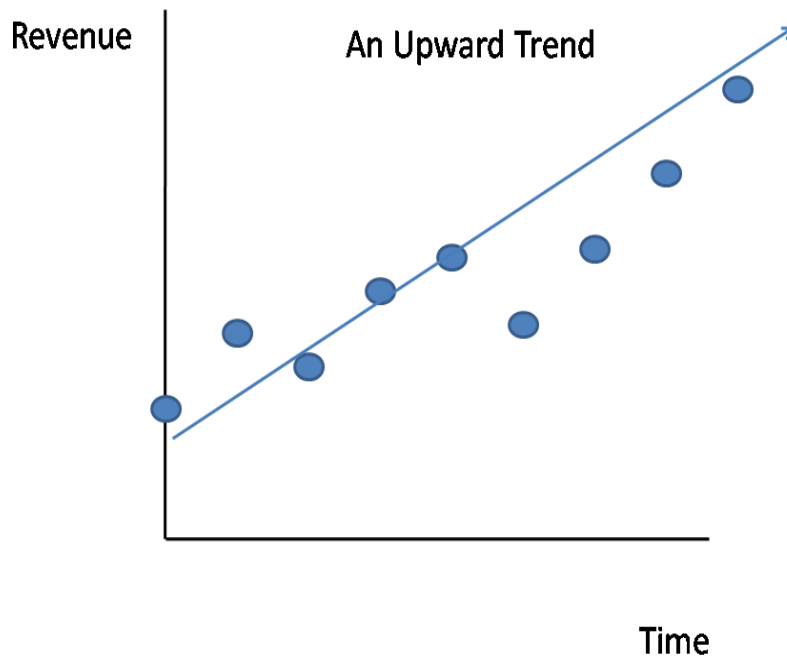


Figure 7.4: An upward trend

- Unpredictable and random, like the error term in a general linear model.

There are various techniques for analyzing time series data. This section covers only one way of analyzing time series data. This section focuses on regression models for analyzing time series data.

