

Descriptive Statistics

The Basics for Biostatistics: Volume I

Mohammed A. Shayib

bookboon
The eBook company

MOHAMMED A. SHAYIB

**DESCRIPTIVE
STATISTICS – THE BASICS
FOR BIOSTATISTICS
VOLUME I**

Descriptive Statistics – The Basics for Biostatistics: Volume I

1st edition

© 2018 Mohammed A. Shayib & bookboon.com

ISBN 978-87-403-2125-8

Peer review by Prof. Heather Gamber Ph.D., Lone Star College, USA

CONTENTS

	Preface	8
1	Describing Data Graphically	10
1.1	Introduction	10
1.2	Data Types and Collection	13
1.3	Descriptive Statistics	14
1.4	Frequency Distributions	15
1.5	Graphical Presentation	19
1.6	Summation Notation	29
1.7	Methods of Counting	32
2	Organizing And Summarizing Quantitative Data	50
2.1	Introduction	50
2.2	Numerical Methods for Summarizing Quantitative Data	50
2.3	Some Properties of the Numerical Measures of Quantitative Data	63
2.4	Measures of Position for Quantitative Data	65
2.5	Description of Grouped Data	70

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16

I was a construction
supervisor in
the North Sea
advising and
helping foremen
solve problems

Real work
International opportunities
Three work placements



 **MAERSK**

3	Probability	84
3.1	Introduction	84
3.2	Probability Axioms	86
3.3	Operations and Probability calculation on Events	89
3.4	Bayes Formula and Total Probability	96
4	Discrete Probability Distributions	106
4.1	Introduction	106
4.2	The Expectation and Variance of a Random Variable	108
4.3	Discrete Probability Distributions	113
4.4	More Discrete Random Variables	124
4.5	Bivariate Random Variables	129
5	Continuous Probability Distributions	147
5.1	Introduction	147
5.2	Continuous Probability Distributions	149
5.3	Normal Approximation to Discrete Distributions	172
5.4	More Continuous Distributions	176
5.5	Bivariate Random Variables (Optional)	186
6	Sampling Distributions	204
6.1	Introduction	204
6.2	Sampling Distribution of the Sample Mean	205
6.3	Central Limit Theorem	207
6.4	Distribution of the Sample proportion	212
6.5	Sampling Distribution of the Sample Variance	214
7	Simple Linear Regression and Correlation	223
7.1	Introduction	223
7.2	Regression Models	228
7.3	Fitting a straight line (First order Model)	230
7.4	Correlation	236

Appendix A	246
Tables	246
Appendix B	267
Answers to Selected Exercises	267
Appendix C	274
Formulas	274
Appendix D	293
References	293

**This book is dedicated to my Grandchildren
Mariana, Sophia, and Yousef**

PREFACE

This manuscript on **Descriptive Statistics – The Basics for Biostatistics**, has been written to provide the necessary background to the Professional Medical staff, whether they are, Medical Doctors, Nurses, Assistants or Medical Quality managers. It is already encrypted on the American currency “IN GOD WE TRUST”. The second part says “EVERYTHING ELSE NEEDS DATA”. It is the 21st century and the technology age, and data is abundant in every way and field. Statistics is not an exception since it deals with data all the time. How to make sense of that information was the motive behind writing this treatise.

The material presented in: **Descriptive Statistics – The Basics for Biostatistics**, is a compendium, an introduction to the growing field of statistics. In this volume, we emphasize on the concepts, definitions and terminology. With no doubt in mind, linking the three building blocks, mentioned above, will provide any person with a strong hold on the subject of statistics. In addition to the building blocks, the material is presented in such a way that any motivated reader can follow the outlined procedures included herein. Special attention was taken in order to provide a clear methodology for applying, and interpreting the results once the analysis had been carried on.

The material had been presented in such a way that only a course in College Algebra can be a prerequisite. The material is presented in **TWO** major volumes that complete each other.

The two volumes are:

Descriptive Statistics – The Basics for Biostatistics

Inferential Statistics – The Basics for Biostatistics

Without any doubt the two volumes form one body for understanding **The Basics for Biostatistics**. Within each part, the presentation goes like this.

Descriptive Statistics – The Basics for Biostatistics

This volume has the first 7 chapters. This is the required material for understanding the rest of the topics in the second Volume.

Chapter 1 Describing Data Graphically After Data Collection, graphical presentation of data, descriptive statistics as frequency distributions as well as giving the levels of measurements, and graphs.

Chapter 2 Describing Data Numerically: Organizing and Summarizing Quantitative data numerically regardless of the data type as discrete or continuous.

Chapter 3 Probability This chapter introduces the notion of probability, its axioms, its rules, and applications.

Chapter 4 Discrete Probability Distributions This Chapter contains material on Random Variables and their probability distributions for discrete cases.

Chapter 5 Continuous Probability Distributions This Chapter contains material on Random Variables and their probability distributions for continuous cases.

Chapter 6 Sampling Distributions is about the sampling distributions of the sample mean, sample proportion and the central limit Theorem.

Chapter 7 Simple Linear Regression and Correlation, It is restricted to the simple linear regression between an explanatory variable and a response variable related to it. Moreover, the correlation concept and definition are introduced in order to measure the strength of that linear relationship which was found earlier. The last two sections of this Chapter will be covered in Part II.

The Treatise is wrapped up with **3 Appendices** that have **10 Tables** for use when reading the textbook, **Answers for Selected Exercises**, and **Formulas** for quick reference.

The two volumes: Descriptive Statistics – The Basics for Biostatistic, and Inferential Statistics – The Basics for Biostatistic, form one manuscript for understanding The Basics for Biostatistics

Acknowledgements the Author likes to exepress his gratitude and being thankful to Dr. Heather Gamber, Professor, Department of Mathematics, Lone Star College Cy-Fair, for her help and efforts in reviewing the manuscript, and for her remarks that will make this material more appealing to the readers in the medical field.

The Author is grateful and heartfelt thankful for Ms. Jakobsen and her Staff at bookboon.com for taking time and putting the efforts for this book project to be done.

Mohammed A. Shayib
Monday, 7/10/2017

1 DESCRIBING DATA GRAPHICALLY

1.1 INTRODUCTION

Probability and statistics act as inseparable twin. Regardless, whether you had done your experiment, or planning to carry it, the following question is always hanging there: What is the chance for success?

The subject of statistics may be presented at various levels of mathematical difficult. It may be directed toward applications in various fields of inquiry. In the present manuscript, it is assumed that the reader is familiar with the basic calculus courses that include differentiation and integration. In addition to calculus, a basic knowledge of matrices will be helpful for some part of the material. As for any course, and probability and statistics is not an exception, the building blocks are: **Definitions**, **Concepts** and **Terminology**. Based on that the material will start with how to present and summarize data for any project, or trial, or an experiment. Thus, some definitions will be given first.

Definition 1.1 Statistics is that branch of science that deals with:

- 1) Collecting,
- 2) Organizing,
- 3) Summarizing,
- 4) Analyzing of data, and
- 5) Making inferences, or decisions and predictions, about populations based on data in samples.

Step 5 is the **Objective of Statistics**, i.e., making inferences about a population based on information contained in a representative sample taken from that population.

Definition 1.2 A population is a group, or a set of objects, or individuals, that share a certain property, and it is the entire interesting group to be studied. For our purposes, here, from most populations, we will be collecting some measurements or observations on the subjects of interest, which will be represented by numbers, most of the time. For instance, if we talk about students, the population could be all the students, and their characteristics that you can think of.

Since we cannot, and sometimes it is impossible to deal with all the population, a smaller and representative part, or a subset, of that population is considered.

Definition 1.3 A smaller, and a representative part, or a subset of that population is called a **sample**. Samples are collected from populations which are collections of individuals with certain individual items of particular interest.

Let us give some examples to explain the aforementioned concepts.

EXAMPLE 1.1 The students enrolled in any class, at an institution, form a population, since there are no more students that will have the same property.



EXAMPLE 1.2 Consider the students in a particular class, and try to choose, at random, a committee of three students. This committee is a sample of that population.



The elements in a population, or in a sample, are called observations, measurements, scores, or just data. Based on that, we have 4 **Levels of Measurements**. We define them below.

Definition 1.4 Data may be classified into the following **four levels of measurement**.

- **Nominal data** consists of names, labels, or categories, gender, major at college. There is no natural or obvious ordering of nominal data. (Such as high to low). Arithmetic cannot be carried out on nominal data.
- **Ordinal data** can be arranged in any particular order. However, no arithmetic can be done or performed on ordinal data.
- **Interval Data** are similar to ordinal data, with the extra property that subtraction may be carried out on an interval data. There is no natural zero for interval data.
- **Ratio data** are similar to interval data, with the extra property that division may be carried out on ratio data. There exists a natural zero for ratio data.

EXAMPLE 1.3 Identify which level of measurement is represented by the following data:

- a) Years covered in **American History: 1776–1876**
- b) Annual income of students in a Math/Statistics Class: \$0–\$10,000
- c) Course grades in any course: A, B, C, D, F, P, W, I
- d) Student gender: male, female

Solution: The years 1776 to 1876 represent an interval data. There is no natural zero (No “year zero”).

- a) Also, division does not make any sense in terms of years. What is $1776/1876$? So, that data is not a ratio. However, subtraction does make sense, the period covers: $1876 - 1776 = 100$ years.
- b) Student’s income represents a ratio data. Here division does make sense. That is, someone who made \$4000 this year compared to what was made last year of \$2000. Also, some student probably had no income last year, so that \$0, the natural zero, makes sense.
- c) Course-grades represent an ordinal data, since (i) they may be arranged in a particular order, and (ii) arithmetic cannot be done or performed on them. The quantity $A-B$ makes no sense.
- d) Student gender represents nominal data, since there is no natural or obvious way that the data may be ordered. Also no arithmetic can be done on students’ gender.



www.job.oticon.dk

oticon
PEOPLE FIRST



1.2 DATA TYPES AND COLLECTION

Data will be collected on individuals from the population and termed as variables, since it changes. Thus, variables are some characteristics of the individuals within the population. Variables can take on various types of values, some of them are numbers and some are categories. For example, the number of doors in a house is 10, and its area is 2975 square feet, each of which is numeric. On the other hand, this house is a single-family house, which in essence has no numerical value to it. This leads us to classify variables into two groups: **Qualitative** and **Quantitative**.

Definition 1.5 A **Qualitative, or Categorical**, variable allows listing the individuals' characteristics into categories.

EXAMPLE 1.4 The data in **EXAMPLE 1.3** above, in parts c. and d. are qualitative data.



Definition 1.6 A **Quantitative** variable is a variable that takes numerical measures upon which arithmetic operations can be carried out on the characteristics of the individuals. With no doubt, arithmetic operations can be carried out on quantitative variables, and thus providing meaningful results.