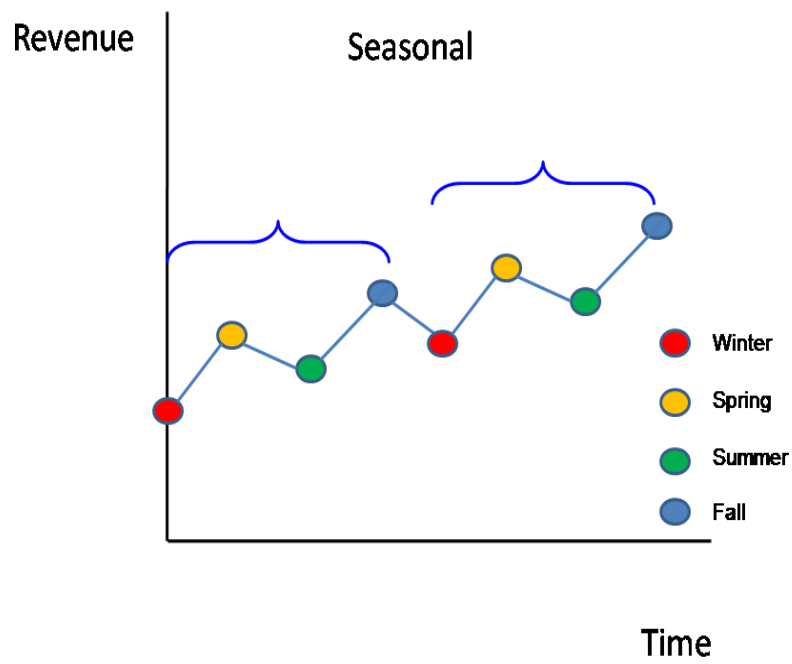


Figure 7.5: A seasonal component

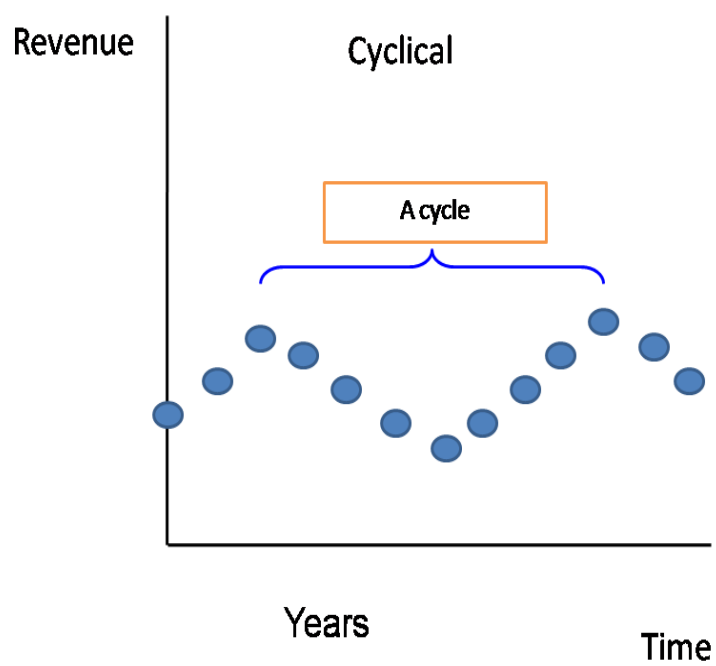


Figure 7.6: A cyclical component

7.5.2 Simple Linear Regression Applied to Time Series Data

7.5.2.1 Time Series Model with Time Period as the Dependent Variable

The techniques learned within general linear models can be applied to time series data (?). One of the simplest models within time series analysis for forecasting is simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where x_i represents time period t and may be denoted by t . A more complex model could incorporate a nonlinear relationship as well. For example, a quadratic model,

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t.$$

Time series models where the only independent variable is time, t , can be calculated in the same manner as a general linear model by simply substituting in the time period for the independent variable. A common practice is to set the first time period for the independent variable. A common practice is to set the first time period to zero. For example, looking at yearly data t would equal zero for the first year observed and $t = 1$ for the second time period, etc. Thus if the first time period observed is the year 2001, for the year 2001, t would be set to zero. This only effects the intercept in the general linear model, not the slope(s). In order to predict the dependent variable for the next time period, the value for the present time period plus one would be entered into the equation. For example, data collected from the year 1999 to the year 2007, setting the time period 1999 to zero, with the time series

Year	Time period t	Year	Time period t
1999	0	2004	5
2000	1	2005	6
2001	2	2006	7
2002	3	2007	8
2003	4	2008	9

Table 7.1: Illustrating the conversion of year into a time period starting with zero.

model $\hat{y} = 10 + 5t$, would yield an estimate of $10 + 5 * 9 = 55$ for the year 2008. See Table ?? to understand why the year 2008 is replaced by the value 9.

7.5.2.2 Time Series Model with the Previous Time Period Data as the Independent Variable

Often the past can tell us about the future. An autoregressive time series model uses historical data of the dependent variable as the independent. The assumption of an autoregressive model is that the dependent variable is in essence a function of previous data of the dependent variable plus an error term, i.e. $y_t = f(y_{t-1}, \dots, y_0) + \epsilon_t$.

A first order autoregressive model is a model which only uses the previous time period as the independent variable in the model and is written as:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t.$$

In Table ?? the previous year's EPS is used as the independent variable for understanding EPS. In the example from Table ??, the estimate for the intercept is 0.1 and the estimate for the slope is 1.0. The estimates are obtained the same way as in

Year	EPS	Previous Year's EPS	Year	EPS	Previous Year's EPS
1999	1.0	—	2004	1.5	1.4
2000	1.1	1.0	2005	1.6	1.5
2001	1.2	1.1	2006	1.7	1.6
2002	1.3	1.2	2007	1.8	1.7
2003	1.4	1.3	2008	1.9	1.8

Table 7.2: Illustrating the dependent variable in a first order autoregressive model.

simple linear regression, using the least squares method. The regression equation is

$$\hat{y}_t = 0.1 + 1.0 \times y_{t-1},$$

and thus the prediction for the EPS for the year 2009 equals $0.1 + 1.0 \times 1.9 = 2.0$, where the value of 1.9 is from the year 2008, the year prior to 2009.

A p^{th} order autoregressive model uses p previous time periods and written

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \epsilon_t.$$

In this section only the very basic time series models were covered for introductory purposes. There exist non-linear time series models and other more complicated time series models than those presented here which go beyond the scope of this text.

7.6. Examples

7.6.1 Simple Linear Regression - Calculate the Slope and Intercept - Examples

EXERCISE 7.6.1. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-12.0	3.0
2	-12.3	3.0
3	-23.0	9.0
4	-10.0	1.0

Table 7.3: The results from the experiment.

EXERCISE 7.6.2. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	7.7	9.0
2	20.2	0.0
3	16.1	4.0
4	16.6	4.0

Table 7.4: The results from the experiment.

EXERCISE 7.6.3. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-7.2	7.0
2	-2.6	6.0
3	-8.3	7.0
4	-5.9	8.0

Table 7.5: The results from the experiment.

EXERCISE 7.6.4. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-11.9	1.0
2	-2.3	3.0
3	-14.1	1.0
4	15.5	7.0

Table 7.6: The results from the experiment.

EXERCISE 7.6.5. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	16.7	3.0
2	17.7	4.0
3	21.9	7.0
4	24.6	9.0

Table 7.7: The results from the experiment.

7.7. Exercises

7.7.1 Simple Linear Regression - Calculate the Slope and Intercept - Exercises

EXERCISE 7.7.1. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-3.8	5.0
2	-5.7	2.0
3	-3.1	4.0
4	2.8	1.0

Table 7.8: The results from the experiment.

EXERCISE 7.7.2. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-36.6	9.0
2	-35.7	9.0
3	0.3	1.0
4	2.1	0.0

Table 7.9: The results from the experiment.

EXERCISE 7.7.3. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-22.1	1.0
2	-39.3	4.0
3	-40.1	4.0
4	-55.4	7.0

Table 7.10: The results from the experiment.

EXERCISE 7.7.4. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-23.9	7.0
2	-17.7	2.0
3	-22.9	4.0
4	-31.1	9.0

Table 7.11: The results from the experiment.

EXERCISE 7.7.5. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-28.8	8.0
2	-6.3	3.0
3	-17.4	5.0
4	-2.8	3.0

Table 7.12: The results from the experiment.

EXERCISE 7.7.6. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-11.9	4.0
2	-0.3	0.0
3	-15.0	7.0
4	-5.8	1.0

Table 7.13: The results from the experiment.

EXERCISE 7.7.7. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	25.3	7.0
2	18.7	4.0
3	20.5	6.0
4	30.9	9.0

Table 7.14: The results from the experiment.

EXERCISE 7.7.8. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	21.7	3.0
2	23.3	3.0
3	23.2	3.0
4	25.9	6.0

Table 7.15: The results from the experiment.

EXERCISE 7.7.9. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	-15.4	1.0
2	-14.4	4.0
3	-10.6	5.0
4	-12.2	2.0

Table 7.16: The results from the experiment.

EXERCISE 7.7.10. Calculate the slope and intercept for the simple linear regression model.

	y	x
1	11.7	8.0
2	-12.3	0.0
3	11.5	8.0
4	2.0	6.0

Table 7.17: The results from the experiment.

7.7.2 Multiple Choice

Use $\alpha = 0.05$, where appropriate, for answering the following questions. Click "Begin" and when you are finished click "End". Enjoy.

[Begin Multiple Choice Questions](#)

1. If the p-value for the model is 0.34 then the independent variables definitely are informative in terms of the dependent variable.
 - (a) True
 - (b) False
2. $\hat{y} = 10 + 4x_1 + 5x_2$ Using the latter equation the expected changed in y for 5 units change in x_1 is
 - (a) 20
 - (b) 5
 - (c) 37
 - (d) 50
3. $\hat{y} = 10 + 4x_1 + 5x_2$ Using the latter equation the expected changed in y for 10 units change in x_2 is
 - (a) 20
 - (b) 5
 - (c) 37
 - (d) 50

4. $\hat{y} = 10 + 4x_1 + 5x_2$ Using the latter equation the expected changed in y for 1 unit change in x_2 is
- (a) 20
 - (b) 5
 - (c) 37
 - (d) 50
5. $\hat{y} = 10 + 4x_1 + 5x_2$ Using the latter equation the expected value of y if $x_1 = 3$ and $x_2 = 3$ is
- (a) 20
 - (b) 5
 - (c) 37
 - (d) 50

End Multiple Choice Questions

7.8. Assignment

Your client is a the marketing division within wireless phone operator (pretend DTAC). They wish to understand who their present customers are and they wish to increase market share. The way in which they wish to increase market share is by targeting potentially high phone usage customers. They are open to additional suggestions. You will be given the client database and a prospect database. The prospect database is the list of potential customers, but are not customers yet. Give

them a marketing strategy and discuss who they should market to and how you came to your results. This is to be a mail out campaign.

This will be a group project. Groups will be of size approximately 3-4 people. Groups are to work separately. The data can be found at a link on the author's websites www.adryver.com/mybook or www.learnviaweb.com/mybook.