EXAMPLE 1.5 The data in Example 1.3, above, in parts a. and b. are quantitative data.



In addition to the above classification of data as qualitative or quantitative; quantitative variable data can be further classified as: **Discrete or Continuous**.

Definition 1.7 A **Discrete Variable** is a quantitative variable that will assume a finite, or a countable, set of values. A discrete variable cannot take on every possible value in an interval on the real line. Each value can be plotted as a separate point on the real line, with space between any two consecutive points.

EXAMPLE 1.6

- a. The number of children that a family can have is a discrete variable.
- b. The number of friends a student, in college, can have is a discrete variable.



Definition 1.8 A **Continuous Variable** is a quantitative variable that has an uncountable number of values. In other words, a continuous variable can assume any value between any two points on the real line, and thus the possible values of a continuous variable can form an interval on the number line, with no spaces between the points.

EXAMPLE 1.7 The grade point average, GPA, of any student can take an infinite number of possible values, for example in the interval 0.0 to 4.0, hence GPA is a continuous variable.



In order to make the best decision, and get the most information from our data, certain measures for that purpose are needed. Thus, we can think of Statistics as two branches:

1. **The Descriptive Statistics, and**
2. **The Inferential Statistics.**

We will handle the concept of descriptive statistics in the next section, while leaving the inferential statistics for a later section.

1.3 DESCRIPTIVE STATISTICS

Once we have identified our population, and collected the sample data, our goal is to describe the characteristics of the sample in an accurate and unambiguous fashion in such a way that the information will be easily communicated to others. Describing, or just summarizing, the data can be done in two ways:

- i) **Graphically or**
- ii) **Numerically.**

Graphical description of the data depends on the data type. As we know, there are two types of data: Qualitative and Quantitative data. The graphs for describing a **Qualitative Data** include:

- i) **The Bar Graph,**
- ii) **The Pie Chart, and**
- iii) **The Pareto Chart.**

For describing a **Quantitative Data** graphically, we use:

- i) **The Dot Plot,**
- ii) **The Stem-and-Leaf Display, and**
- iii) **The Histogram.**

In the subsequent sections, we will discuss all those methods of presenting the data graphically. Data, in some cases, will be given or presented to the researcher in a table, based on some measures chosen. The next section will deal with data when it is “summarized” in a table.

1.4 FREQUENCY DISTRIBUTIONS

When dealing with large sets of data, a good overall picture and sufficient information can be often conveyed by grouping the data into a number of classes. For instance, the weights of 125 mineral specimens collected on a field trip may be summarized as follows:

Weight (gm)	#of Specimens
0–19.9	19
20.0–39.9	38
40.0–59.9	35
60.0–79.9	17
80.0–99.9	11
100.0–119.9	3
120.0–139.9	2

Table 1

Schlumberger

WHY WAIT FOR PROGRESS?

DARE TO DISCOVER

Discovery means many different things at Schlumberger. But it's the spirit that unites every single one of us. It doesn't matter whether they join our business, engineering or technology teams, our trainees push boundaries, break new ground and deliver the exceptional. If that excites you, then we want to hear from you.

careers.slb.com/recentgraduates

Tables, like the one above, are called frequency distribution. If the data are grouped according to numerical size, as above, the resulting table is called a numerical or quantitative distribution. If the data are grouped in non-numerical categories, the resulting table is called a categorical or qualitative distribution. Frequency distributions present data in a relative compact form. They give a good overall picture, and contain information that is adequate for many purposes. Frequency distributions present raw data in a more readily usable form.

The construction of frequency distributions consists of three steps, particularly for quantitative data:

1. Choosing the classes (intervals, or categories for qualitative data)
2. Tally the data into these classes
3. Count the number of items in each class.

The first step is the most important step, while the others are purely mechanical and depend on step 1. Designing too few classes would obscure the information in the distribution while, on the other hand, designating too many classes would confuse the reader. Generally speaking, the common sense is the best guide here. Generally, there are some formulas for determining the optimal number of classes, especially for quantitative data, like the following: If the number of classes is to be k , then

$$k = \sqrt{n} \quad \text{or} \quad k = 1 + 3.3 \log n,$$

where n is the sample size, and without any doubt, k will be rounded, down or up, to a whole number.

Certain precautions need to be in place:

1. Each item will go into one and only one class,
2. The smallest and the largest values fall within the classification,
3. None of the observations can fall into gaps between successive classes,
4. Successive classes do not overlap.

Whenever possible we make the classes the same width, that is, we make them cover equal ranges of values (look up Table 1). To summarize what had been said, consider the following example.

EXAMPLE 1.8 The following are the grades of 50 students in a statistics class:

75	89	66	52	90	68	83	94	77	60
38	47	87	65	97	49	65	70	73	81
85	77	83	56	63	79	69	82	84	70
62	75	29	88	74	37	81	76	74	63
69	73	91	87	76	58	63	60	71	82

We like to construct a frequency distribution for these data. Since no grade is less than 20, and no one greater than 100, the following classes are considered, and the summary is given in **Table 2**:

Classes	Tally	Frequency	Relative Frequency (%)	Cumulative Rel. Freq (%)
20–29	/	1	2	2
30–39	//	2	4	6
40–49	//	2	4	10
50–59	///	3	6	16
60–69	//// //	12	24	40
70–79	//// //	14	28	68
80–89	//// //	12	24	92
90–99	///	4	8	100
Total		50	100%	

Table 2

The numbers in the frequency column show how many items fall into each class, and they are called the frequencies of those classes. The smallest and the largest values that can fall into any given class are called the class limits. These limits are given in column 1 under classes in Table 2. Thus, the limits for the first class are 20 and 29, and as it is clear, 20 is the lower limit while 29 is the upper limit of the first class, and so on for the other classes. The classes' boundaries for the data are: 19.5–29.5, 29.5–39.5, ..., 89.5–99.5. These boundaries are carried to one more decimal place than the data is presented by in order not to have any point on the boundary between any two classes. The data dictate how many decimal places are needed to have separate classes, and no data point is shared between two classes.

Numerical distributions also have what we call class marks and class intervals. Class marks are simply the midpoints of the classes, and the class interval is the width, of the class. For our data above the class marks are: $(20+29)/2 = 24.5$, $(30 + 39)/2 = 34.5$, $(90 + 99)/2 = 94.5$, and the class widths are: $29.5 - 19.5 = 10$, $39.5 - 29.5 = 10$, and $99.5 - 89.5 = 10$, which is the same for all the classes. It is to be noted that the class interval is not given by the difference between the class limits but rather by the difference between the class boundaries.

There are two ways in which frequency distributions can be modified to suit particular needs. One way is to convert a distribution into a percentage distribution by dividing the frequency in each class by the total number of observations, and express it as a percentage, see column 4 in Table 2. This column has what we call the relative frequency. The other way of modifying a frequency distribution is by presenting it as a cumulative relative frequency distribution by adding the relative frequencies as we go down the classes, and this will generate column 5 in Table 2, (check the **Ogive** line, see **Figure 5C**).



PREPARE FOR A LEADING ROLE.

English-taught MSc programmes in engineering: Aeronautical, Biomedical, Electronics, Mechanical, Communication systems and Transport systems. No tuition fees.

→ liu.se/master

li.u LINKÖPING UNIVERSITY



1.5 GRAPHICAL PRESENTATION

1.5.1 GRAPHICAL PRESENTATION OF QUANTITATIVE DATA

The most common form of graphical presentation of quantitative data is the histogram. An example of which is shown in **Figure 1B**, or **Figure 2**.

There are some steps to be followed in order to construct a **Histogram**. The steps are:

1. Identify the minimum and maximum in the data, and know how many data points there are on hand.
2. Calculate the difference between the max and the min of the data, and call it R,
 $R = \text{max} - \text{min}$.
3. Decide on the number of classes, k, needed for your histogram, usually by one of the formulas cited above. Make sure that k is an integer you can work with. The number of classes, k, will have the following range of values: $5 < k < 20$ as its limits.
4. Calculate the class width, w, by $w = R/k$, and round it to a number that will not interfere with the data points as well as with the class limits and boundaries.
5. Choose the lower limit LL_1 of the first class in order to accommodate for the minimum.
6. Find the upper-class limit of the first class by adding w to LL_1 to make the UL_1 , and at the same time it is LL_2 , and so on until you cover all the data. Be sure there are no data points on the boundaries of two classes. Data points have to be clearly identified by one and only one class.
7. Make a Tally for the classes made above, by going over the data once.
8. Get the frequencies in each class, and check if all data points have been counted for.
9. Graph the histogram by using the horizontal axis for data classes and the vertical axis for class frequencies. The width of the classes is the same, and the height of each class will depend on the frequency in that class.

In general, a **Histogram** is a bar graph with no spaces between the bars (**Figure 1B**).

Figure 1A, below, shows the bar graph of the data in **EXAMPLE 1.8**. Bar graphs are usually for Qualitative data. The display below is for showing the excel output. You can make that a histogram by eliminating the width between classes (**Figure 1B**).

Qualitative data can be summarized by using a bar graph, a pie chart, or a Pareto Chart. The classes here have no boundaries, and no limits. They are the categories used to classify the data. Frequency, relative frequency, or cumulative relative frequency distributions can be done on qualitative data. For a complete histogram that is generated from **Figure 1A**, you can check **Figure 1B**. For the explanation of what was said here, let us look at another example below, **EXAMPLE 1.9**.

EXAMPLE 1.9 A Typical Histogram

Pulse rates, in beats per minute, were calculated for 192 students enrolled in a statistics course. The first step in creating a histogram is to create a [frequency table](#), as shown in Table 3, and the histogram below, **Figure 2**.

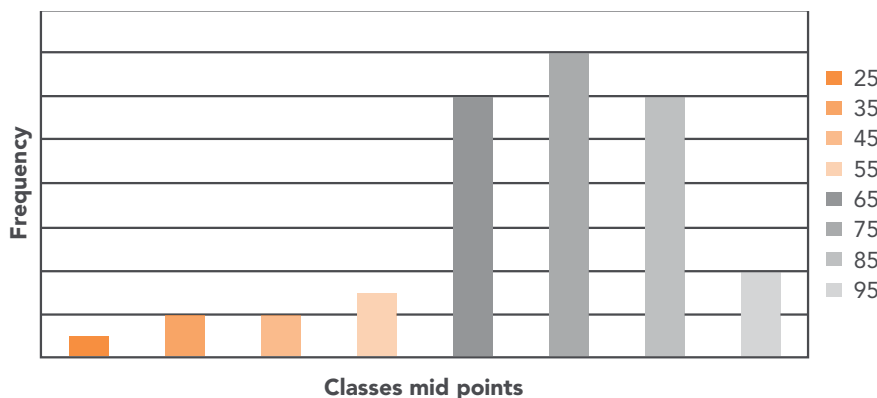


Figure 1A Bar Graph for the Data in Table 2

The broken lines joining the middle points on the tops of the bars will form what we call a polygon graph for the data. There are three cases to be taken into consideration when we look at the polygon.