Note: This project has two purposes:

1. Increase your knowledge of general linear models.
 - This is the main purpose.
 - Within the presentation you must explain the performance of your model, how good or bad, in a manner that a non-statistician would understand.
 - This is a very difficult task, but this will increase your understanding greatly.
 - This will also let me know the level of understanding you have of what you are doing.
2. Give you an example of a targeted marketing strategy

The Deliverable:

- You will present the project when it is due.
 - The presentation should be done in PowerPoint.
 - Graphs, etc. must be done in Microsoft Excel.

- You will pretend you are from a consulting firm presenting to a client with very little to no statistical background.
- Finally, you will hand in the PowerPoint presentation and all supporting materials.
 - The appendix should contain the statistical procedures used and the output.
- You will be graded mostly on the PowerPoint Presentation, but not solely.
 - The presentation should be made as a stand alone document.
 - It should be made in a manner that I can read it without you needing to present for me to understand.

The Client Data File Contains

- Name
- Gender (male=1)
- Home owner or not (own=1)
- Cell phone plan type there are 5 plan types. The minimum minutes charged for:
 1. no minutes charge 4.0 baht/minute (No minimum payment)
 2. 200 minutes charge 3.5 baht/minute (Pay at least 700 baht)
 3. 400 minutes charge 3.0 baht/minute (Pay at least 1200 baht)

4. 600 minutes charge 2.5 baht/minute (Pay at least 1500 baht)
 5. 800 minutes charge 2.0 baht/minute (Pay at least 1600 baht)
- Customer income
 - Government job or not (if work in government=1)
 - Age category (age categories go youngest to oldest)
 - Location (place of residence ordered, to be considered an actual address)
 - Minutes (the total number of minutes used for the most recent month)
 - Payment history. You are given the past 12 months on whether the customer paid on time or was late and how late.
 1. 0=current, not late on payment
 2. 1=30 days late
 3. 2=60 days late
 4. 3=90 days late
 5. 4= in default, and are not expected to pay, very bad

Note: The Prospect Data Does Not Have As Much Information As Your Client Database.

8

Supplemental Material

8.1. What To Do When

This appendix material is to help the reader determine what test to perform when. It is only to be used as a rough guideline and is accurate most of the time but not always. In particular, the case of a small sample this section could be misleading. This section will aid greatly in understanding why and when a statistician may perform certain statistical tests. It is the opinion of the author that often it is important to consult a statistician and that most if not all introductory courses are not enough for a researcher to lead the statistical aspects of the project.

What to do when is mainly determined by the number of variables and the type of variables you are working with. Yes, it is often this simple, for the general situation. It is necessary to check assumptions but below is a starting point.

First off, what is the number of variables and the type of variable(s):

1. One variable: categorical
2. One variable: continuous
3. Two variables: categorical and categorical
4. Two variables: categorical and continuous
5. Two variables: continuous and continuous
6. Multiple variables: Continuous dependent
7. Multiple variables: Categorical dependent (beyond the scope of this text)
8. Exception: Time series data - use time series techniques

The next sections in the appendix cover the above in more detail.

8.2. One Variable

1. Categorical
 - (a) Two Categories
 - Binomial Test or Z approximation for test of proportions
 - (b) More than two categories
 - Chi-square test

2. Continuous

- t-test

8.3. Two Variables

1. Categorical and Categorical

- (a) Two Categories For Each Variable

- i. If one or two sided test: Two Sample Z-test of proportions
 - ii. If two sided test: Chi-Square test

- (b) More than two categories for at least one of the variables.

- Chi-square test

2. Categorical and Continuous

- (a) Two Categories

- Two sample t-test
- Paired t-test (if data is paired)

- (b) More than two Categories

- ANOVA

3. Continuous and Continuous

- Correlation and simple linear regression

8.4. Multiple Variables

1. Multiple variables: Continuous dependent
 - General linear model
2. Multiple variables: Categorical dependent (beyond the scope of this text)
 - (a) Two Categories
 - Logistic regression
 - (b) More than two Categories
 - Discriminant analysis

8.5. Time Series Data

- Time series data - use time series techniques

Tables

8.5.1 The Cumulative Standardized Normal Distribution

The area under the cumulative standardized normal distribution from $-\infty$ to Z

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

8.5.2 Critical Values of t

For a given d.f., entry represents the critical value of t corresponding to a specified upper-tail area (α)