1. The graph is skewed to the right,
2. The graph is skewed to the left, or
3. The graph is symmetric.

These cases will be looked at again after we present the numerical summary for the data.

In the table below, **Table 3**, the pulse rate is taken as an open – closed interval by the notation. In other words the interval $(34-41]$ stands for the range of values $34 < \text{Pulse Rate} \leq 41$, and

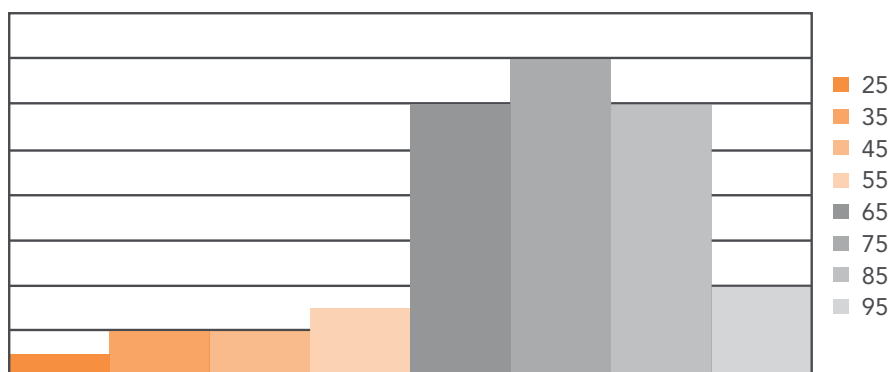


Figure 1B Histogram for data in Table 2

(41–48] is for the range of values given by $48 < \text{Pulse rate} \leq 55$, and so on.

Using the class frequencies (the number of observations in each [class interval](#)) shown in the [frequency table](#), the following histogram **Figure 2** was created.

Pulse Rate for a sample of Students

Pulse Rate	Frequency
(34–41]	2
(41–48]	2
(48–55]	4
(55–62]	19
(62–69]	40
(69–76]	53

Click here to learn more

TAKE THE
RIGHT TRACK

Give your career a head start
by studying with us. Experience the advantages
of our collaboration with major companies like
ABB, Volvo and Ericsson!

Apply by
15 January

World class
research

www.mdh.se

MÄLARDALEN UNIVERSITY
SWEDEN

Pulse Rate	Frequency
(76–83]	30
(83–90]	27
(90–97]	10
(97–104]	5

Table 3

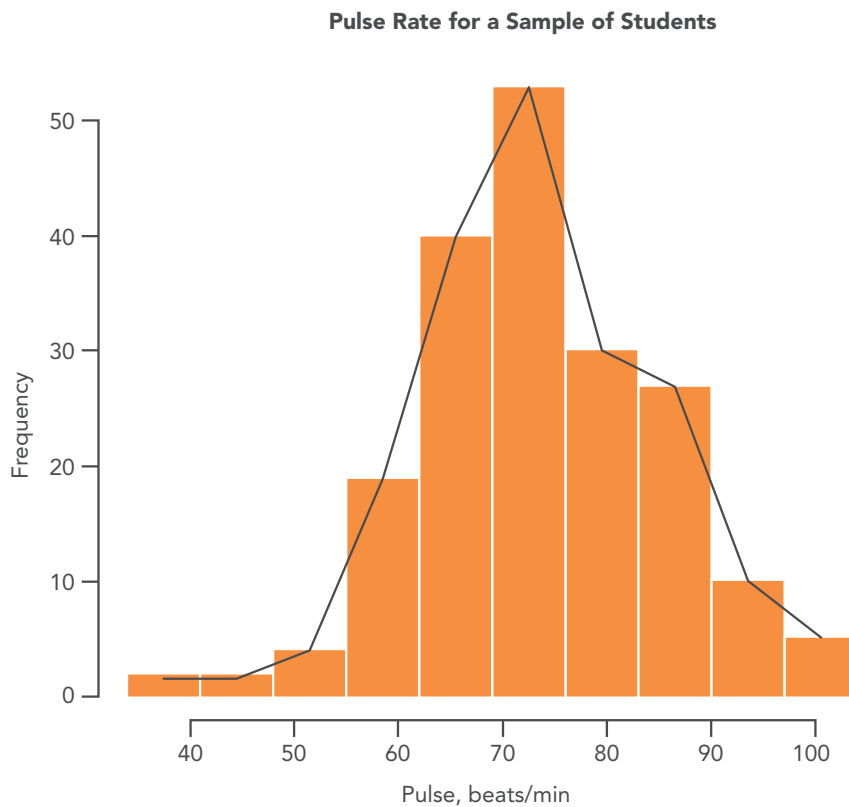


Figure 2 Histogram and Frequency Polygon for Pulse data

On any histogram, the top of each bar can be joined to the next bar top-point by a broken line. This line will represent the frequency polygon, see **Figure 2**.

In addition to the histogram for presenting the quantitative data graphically we have:

1. The Dot Plot
2. The Stem-and-Leaf.

We are not concerned much about the dot plot for quantitative data. Essentially it is the data line, drawn horizontally, with dots above the values once they are marked on the data line.

For the stem-and-leaf, it could be used as a graphical (or numerical) summary at the same time. It is highly related on how the data is presented. Is it all one digit, two digits, or 3 digits' data? This setup determines how to make a stem-and-leaf diagram. There could be one stem, or split stem for the same data.

EXAMPLE 1.10 consider the following data, and construct a Stem-and-Leaf for it.

27	17	11	24	36	13	29	22	18	17
23	30	12	46	17	32	48	11	18	23
18	32	26	24	38	24	15	13	31	22
18	21	27	20	16	15	37	19	19	29

Table 4

Solution:

Clearly, the data is of two digits, and varies between 11 and 48, and there are 40 data points. Based on this information, from the data, the stem will be the tens digit while the leaf will be the unit's digit. Thus, we the following stem-and-leaf diagram

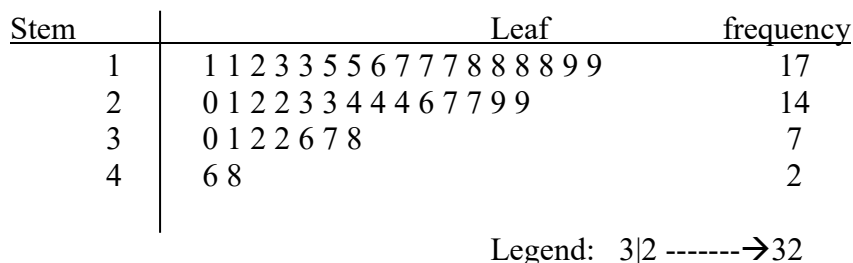


Figure 3 Stem-and-Leaf for Data in Table 4

Any stem-and-leaf diagram should be accompanied by a legend saying what the stem and leaf stand for. It is recommended to go over the data once, and put the leaf next to the stem, and then get another diagram with the leaf in order of magnitude. It is quite useful to have the frequency column, so you know you have not missed any data, see **Figure 3**.

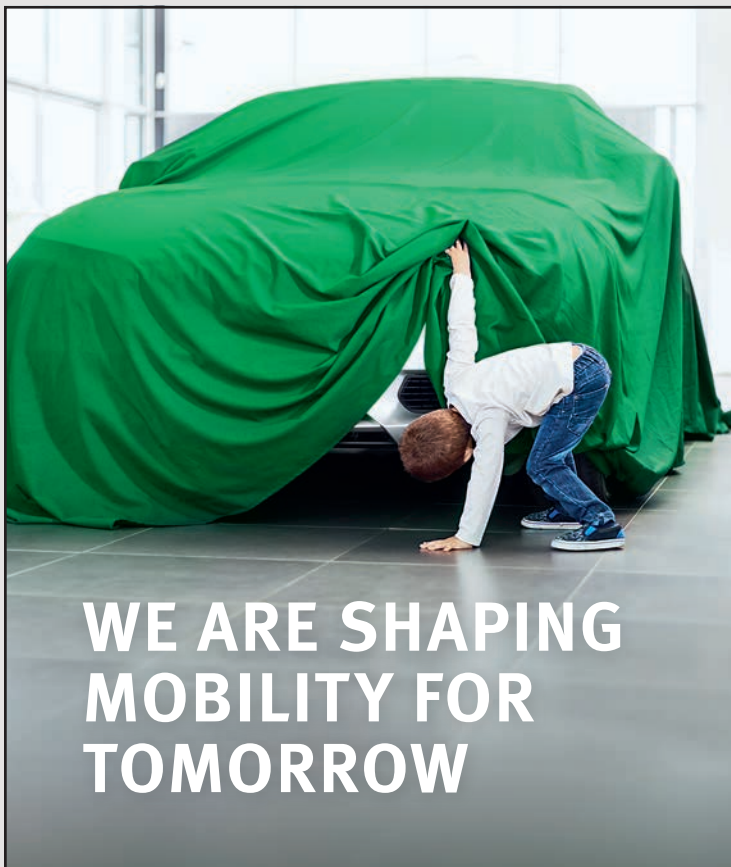
In addition to the display, if you could turn the sheet 90 degrees, counter clockwise, and draw a bar along each stem, you get yourself a histogram. It is very visible that there are too many “leaf” on one stem. In this case, we can split the stem in two parts, with the first part

for 0–4, and the second part for the digits 5–9. Doing what just had been said we have the following diagram for the stem-and-leaf for the same data presented in **EXAMPLE 1.10**.

Stem	Leaf	frequency
1	1 1 2 3 3	5
1	5 5 6 7 7 7 8 8 8 8 9 9	12
2	0 1 2 2 3 3 4 4 4	9
2	6 7 7 9 9	5
3	0 1 2 2	4
3	6 7 8	3
4		0
4	6 8	2

Legend: 3|2 -----→32

Figure 4 Stem-and-Leaf, Double stem, for data in EXAMPLE 1.10



**WE ARE SHAPING
MOBILITY FOR
TOMORROW**

How will people travel in the future, and how will goods be transported? What resources will we use, and how many will we need? The passenger and freight traffic sector is developing rapidly, and we provide the impetus for innovation and movement. We develop components and systems for internal combustion engines that operate more cleanly and more efficiently than ever before. We are also pushing forward technologies that are bringing hybrid vehicles and alternative drives into a new dimension – for private, corporate, and public use. The challenges are great. We deliver the solutions and offer challenging jobs.

www.schaeffler.com/careers

SCHAEFFLER



1.5.2 GRAPHICAL PRESENTATION OF QUALITATIVE DATA

As we have seen before, data can be qualitative or quantitative. It is to be recalled that qualitative data provide measures that categorize or classify an individual. When qualitative data is collected, we are interested in the number of items, or individuals, that fall within each category. Qualitative data can be summarized by using a

1. **Bar graph,**
2. **A pie-chart,**
3. **A Pareto Chart.**

The classes here have no boundaries. They are the categories that the data has been classified based upon. Frequency, cumulative frequency, relative frequency distributions can be drawn to present the data graphically. A graph called the **Ogive (pronounced ojive)**, a broken line joining the tops of the bars on the bar graph, which comes from the categories and their cumulative relative frequencies, can be made, see **Figure 5C**. The broken line joining the tops of the bars in a frequency distribution bar graph is called a frequency polygon. The broken line in **Figure 2**, above, is a frequency polygon for the data in **EXAMPLE 1.9**. As it was with quantitative data a frequency distribution can be constructed which will list each category of data and the number of occurrences in the category.

EXAMPLE 1.11 Consider a class in **Statistics 1350**, of 50 students. How do we classify those students based on their major in college?

Solution: No doubt those students in one course are not all have the same major, unless that class was a very specialized one. In Stat 1350, we find students majoring in Biology, Psychology, Elementary Education, Nursing, and social sciences. Those majors are categories and make the classes for our frequency distribution, as follows:

Major	Frequency	Relative frequency (%)
Biology	10	20
Psychology	15	30
Elem. Educ.	12	24
Nursing	3	6
Social Sc	10	20

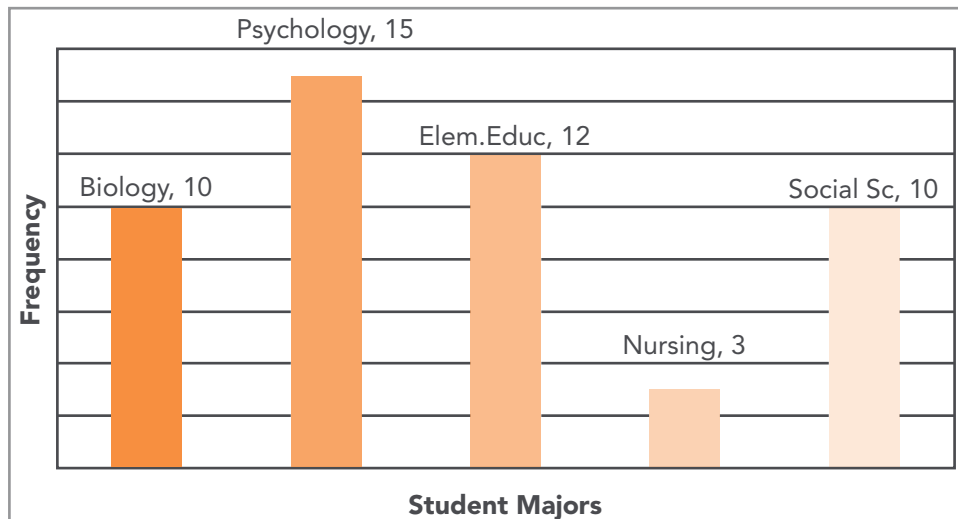


Figure 5A Bar Graph and Frequency Distribution

There is another way of describing the data in **EXAMPLE 1.11**, by depicting the relative frequency bar graph as shown in **Figure 5B**. It is to be noticed in **Figure 5B**, that you do not need to squeeze the bars to stand for percentages. The bars should be clear enough to give a complete picture and information. In addition to the two graphs in **Figure 5A** and **Figure 5B** we can display the information in a pie chart as shown in Figure 5.

Figure 6 which is displaying a **Pie Chart** for the data in **EXAMPLE 1.11**, in general, can be used when the number of categories is between 5 and 10 inclusive. Beyond those limits, fewer than 5, the graph might hide some information from the data or exploit that information with too many classes.

So far, we have displayed a bar graph, and a pie chart for qualitative data. What is that Pareto chart? **A Pareto Chart is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.** It is mainly a quality-tool for industry and business alike. It helps reduce the cost and increase the profit if used effectively.

Applying the definition of the Pareto chart on the data in **EXAMPLE 1.11**, we will have the Pareto chart shown in **Figure 7**.

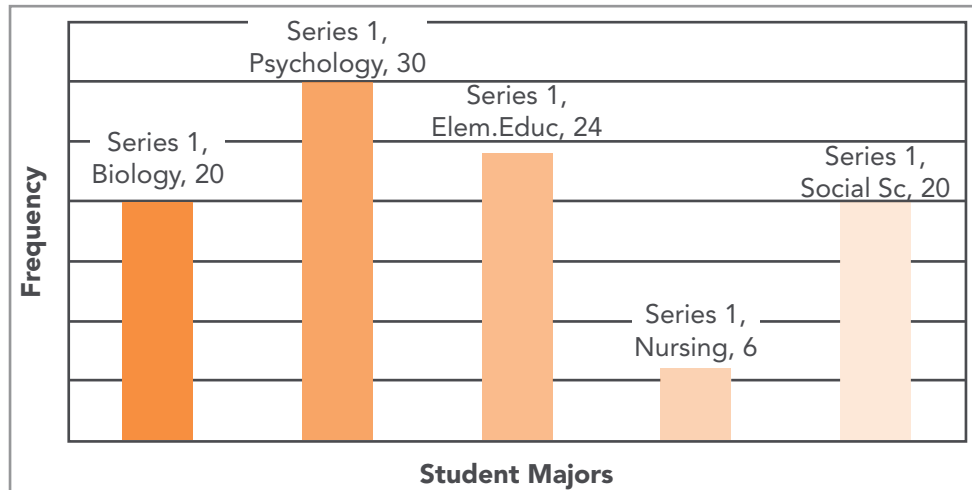


Figure 5B Bar Graph and Relative Frequency Distribution

STUDY FOR YOUR MASTER'S DEGREE IN THE CRADLE OF SWEDISH ENGINEERING

Chalmers University of Technology conducts research and education in engineering and natural sciences, architecture, technology-related mathematical sciences and nautical sciences. Behind all that Chalmers accomplishes, the aim persists for contributing to a sustainable future – both nationally and globally.

Visit us on Chalmers.se or [Next Stop Chalmers](#) on facebook.



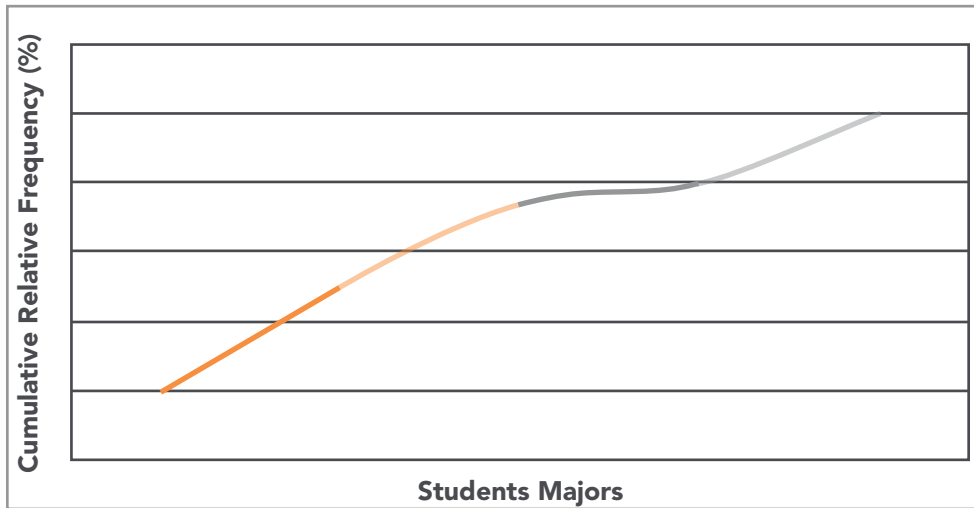


Figure 5C Cumulative Relative Frequency Distribution (Ogive)

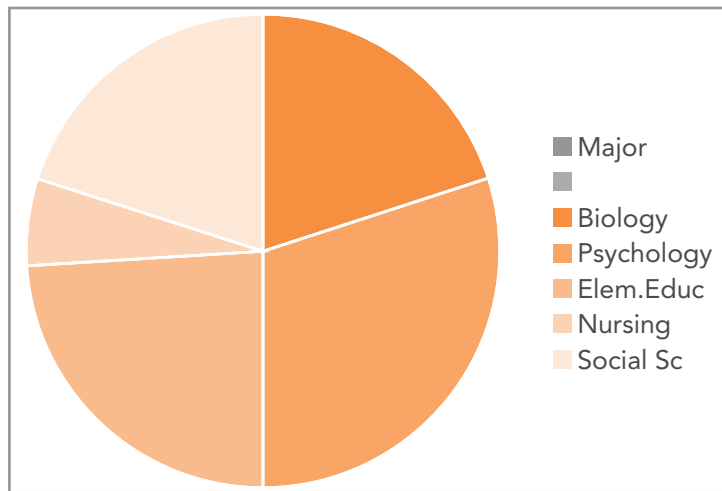


Figure 6 Pie Chart for the Majors in the Class

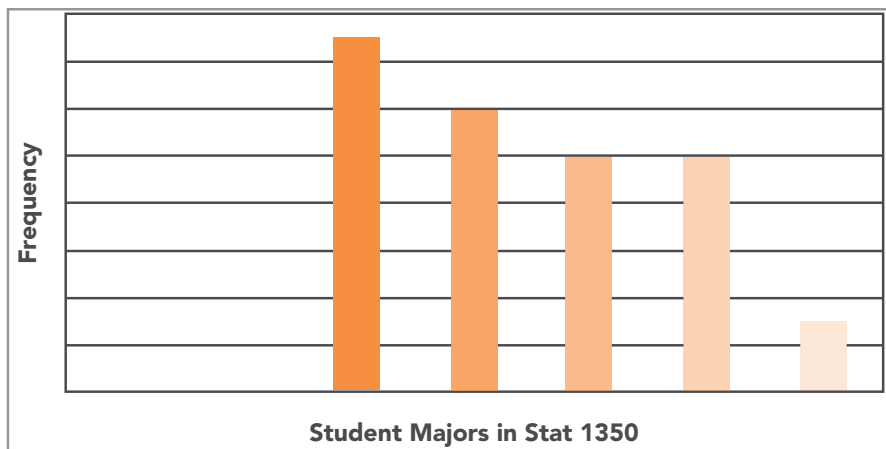


Figure 7 Pareto Chart for Enrollment by major in Stat 1350

As mentioned above, the **Pareto Chart** is a quality-tool and it can list the sources of scrap at a factory. It is a graphical tool for qualitative data. It is useful in any type of qualitative data, once the data is given in separate and non-overlapping categories, or classes. It is usually displayed starting with the class (or category) with the highest frequency, being first class, and subsequently in a descending order for the other classes. Thus, a quality engineer will know where the major source for scrap is and go for it. In another situation, a business manager can make a **Pareto Chart** on the items sold most with high demand. He can get on that and ask for more supply, and thus he will make more money then.