<i>d.f./α</i>	0.25	0.10	0.05	0.025	0.01	0.005
1.00	1.00	3.08	6.31	12.71	31.82	63.66
2.00	0.82	1.89	2.92	4.30	6.96	9.92
3.00	0.76	1.64	2.35	3.18	4.54	5.84
4.00	0.74	1.53	2.13	2.78	3.75	4.60
5.00	0.73	1.48	2.02	2.57	3.36	4.03
6.00	0.72	1.44	1.94	2.45	3.14	3.71
7.00	0.71	1.41	1.89	2.36	3.00	3.50
8.00	0.71	1.40	1.86	2.31	2.90	3.36
9.00	0.70	1.38	1.83	2.26	2.82	3.25
10.00	0.70	1.37	1.81	2.23	2.76	3.17
11.00	0.70	1.36	1.80	2.20	2.72	3.11
12.00	0.70	1.36	1.78	2.18	2.68	3.05
13.00	0.69	1.35	1.77	2.16	2.65	3.01
14.00	0.69	1.35	1.76	2.14	2.62	2.98
15.00	0.69	1.34	1.75	2.13	2.60	2.95
16.00	0.69	1.34	1.75	2.12	2.58	2.92
17.00	0.69	1.33	1.74	2.11	2.57	2.90
18.00	0.69	1.33	1.73	2.10	2.55	2.88
19.00	0.69	1.33	1.73	2.09	2.54	2.86
20.00	0.69	1.33	1.72	2.09	2.53	2.85
21.00	0.69	1.32	1.72	2.08	2.52	2.83
22.00	0.69	1.32	1.72	2.07	2.51	2.82
23.00	0.69	1.32	1.71	2.07	2.50	2.81
24.00	0.68	1.32	1.71	2.06	2.49	2.80
25.00	0.68	1.32	1.71	2.06	2.49	2.79
26.00	0.68	1.31	1.71	2.06	2.48	2.78
27.00	0.68	1.31	1.70	2.05	2.47	2.77
28.00	0.68	1.31	1.70	2.05	2.47	2.76
29.00	0.68	1.31	1.70	2.05	2.46	2.76
30.00	0.68	1.31	1.70	2.04	2.46	2.75
∞	0.67	1.28	1.64	1.96	2.33	2.58

8.5.3 Critical Values of χ^2

For a specified degrees of freedom ($d.f.$), the critical value of χ^2 corresponding to a specified upper-tail area (α)

	UPPER-TAIL AREAS							
$d.f./\alpha$	0.990	0.975	0.950	0.9000	0.100	0.050	0.025	0.010
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000

8.5.4 Critical Values of F-distribution

For a given numerator d.f. (across) and denominator d.f. (down), entry represents the critical value of F equals the 95% probability less than the value given. (α) Note: $F(A; df1, df2) = \frac{1}{F((1-A); df2, df1)}$ where $P[F(df1, df2) \leq F(A; df1, df2)] = A$.

Den. df / Num. df	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

About the Author

CONTACT INFORMATION:

- Address: Graduate School of Business Administration, NIDA, 118 Seri Thai Road, BKK Thailand 10240
- Email: dryver@gmail.com
- Url: www.learnviaweb.com

EDUCATION:

- **Ph.D. in Statistics** (August, 1999)
 - *The Pennsylvania State University*, State College, PA, USA
 - * Dissertation Topic: Adaptive Sampling
 - * Advisor: Steven K. Thompson, Ph.D.
- **B.A. in Mathematical Sciences/Statistics** (May, 1993)
 - *Rice University*, Houston, TX, USA

EMPLOYMENT HISTORY:

- Assistant Professor (Apr., 2006 to Present) and Lecturer (Oct., 2003 to Mar., 2006)
 - **National Institute of Development Administration**, BKK, TH
- Statistical Analyst Consultant (Jan., 2002 to Sept., 2003)
 - **Scorex an Experian Company**, CA, USA
- Project Manager (Dec., 2000 to Jan., 2002)
 - **AnaBus Inc.**, PA, USA
- Statistical Consultant (Aug., 1999 to Dec., 2000)
 - **PricewaterhouseCoopers**, DC, USA
- Instructor, Research Assistant, and Teaching Assistant (Aug., 1993 to Aug., 1999)
 - **The Pennsylvania State University**, PA, USA

CONSULTING EXPERIENCE:

- *Strategic Consulting - Sample Projects*
 - Data Quality Studies - Investigated client data for its accuracy and usefulness
 - Fraud Detection Models - Built statistical models in order to rank individuals applying for credit in terms of likelihood to commit fraud