1.6 SUMMATION NOTATION

The Greek Capital letter, Σ Sigma, is used to indicate summation of elements in a set or a sample or a population. It is usually indexed by an index to show how many elements are to be summed. The lower case Greek letter, σ sigma, is used for a quite different number in the statistics sequel, as we will see later. Let us consider an example.

EXAMPLE 1.12 Consider the following set of numbers: 2, 5, 6, 7, 11, 15, 20, 22, and 23. Find the sum of the first 3 numbers.

Solution: This set of numbers forms an array, since they are listed in order from the smallest to the largest. To sum the first three numbers, we write

$$\sum_{i=1}^3 X_i = X_1 + X_2 + X_3 = 2 + 5 + 6 = 13$$

In reference to the expression $i = 1$, below the summation sign, is called the lower limit of the summation, and the number 3, in this case, is called the upper limit. In general, in case we like to add all the numbers in the array, the order here does not matter. We can add them in any order they are given. There is no need to arrange them in an array.

Theorems on Summation

1. If all the values in an array are equal, the sum of their values equals the number of them times that constant value.

$$\sum_{i=1}^3 C = C + C + C = 3C,$$

Proof: $\sum_{i=1}^n C = C + C + C + \dots + C$, n times, we find that $\sum_{i=1}^n C = nC$.

2. The sum of a constant times a variable (CX_i , $i = 1, 2, \dots, n$) is equal to the constant times the sum of those variables.

$$\sum_{i=1}^n CX_i = C \sum_{i=1}^n X_i,$$

Proof: since $\sum_{i=1}^n CX_i = CX_1 + CX_2 + \dots + CX_n = C(X_1 + X_2 + \dots + X_n) = C \sum_{i=1}^n X_i$.

3. The sum of a sum (or a difference) is equal to the sum (or the difference) of the sums.

$$\sum_{i=1}^n (X_i \pm Y_i) = \sum_{i=1}^n X_i \pm \sum_{i=1}^n Y_i. \text{ The proof is left to the reader.}$$

4. Some common misconceptions are considered when they should not be confused. Here are two cases:

A. $\left(\sum_{i=1}^n X_i\right)^2 \neq \sum_{i=1}^n X_i^2.$

Let us give an example.



Scholarships

Lnu.se

Open your mind to new opportunities

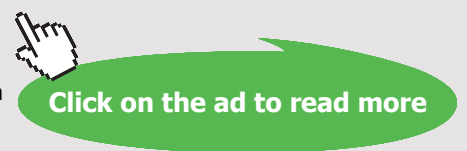
With 31,000 students, Linnaeus University is one of the larger universities in Sweden. We are a modern university, known for our strong international profile. Every year more than 1,600 international students from all over the world choose to enjoy the friendly atmosphere and active student life at Linnaeus University. Welcome to join us!

Linnæus University
Sweden

Bachelor programmes in
Business & Economics | Computer Science/IT | Design | Mathematics

Master programmes in
Business & Economics | Behavioural Sciences | Computer Science/IT | Cultural Studies & Social Sciences | Design | Mathematics | Natural Sciences | Technology & Engineering

Summer Academy courses



EXAMPLE 1.13 Consider the data in **EXAMPLE 1.12**, above, and let us compare the sum of the squares of the first three numbers to the square of the total of those three numbers. We have

$$\sum_{i=1}^3 X_i = X_1 + X_2 + X_3 = 2 + 5 + 6 = 13, \text{ and thus } \left(\sum_{i=1}^3 X_i \right)^2 = 13^2 = 169.$$

$$\text{While } \sum_{i=1}^3 X_i^2 = 2^2 + 5^2 + 6^2 = 4 + 25 + 36 = 65.$$

Clearly the above two sums are not the same.



Clearly, the order of operations makes a big difference. In $\sum_{i=1}^3 X_i^2$, squaring each number is done first and then we add those squared values, while in $\left(\sum_{i=1}^3 X_i \right)^2$, we add all the values first, and then we square their total. Quite clear that $\sum_{i=1}^3 X_i^2 \neq \left(\sum_{i=1}^3 X_i \right)^2$.

$$\text{B. } \left(\sum_{i=1}^n X_i Y_i \right) \neq \left(\sum_{i=1}^n X_i \right) \cdot \left(\sum_{i=1}^n Y_i \right).$$

Clearly the left-hand side above can be expressed as

$$\sum_{i=1}^n X_i Y_i = X_1 \cdot Y_1 + X_2 \cdot Y_2 + \dots + X_n \cdot Y_n,$$

While the right-hand side is given by

$$\left(\sum_{i=1}^n X_i \right) \cdot \left(\sum_{i=1}^n Y_i \right) = (X_1 + X_2 + \dots + X_n) \cdot (Y_1 + Y_2 + \dots + Y_n).$$

Again, let us give an example.

EXAMPLE 1.14 Consider the X array as 2, 4, 6, and 8; while the Y array to be given by 3, 5, 7, and 9.

Solution:

$$\sum_{i=1}^4 X_i Y_i = 2(3) + 4(5) + 6(7) + 8(9) = 6 + 20 + 42 + 72 = 140$$

$$\left(\sum_{i=1}^4 X_i \right) \cdot \left(\sum_{i=1}^4 Y_i \right) = (2 + 4 + 6 + 8) \cdot (3 + 5 + 7 + 9) = (20) \cdot (24) = 480.$$

No doubt, we see that $140 \neq 480$.



1.7 METHODS OF COUNTING

Counting plays an important role in many major fields, including probability. In this section, we will introduce special types of problems and develop general techniques for their solutions. We begin with The Multiplication Rule of Counting.

1.7.1 THE MULTIPLICATION RULE OF COUNTING

If a job consists of a sequence of choices to be done in which there are p selections for the first choice, q selections for the second choice, and r selections for the third choice, and so on, then the job of making these selections can be done in $p \cdot q \cdot r \cdots$, different ways.

EXAMPLE: 1.15 A three member-committee from a class of 25-students is to be randomly selected to serve as chair, vice-chair, and secretary. The first selected is the chair; the second is the Vice-chair; and the third is the secretary. How many different committee structures are possible?

Solution: There will be three selections. The first selection requires 25 choices. Because once the first student is chosen cannot be chosen again, we get left with 24 choices for the vice-chair. Similarly, we have 23 choices for a secretary. Using the Multiplication Rule we found that there are $25 \cdot 24 \cdot 23 = 13800$ different committee structures.



1.7.2 PERMUTATIONS AND COMBINATIONS

It is of interest, to the reader, to know in how many ways we can arrange three books on a shelf in the library. It is well known that books in the libraries are put in order with respect to the subject matter of that book. Any misplacement of any book outside its place will be considered as a lost copy. Thus, we see that 3 books can be put in order in 6 different ways, i.e., the first slot on the shelf can be filled out by 3 different books, the second slot by two different books, and the third by one book, the one that was left. Hence the number of ways is $3! = 3 \cdot 2 \cdot 1 = 6$. Hence, we have the following definition.

A Permutation is an arrangement of all, or part, of the objects in a set. Hence the number of permutations of n distinct objects is $n!$ (Read as n -factorial).

Definition 1.9 For a positive integer n , **$n!$ (Read n -Factorial)**, is given by

$$n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdots 3 \cdot 2 \cdot 1, \text{ with the convention that } 0! = 1.$$

In case we like to arrange r distinct objects from n distinct ones we see that the number of ways is given by:

$${}_n P_r = n! / (n - r)!, \quad 0 \leq r \leq n.$$

EXAMPLE 1.16 In how many ways can 4 boys and 5 girls sit in a row if the boys and girls must alternate?

Solution: There are $5!$ ways for the girls to sit, while there are $4!$ ways for the boys to take their seats. Thus, there are $5! \cdot 4! = 2880$ ways.



So far we considered permutations of distinct objects. That is, there are no two elements in the set that are alike. In the three books arrangement, if two of the books are the same text, denoted by a , b and c , then the arrangements of the six permutations of the letters a , b , c , where $c = b = x$, become axx , axx , xax , xax , xxa , and xxa , of which only three are distinct. Therefore, with three letters, two being the same, we have $3! / [(2!) \cdot (1!)] = 3$ distinct permutations. Hence, we have the following Rule:

e-learning for kids

- The number 1 MOOC for Primary Education
- Free Digital Learning for Children 5-12
- 15 Million Children Reached

About e-Learning for Kids Established in 2004, e-Learning for Kids is a global nonprofit foundation dedicated to fun and free learning on the Internet for children ages 5 - 12 with courses in math, science, language arts, computers, health and environmental skills. Since 2005, more than 15 million children in over 190 countries have benefitted from eLessons provided by EFK! An all-volunteer staff consists of education and e-learning experts and business professionals from around the world committed to making difference. eLearning for Kids is actively seeking funding, volunteers, sponsors and courseware developers; get involved! For more information, please visit www.e-learningforkids.org.



The number of distinct permutations of n things of which n_1 are of one kind, n_2 of a second kind, ..., n_k of a k^{th} kind, is

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!},$$

With the condition that the sum of the numbers for the different kinds equals the total on hand, i.e., $n_1 + n_2 + \dots + n_k = n$.

In many problems, we are interested in the number of ways of selecting r objects from n distinct objects without regard to order. These selections are called **Combinations**.

A Combination is actually a partition of the objects in two cells, the one cell containing the r objects and the other cell containing the $(n - r)$ objects that are left. The number of such **Combinations** is denoted, for short, by $\binom{n}{r}$, or by ${}_n C_r$, in case we like to use the notation in a TI calculator. Other notations are in use, but not frequently, such as C_n^r or C_r^n with the understanding that $0 \leq r \leq n$. The number of **Combinations** of n distinct objects taken r at a time is given by

$${}_n C_r = \binom{n}{r} = \frac{n!}{(n-r)! \cdot r!}, \quad 0 \leq r \leq n.$$

EXAMPLE 1.17 From 4 mathematicians and 3 statisticians find the number of committees that can be formed consisting of 2 mathematicians and 1 statistician.

Solution: The number of selecting 2 Mathematicians from 4 is $\binom{4}{2} = \frac{4!}{2! \cdot 2!} = 6$.

The number of ways of selecting 1 statistician from 3 is $\binom{3}{1} = \frac{3!}{1! \cdot 2!} = 3$.

Using the multiplication rule with $p = 6$ and $q = 3$, we can form $p \cdot q = (6) \cdot (3) = 18$ committees.



Understanding the Concepts Exercises CHAPTER 1

1. Define Statistics.
2. Explain the difference between a population and a sample.
3. Why sampling is used in Statistics?
4. Discuss the difference between sampling with replacement and sampling without replacement.
5. Is there a difference between an observational study and an experiment?