- Optimal Sample Allocation - Cut sampling costs without decreasing precision
 - Retail Equipment Comparison - Comparison of retail checkout counter equipment and design in terms of efficiency
 - E-Commerce (Return on Investment) - Compared different on-line advertising tools in terms of revenue generation
 - Targeted Marketing - Created a target list of individuals, who were expected to be the most profitable to acquire as future clients from a larger list of potential clients
- *Strategic Consulting - Duties*
 - Design: Discussed and helped put together the design of the project
 - Analysis: Multiple linear, logistic, and piecewise regression, decision tree, ANOVA, ANCOVA, simulation, Neyman allocation, post stratification, ...
 - Presentation: Helped make the presentations and presented to senior management
 - *Process Improvement*
 - Benchmark model is a model created in order to obtain an estimate of expected performance should a final model be developed. Improved and coded the sampling, variable selection, and performance chart steps used in the benchmark model process

- Developed programs that create hundreds of statistical reports in HTML to be viewed in Internet Explorer with hyperlinks. Saved numerous labor hours while producing more elegant reports for our clients
- Developed Excel macros written in visual basic. Used macros to facilitate the importing and formatting of multiple text files into excel spreadsheets. Also used macros to create over a hundred formatted spreadsheets for each project with hyperlinks

SELECTED COURSES TAUGHT:

- National Institute of Development Administration
 - Data Mining
 - Quantitative Analysis for Business Decisions
 - Quantitative Research Methodology I
 - Quantitative Research Methodology II
 - Regression
 - Sampling Techniques
 - Sampling Theory
 - Statistical Methods for Population and Development Research I
 - Statistical Quality Control
 - Theory of Multivariate Statistics
- Dhurakij Pundit University International College

- Advanced Statistics and Business Modeling
- Thammasat University, Chulalongkorn University, and NIDA
 - Part of an intensive course on statistics for the JDBA
- The Pennsylvania State University
 - Elementary Statistics

RESEARCH GRANTS:

- Head of research grants type 3 at NIDA, titled:
 - An In-Depth Look at Validating Logistic Regression Models in Relation to Credit Scoring
 - Ratio Estimators in Adaptive Cluster Sampling

SELECTED PUBLICATIONS:

- Dryver, A.L. and Sukkasem, J. (In Press). Validating Risk Models With a Focus on Credit Scoring Models. *Journal of Statistical Computation and Simulation*
- Dryver, A.L. and Chao, C.T. (2007). Ratio estimators in adaptive cluster sampling. *Environmetrics* **18**(6) 607-620
- Dryver, A.L. and Thompson, S.K. (2007). Adaptive sampling without replacement of clusters. *Statistical Methodology* **4** 35-43

- Dryver, A.L. and Thompson, S.K. (2005). Improved unbiased estimators in adaptive cluster sampling. *Journal of Royal Statistical Society B* **67**(1), 157-166
- Dryver, A.L. (2003) Performance of adaptive cluster sampling estimators in a multivariate setting. *Environmental and Ecological Statistics* **10**(1), 107-113

INVITED PRESENTATIONS:

- Improving ratio estimators in adaptive cluster sampling using the Rao-Blackwell theorem
 - Survey Research Methodology Conference, 2006
 - Center for Survey Research, Academia Sinica, Taiwan
 - Support provided by the Center for Survey Research Taipei, Taiwan.
- Data quality and preparation for model building
 - National Conference of Applied Statistics, 2006
 - National Institute of Development Administration, Thailand
- Ratio estimators in adaptive cluster sampling
 - Statistics in the Technological Age, 2005
 - University of Malaya, Malaysia
- Building a fraud detection model
 - Applied Statistics Seminar: Data Mining and its Applications, 2005

- National Institute of Development Administration, Thailand
- A more efficient estimator in adaptive cluster sampling than the standard Hansen-Hurwitz type estimator
 - Future of Statistical Theory, Practice and Education, 2004
 - Indian School of Business, India

AWARDS:

- Papers awarded for outstanding research by National Institute of Development Administration:
 - Improved unbiased estimators in adaptive cluster sampling
 - Ratio estimators in adaptive cluster sampling
 - Validating Risk Models With a Focus on Credit Scoring Models

COMPUTER SKILLS:

- Environments
 - Linux, Mainframe, Unix, and Windows
- Statistical Software and Programming Languages
 - Basic, C++, Excel VBA, Fortran, HTML, JAVA, JCL, JSP, LaTeX, LINDO, MATLAB, Minitab, PAJEK, PASCAL, R, SAS, SIMAN, S-Plus, and SPSS