6. Why convenience samples are ill advised?
7. Is there a difference between cluster and stratified samplings? Explain the differences.
8. What is an open question? What is a closed question? Are they the same?
9. What is replication in an experiment?
10. What are the two main parts of statistics?
11. Why do we need to add the frequencies when we construct a frequency distribution?
12. What is a Pareto Chart?
13. What are the types of data?
14. Why shouldn't classes overlap when one summarizes continuous data?
15. What are the advantages and disadvantages of a histogram versus a stem-and-leaf plots?
16. Contrast the differences between histograms and bar graphs.
17. What is an ogive?
18. What is the value of the cumulative relative frequency for the last Class? Defend your answer.
19. What is the most common mistake when comparing histograms?
20. Are histograms preferred to pie charts? Why?

CHAPTER 1 EXERCISES

- 1.1 The following are the scores made on an intelligence test by a group of children who participated in the experiment:

114	115	113	112	113	132	130	128	122	121	126	117	115
88	113	90	89	106	104	126	127	115	116	109	108	122
123	149	140	121	137	120	138	111	100	116	101	110	137
119	115	83	109	117	118	110	108	134	118	114	142	120
119	143	133	85	117	147	102	117					

- Construct:
- i) A frequency distribution
 - ii) A relative frequency distribution
 - iii) A histogram
 - iv) A frequency Polygon

- 1.2 75 employees of a general hospital were asked to perform a certain task. The time taken to complete the task was recorded. The results (in hours) are as shown below:

1.5	1.3	1.4	1.5	1.7	1.0	1.3	1.7	1.2	1.8	1.1	1.0	1.8
1.6	2.1	2.1	2.1	2.1	2.4	2.9	2.7	2.3	2.8	2.0	2.7	2.2

2.3	2.6	2.8	2.1	2.3	2.4	2.0	2.8	2.2	2.5	2.9	2.0	2.9
2.5	3.6	3.1	3.5	3.7	3.7	3.4	3.1	3.5	3.6	3.5	3.2	3.0
3.4	3.4	3.2	4.5	4.6	4.9	4.1	4.6	4.2	4.0	4.3	4.8	4.5
5.1	5.7	5.1	5.4	5.7	6.7	6.8	6.6	6.0	6.1			

- Construct:
- i) A frequency distribution
 - ii) A relative frequency distribution
 - iii) A histogram
 - iv) A frequency Polygon

1.3 On the first day of classes, last semester, 50 students were asked for their one-way travel from home to college (to the nearest 5 minutes). The resulting data were as follows:

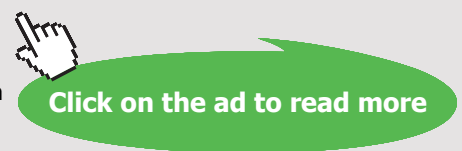
20	20	30	25	20	25	30	15	10	40	35	25	15
25	25	40	25	30	5	25	25	30	15	20	45	25
35	25	10	10	15	20	20	20	25	20	20	15	20
5	20	20	10	5	20	30	10	25	15	25		

.....Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".

Construct a stem-and-leaf display for these data.

1.4 Write each of the following in full expression; that is, without summation sign:

$$\text{a) } \sum_{i=1}^4 x_i \quad \text{b) } \sum_{i=1}^5 x_i f_i \quad \text{c) } \sum_{i=1}^{10} y_i \quad \text{d) } \sum_{i=1}^7 x_i y_i \quad \text{e) } \sum_{i=1}^6 (x_i + y_i)$$

1.5 Write each of the following by using the summation notation:

$$\begin{aligned} \text{a) } & x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4 + x_5 f_5 + x_6 f_6 \\ \text{b) } & y_1^2 + y_2^2 + y_3^2 + y_4^2 \\ \text{c) } & (z_2 + 3) + (z_3 + 3) + (z_4 + 3) + (z_5 + 3) + (z_6 + 3) \end{aligned}$$

1.6 Given: $X_1 = 2$, $X_2 = 3$, $X_3 = 4$, $X_4 = -2$; and $Y_1 = 5$, $Y_2 = -3$, $Y_3 = 2$, and $Y_4 = -1$. Find:

$$\text{a) } \sum_{i=1}^4 x_i \quad \text{b) } \sum_{i=1}^4 Y_i^2 \quad \text{c) } \sum_{i=1}^4 X_i Y_i$$

1.7 Given $X_{11} = 3$, $X_{12} = 1$, $X_{13} = -2$, $X_{14} = 2$; $X_{21} = 1$, $X_{22} = 4$, $X_{23} = -2$, $X_{24} = 5$; $X_{31} = 3$, $X_{32} = -1$, $X_{33} = 2$, and $X_{34} = 3$. Find:

$$\begin{aligned} \text{a) } & \sum_{i=1}^3 X_{ij} \text{ Separately for } j = 1, 2, 3, \text{ and } 4, \\ \text{b) } & \sum_{j=1}^4 X_{ij} \text{ Separately for } i = 1, 2, \text{ and } 3. \end{aligned}$$

1.8 A TRUE-FALSE test consists of 12 questions. In how many ways can a student mark one answer to each question?

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

$$\begin{aligned} \text{a) } 20! &= 20 \cdot 19 \cdot 18 \cdot 17! & \text{b) } 3! + 4! &= 7! & \text{c) } 4! \cdot 3! &= 12! & \text{d) } 16! &= 17! / 17 \\ \text{e) } 1/2! + 1/2! &= 1 & \text{f) } 9! / (7! \cdot 2!) &= 72 & \text{g) } 4! + 0! &= 25 \end{aligned}$$

1.10 If there are 8 horses in a race, in how many different ways can they be placed First, Second and Third?

1.11 Determine the original of the data set below. The stem represents tens digit and the leaf represents the ones digit.

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

1.12 Determine the original of the data set below. The stem represents ones digit and the leaf represents the tenths digit.

Stem	Leaf
1	2 4 6
2	4 4 7 7 9
3	3 5 7 7 8
4	1 1 3 6 6 8 9 9
5	3 4 5 8
6	2 4

1.13 Construct a stem-and-leaf diagram for the data in Exercise 1.1,

- a) By using the units as the leaf,
- b) By using a split stem.

1.14 Construct a stem-and-leaf diagram for the data in Exercise 1.2, by using a split stem and the leaf represents the on the tenths digit.

1.15 Determine the original set of data below. The stem represents ones digit and the leaf represents the tenths digit.

Stem	Leaf
12	3 7 7 9
13	0 4 5 4 7 8 9
14	2 4 4 7 7 8 9
15	1 2 2 5 6 7
16	0 3 4 5 8
17	1 2 4

Classify the variable as **Qualitative** or **Quantitative** in the following Exercises: 16–23.

1.16 Nation of origin.

1.17 Number of friends.

1.18 Eye color.

Nido

Luxurious accommodation

Central zone 1 & 2 locations

Meet hundreds of international students

BOOK NOW and get a £100 voucher from voucherexpress

Nido Student Living - London

Visit www.NidoStudentLiving.com/Bookboon for more info.

+44 (0)20 3102 1060

- 1.19 Grams of sugar in a meal.
- 1.20 Number of left turns you made while driving home today.
- 1.21 The value of your car.
- 1.22 Your phone number.
- 1.23 Your student ID on any campus.

Classify the quantitative variable in the Exercises: 24–30, as **Discrete** or **Continuous**.

- 1.24 The distance from your house to school.
- 1.25 The time to run a marathon.
- 1.26 The number of questions you will get wrong on a multiple-choice exam.
- 1.27 The number of seats in a classroom.
- 1.28 The time you take to finish a pop quiz.
- 1.29 The amount of gas in the tank of your car.
- 1.30 The number of cylinders in the engine of your car.

- 1.31. Births are not evenly distributed across the days of the week. The table below shows the average number of babies delivered on each day of the week in recent years:

Day	Number of Births
Sunday	7374
Monday	11706
Tuesday	13170
Wednesday	13150
Thursday	13014
Friday	12664
Saturday	8455

- Present the data above in a bar graph.
- Make a pie chart for the data, and justify it.
- Give a reasonable reason why there are fewer births on the weekends.

- 1.32 **Weight at Birth** (oz.) for 40 children is given below:


58 118 92 120 86 115 123 134 94 104
 132 98 121 68 107 111 121 124 91 122
 138 104 115 138 128 106 125 133 115 127
 108 118 67 146 122 104 99 105 108 135

- Present the data above in a stem-and-leaf graph.
- Make a histogram for the data, and justify it.
- Give a reasonable reason why there are fewer light weights.
- Could you guess what is the normal weight at birth for a full-term pregnancy?


- 1.33 Hypertension and Blood Pressure. The data, in the Table below, shows the two types of Blood Pressure, i.e. Systolic and Diastolic, for two different positions.

- Present a stem-and-leaf graph for each position and type of data.
- Could you guess what is the normal BP for each type, and position?

Case #	Recumbent, Arm at the side		Standing, Arm at Heart Level	
	SBP	DBP	SBP	DBP
1	99	71	105	79
2	126	74	124	76
3	108	72	102	88
4	122	68	114	72
5	104	64	96	62
6	108	60	96	56
7	116	70	106	70
8	106	74	106	76
9	118	82	120	90
10	92	58	88	60

SIMPLY CLEVER


WE WILL TURN YOUR CV INTO AN OPPORTUNITY OF A LIFETIME



Do you like cars? Would you like to be a part of a successful brand?
 As a constructor at ŠKODA AUTO you will put great things in motion. Things that will ease everyday lives of people all around Send us your CV. We will give it an entirely new new dimension.

Send us your CV on
www.employerforlife.com

Case #	Recumbent, Arm at the side			Standing, Arm at Heart Level	
	SBP	DBP		SBP	DBP
11	110	78		102	80
12	138	80		124	76
13	120	70		118	84
14	142	88		136	90
15	118	58		92	58
16	134	76		126	68
17	118	72		108	68
18	126	78		114	76
19	108	78		94	70
20	136	86		144	88
21	110	78		100	64
22	120	74		106	70
23	108	74		94	74
24	132	92		128	88
25	102	68		96	64
26	118	70		102	68
27	116	76		88	60
28	118	80		100	84
29	110	74		96	70
30	122	72		118	78
31	106	62		94	56
32	148	90		138	94

Reference: C.E. Kossman (1946), "Relative importance of certain variables in the clinical determination of blood pressure." *American Journal of Medicine*, 1, 464–467.

CHAPTER 1 TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP Obtaining a Simple Random Sample

TI-83/84 Plus

1. Enter any nonzero number (the seed) on the HOME screen.
2. Press the *sto* > button.
3. Press the *MATH* button.
4. Highlight the *PRB* menu and select 1: *rand*.
5. From the **HOME** screen press *ENTER*.
6. Press the *MATH* button. Highlight *PRB* menu and select 5: *randInt* (.
7. With *randInt* (on the **HOME** screen, enter 1, *n*, where *n* is the sample size. For Example, *n* = 500, enter the following *randInt* (1, 500)
Press *ENTER* to obtain the first individual in the sample. Continue pressing *ENTER* until the desired sample size is obtained.

Excel

1. Be sure the Data Analysis Tool Pak is activated. This is done by selecting the **TOOLS** menu and highlighting **Add-Ins...** Check the box for the **Analysis Tool Pac** and select **OK**.
2. Select **TOOLS** and highlight **Data Analysis...**
3. Fill in the windows with the appropriate values. To obtain a simple random sample for the situation on hand (see example 2, choosing a committee of 5 from a class of 30 students). When you see the excel screen you fill in the following:

Number of Variables: 1	OK
Number of Random Numbers: 10	CANCEL
Distribution: Uniform	HELP
Parameters	
Between 1 grid 31	
Random Seed 34	
Out options	
Output range	
• New Window	
New workbook	

The reason we generate 10 rows of data (instead of 5) is in case any of the random numbers repeat. Notice also that the parameter is between 1 and 31, so any value greater than

or equal 1 and less than or equal 31 is possible. In the unlikely event that 31 appeared simply ignore it. Select OK and the random numbers will appear in column 1(A1) in the spreadsheet. (Ignore any values to the right of the decimal place.)

TECHNOLOGY STEP-BY-STEP Drawing Bar Graphs and Pie Charts

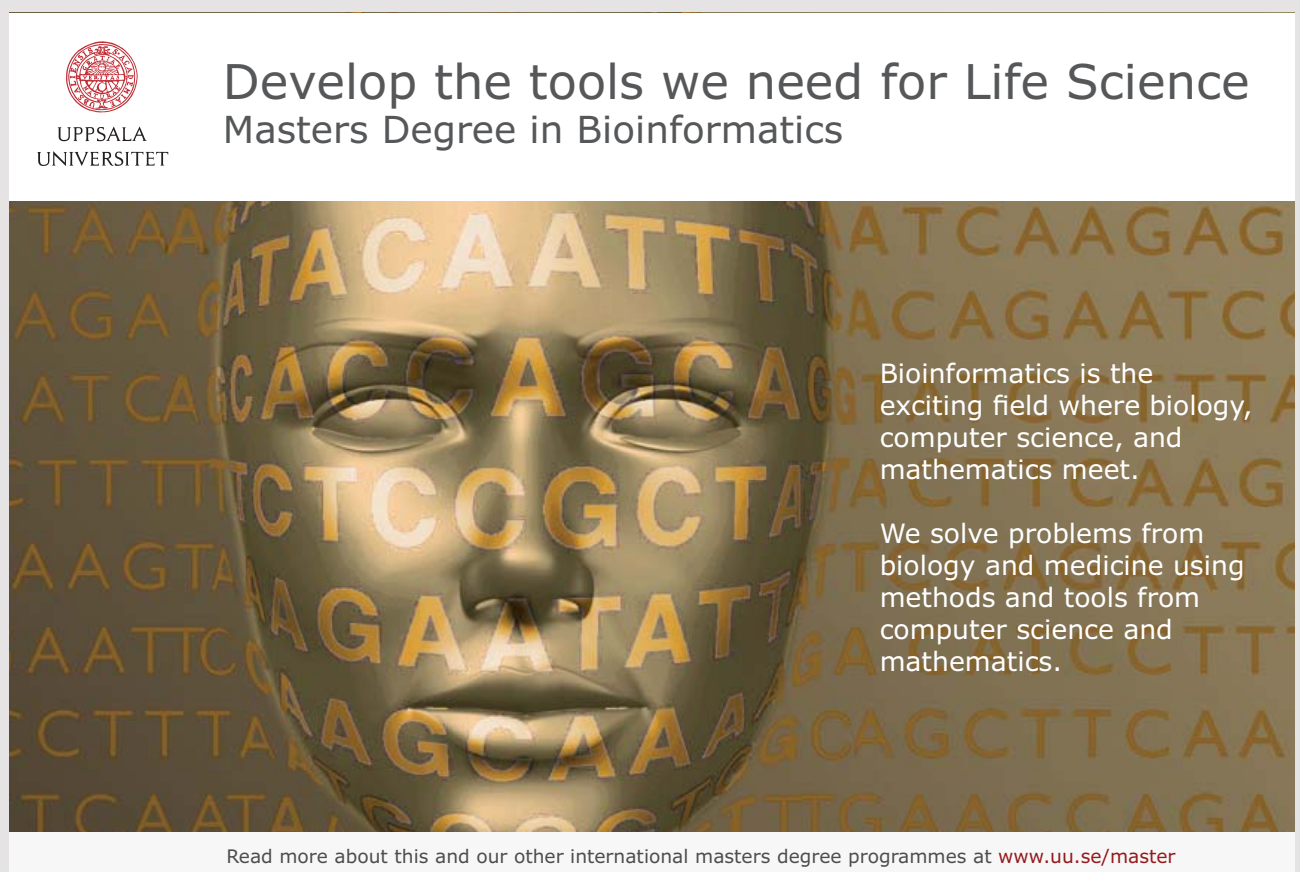
TI-83/84 Plus


The TI-83 or TI-84 does not have the ability to draw bar graphs or Pie charts.

Excel

Bar Graph from Summarized Data

1. Enter the categories in column A and the frequencies or relative frequencies in column B.
2. Select the chart wizard icon. Click the “column” chart type. Select the chart type in the upper-left corner and hit “Next”.




UPPSALA
UNIVERSITET

Develop the tools we need for Life Science Masters Degree in Bioinformatics

Bioinformatics is the exciting field where biology, computer science, and mathematics meet.

We solve problems from biology and medicine using methods and tools from computer science and mathematics.

Read more about this and our other international masters degree programmes at www.uu.se/master

3. Click inside the data range cell. Use the mouse to highlight the data to be graphed. Click "Next".
4. Click the "Titles" tab to include the x-axis, y-axis, and chart titles. Click "Finish."

Pie Charts from Summarized Data

1. Enter the categories in column A and the frequencies in column B. Select the chart wizard icon and Click the "Pie" chart type. Select the pie type in the upper-left corner and hit "Next".
2. Click inside the data range cell. Use the mouse to highlight the data to be graphed. Click "Next".
3. Click the "Titles" tab to the chart titles. Click "Data Labels", tab and select "show label and percent." Click "Finish."

TECHNOLOGY STEP-BY-STEP Drawing Histograms and Stem-and-Leaf Plots

TI-83/84 Plus

The TI-83 and TI-84 do not have the ability to draw stem-and-leaf plots or dot plots.

TI-83/84 Plus

Histograms

1. Enter the raw data in L1 by pressing **Stat** and selecting **1: Edit**.
2. Press **2nd Y =** to access Stat-Plots menu. Select **1: plot1**.
3. Place the cursor on "**ON**" and press **ENTER**.
4. Place the cursor on the histogram icon (check your calculator) and press **ENTER**. Press **2nd Quit** to exit Plot 1 menu.
5. Press **Window**. Set Xmin to the lower-class limit of the first class, or lower. Set Xmax to the upper-class limit of the last class or higher. This will take care of the min and max in the data. Set Xscal to the class width. Set Ymin to -3 (so you can read below the x-axis later). Set Ymax to a value larger than the frequency of the class with the highest frequency.
6. Press **GRAPH**.

Helpful Hints: To determine each class frequency, press **TRACE** and use the arrow keys to scroll through each class. If you decrease the value of Ymin to a value such as -3, you can see the values displayed on the screen easier.

The TI graphing calculators do not draw stem-and-leaf plots or dot Plots.

Excel

Excel does not draw stem-and-leaf plots. Dot plots can be drawn in Excel using the DDXL plug-in. See the Excel Technology manual.

Histogram

1. Enter the raw data in column A.
2. Select **TOOLS** and **Data Analysis...**
3. Select the histogram from the list.
4. With the cursor in the Input Range cell, use the mouse to highlight the raw data. Select the Chart output box and press OK.
5. Double-click on one of the bars in the histogram. **SELECT THE** Options tab from the menu that appears. Reduce the gap width to zero.

OR: You can use the following setup

Excel

Histograms

1. Load the XLSTAT Add-in.
2. Enter the raw data in column A.
3. Select XLSTAT. Click Describing data, and then select Histograms.
4. With the cursor in the Data cell, highlight the data in Column A.
5. Click either the Continuous or Discrete radio button.
6. Click the Options tab. Decide on either a certain number of intervals or enter your own lower class limits. To enter your own intervals, enter the lower-class limits in Column B.
7. Click the Charts Tab. Choose either Frequency or Relative Frequency. Click OK.

TECHNOLOGY STEP-BY-STEP Drawing Histograms, Stem-and-leaf plots, and Dot plots

Minitab

Histograms

1. Enter the raw data in C1.
2. Select the **Graph** menu and highlight

Histogram p

3. Highlight the “simple” icon and press OK.
4. Put the cursor in the “Graph variables” box. Highlight C1 and press Select. Click SCALE and select the Y-Scale Type tab. For a frequency histogram, click the frequency radio button. For a relative frequency histogram, click the percent radio button. Click OK twice.

Note: To adjust the class width and to change the labels on the horizontal axis to the lower-class limit, DoubleClick inside one of the bars in the histogram. Select the “binning” tab in the window that opens. Click the cut point button and the midpoint/cut point positions radio button. In the midpoint/cut point box, enter the lower-class limits of each class. Click OK.

Stem-and-Leaf Plots

1. With the raw data entered in C1, select the **Graph** menu and highlight **Stem-and-Leaf**.
2. Select the data in C1 and press OK.

UNIVERSITY OF COPENHAGEN



*Copenhagen
Master of Excellence*

Copenhagen Master of Excellence are two-year master degrees taught in English at one of Europe's leading universities

Come to Copenhagen - *and aspire!*

Apply now at
www.come.ku.dk



cultural studies



religious studies

science

Dot Plots

1. Enter the raw data in C1.
2. Select the **Graph** menu and highlight **Dot plot**.
3. Highlight the “simple” icon and press OK.
4. Put the cursor in the “Graph variables” box. Highlight C1 and press Select. Click OK.

Statcrunch

Histograms

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graphics** and highlight **Histogram**.
3. Click on the variable you wish to summarize and click Next>.
4. Choose the type of histogram (frequency or relative frequency) You have the option of choosing a lower class limit for the first class by entering a value in the cell marked “Start bins at:”. You have the option of choosing a class width by entering a value in the cell marked “Binwidth:” Click Next>.
5. You could select a probability function to overlay on the graph (such as Normal – see Chapter 7). Click Next>.
6. Enter labels for the x - and y -axes, and enter a title for the graph. Click Create Graph!

Steam-and-Leaf Plots

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graphics** and highlight **Stem and Leaf**.
3. Click on the variable you wish to summarize and click next>.
4. Select None for Outlier trimming. Click Create Graph!

Dot Plots

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graphics** and highlight **Dot plot**.
3. Click on the variable you wish to summarize and click Next>.
4. Enter labels for the x - and y -axes, and enter a title for the graph. Click Create Graph!

2 ORGANIZING AND SUMMARIZING QUANTITATIVE DATA

2.1 INTRODUCTION

In **Chapter 1**, we have seen the types of data that were classified as **Quantitative**, or **Numerical**, and **Qualitative**, or **Categorical**. In addition to that classification we presented some graphical methods to summarize both types of data. In this Chapter, the discussion will be on how to summarize **Quantitative Data** numerically.

2.2 NUMERICAL METHODS FOR SUMMARIZING QUANTITATIVE DATA

As Described above, data can be one of two types: **Qualitative** (Categorical) or **Quantitative (Numerical)**. In this section, we will present some measure to summarize quantitative data. Those measures include:

1. **Measures of Central Tendency (or Measures of Center)**
2. **Measures of Variation (or Measures of Dispersion)**
3. **Measures of Position,**
4. **Measures of Quality and Outliers, and**
5. **The Five Numbers Summary with the Box-Plot**

In the following subsections, we will address the aforementioned measures separately for quantitative data.

2.2.1 MEASURES OF CENTRAL TENDENCY

At first, we will deal with raw data as presented the first time, i.e., not grouped in any way. Starting with the **Measures of Central Tendency**, there are 4 measures:

1. **The Mean,**
2. **The Mode,**
3. **The Median,** and
4. **The Mid-Range.**

Let us give the definition for each of those measures.

Definition 2.1 The Mean, or the arithmetic average, of a set of numbers is computed by adding all the values in the data set and divide by the number of observations.

A point of warning is due here. Is the data set a sample or a population? To set the distinction between a population and a sample, we will reserve the number of observations in the sample to be n , while for a finite population that number will be denoted by N . Thus, there is a difference in presenting the mean of the data if it were a sample or a population. For sample data, the mean is denoted by \bar{x} (pronounced “x-bar”), while the mean of a population data is denoted by μ (pronounced “mew”). For a sample data set given by x_1, x_2, \dots, x_n , the sample mean, \bar{x} , is calculated as:

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n = \sum x_i / n.$$

In case the data set was considered as a population presented as x_1, x_2, \dots, x_N , then the population mean, μ is given by

$$\mu = (x_1 + x_2 + \dots + x_N) / N = \sum x_i / N.$$

Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

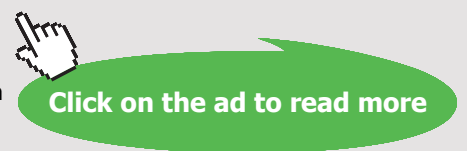
Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF



It is needless to say that we have to set a distinction between a population and a sample, when we calculate the mean. Any set of data should be considered as a sample until it is clearly specified that data is the whole population. Thus, if a set of data is consisting of all conceivably possible (or hypothetically possible) observations of a given phenomenon, we call that set a population. If the data set consists of only a part of these observations, we call that set a **Sample**.

Aside from the fact that the **Mean** or the average as it is frequently called and used, is a simple and familiar measure. The following are some of its noteworthy properties:

- i. It can be calculated for any set of data, so it always exists.
- ii. A data set of numerical values has one and only one mean. Thus, it is unique.
- iii. It lends itself to further statistical treatment, as we will see later.
- iv. It is relatively reliable in the sense that the means of many samples drawn from the same population usually do not fluctuate, or vary, as wildly as other statistical measures when used to estimate the mean of a population.
- v. It takes into account every item of the data set.
- vi. It is very sensitive to any minor change in the data.

In addition to the mean, or the arithmetic average, defined above, there are two other kinds of averages which are used in some special cases, and they are worth noting here. Those are: The Geometric mean, and the harmonic mean. Based on the data, cited above, we can find those means as follows:

Definition 2.2 The **Geometric Mean**, \bar{G} , Gbar, which is given by $\bar{G} = (X_1 \cdot X_2 \cdot X_3 \cdot \dots \cdot X_n)^{1/n}$, or $\bar{G} = \left(\prod_1^n X_i\right)^{1/n}$.

Definition 2.3: The **Harmonic Mean**, Hbar, \bar{H} , is given by $\bar{H} = n / \sum_1^n (1/X_i)$.

EXAMPLE 2.1 Consider the following set of data: 34, 15, 20, 7, 8, 9, 10, 22, 18, 30, 11, 12, and 19.

Solution: This same data can be looked at as a sample or as a population. In either case let us find the mean in each of those two cases. Clearly, we need to add all the data points in either case and divide by their number. Here $n = 13$, and $N = 13$, so we have $\bar{x} = 16.538$, and $\mu = 16.538$.

EXAMPLE 2.2 Consider the following set of data: 5, 8, 12, 15, and 20. For this data, find

- The geometric mean,
- The harmonic mean.
- Compare the above three means: \bar{x} , \bar{G} , and \bar{H} .

Solution:

- The geometric mean is given by $\bar{G} = (X_1 \cdot X_2 \cdot X_3 \cdot \dots \cdot X_n)^{1/n}$, and we have $n = 5$; and $X_1 = 5$, $X_2 = 8$, $X_3 = 12$, $X_4 = 15$, and $X_5 = 20$. Applying the formula for, \bar{G} we see that with a graphing calculator that $\bar{G} = (5 \cdot 8 \cdot 12 \cdot 15 \cdot 20)^{1/5} = (144000)^{1/5} = 10.7565$.
- The Harmonic mean is given by $\bar{H} = n / \sum (1/X_i)$. From the data, and by using a graphing calculator we find that $\bar{H} = 5 / [1/5 + 1/8 + 1/12 + 1/15 + 1/20] = 2.625$.
- For the comparison, we need to calculate the arithmetic mean \bar{x} . It is easily found that it equals to $60/5 = 12$. Therefore, we have $\bar{H} < \bar{G} < \bar{x}$.



Definition 2.4 The Mode is the most frequent data point in the sample. The mode is considered to be the least informative measure in the central tendency measures.

Generally speaking, there are two cases where the mode is useful. First if the data represents frequency counts in a non-ordered or categorical classes (e.g. Hair color, Geographical data) it should be obvious that one can count the most frequent or popular class, while the mean and median cannot be computed. For Example, what meaning would a statement like “the average washing machine is a Maytag” have? Secondly, one may also cite the mode or modes of a distribution along with the mean and median. “While most people earn less than \$50,000, the median income is \$5000. There might be more than one mode. In case there is only one, the sample will be unimodal, or bimodal when it has two modes, or tri-modal when there are 3 modes, and so on.

- EXAMPLE 2.3:**
- Find the mode for the data in **Example 2.1**.
 - Consider the following Data as presented in classes

Class	1	2	3	4	5	6
Frequency	2	7	3	4	2	4

- Solution:**
- The data in **EXAMPLE 2.1** has no mode. Each data point has appeared once.
 - The data in b) above has what we call a modal class. It is that class with the highest frequency. The modal class is class 2.



Definition 2.5: The Median is that value, in an array of numerical data which separates the array into two equal parts; i.e., 50% above the median and 50% below the median. The definition of the median implies three steps before finding it. First, we need to set the data in order, and it does not matter if the setting is done ascending or descending. Second, look up the value of n , the sample size. Third determine the observations in the middle of the array. Is n even, or is it odd? When n is an odd number, there is one middle value and it is at that point whose rank, in the array, equals $(n+1)/2$. Thus, the median in this case is the $[(n+1)/2]^{\text{th}}$ observation. When n is even, there are two middle values in the array, namely $[n/2]^{\text{th}}$ and $[n/2 + 1]^{\text{th}}$ observations. The median in this case is the average of those observations.

EXAMPLE 2.4: What is the **Median** for the data in **EXAMPLE 2.1**?

Solution: Let us find the median for the data in **Example 2.1**. The number of data points is 13, an odd number. Thus, the median is the data point whose rank in the array is $(13+1)/2 = 7$. Let us arrange the data by an ascending order as: 7, 8, 9, 10, 11, 12, 15, 18, 19, 20, 22, 30, and 34. We see that 15 is the seventh element in the array. Thus, the median = 15.



Trust and responsibility

NNE and Pharmaplan have joined forces to create NNE Pharmaplan, the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries.

Inés Aréizaga Esteva (Spain), 25 years old
Education: Chemical Engineer

– You have to be proactive and open-minded as a newcomer and make it clear to your colleagues what you are able to cope. The pharmaceutical field is new to me. But busy as they are, most of my colleagues find the time to teach me, and they also trust me. Even though it was a bit hard at first, I can feel over time that I am beginning to be taken seriously and that my contribution is appreciated.



NNE Pharmaplan is the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries. We employ more than 1500 people worldwide and offer global reach and local knowledge along with our all-encompassing list of services.
nnepharmaplan.com

nne pharmaplan®



Definition 2.6: The Mid-Range is another value that can be used to check on the center of any data. It is rarely used but it will give an indication when compared to the measures of central tendency. The Midrange is the average of the two extreme values in the data, i.e. the average of the maximum and the minimum in the array, i.e., $\text{Mid-Range} = (\text{Max} + \text{Min})/2$.

EXAMPLE 2.5: What is the Mid-Range of the data in **Example 2.1**?

Solution: Looking at the array in **EXAMPLE 2.4**, we see that the minimum is 7 and the maximum is 34. Therefore, the **Mid-Range** is: $\text{MR} = (7 + 34)/2 = 20.5$.



EXAMPLE 2.6: The following is a sample of time intervals between successive menstrual periods (in days) in college-age students:

24	25	25	26	26	27	27						
27	28	28	28	28	28	29	29	28	29	29	30	30
33	30	31	31	32	33	33	34	34	35	36	36	36
37	38											

Calculate the measures of center for the above data.

Solution: There are 35 data points in the sample. The measures of center include: The mean, The median, The mode, and The Midrange. Thus, we have: Mean = 30.3, Median = 29, Mode = 28, with min= 24 and max = 38, we see that Midrange = $(24 + 38)/2 = 31$.

2.2.2 THE MEASURES OF VARIATION

In **Section 2.2.1** we discussed the measures of central tendency, which they measure a typical value of the variable involved. We would like to know the amount of dispersion, or variation, in the variable. Dispersion is the degree to which the data are spread out. Individual differences or variations exist. This is not only a fact, but it is an interesting fact to all of us. The study of the variability among opinions, buying habits, learning abilities, mating behavior, profits, and so forth, occupies a great deal of scientific energy. In this section, we initiate a discussion of variability with a presentation of the basic methods which numerically describe the variability in the observations of a sample and of a population.

Under our discussion of the measures of central tendency, we implicitly acknowledged that such variability exists or why else would we need to compute a single mean (or a median, or a mode) to summarize data? If the data points are all equal, then the first three measure of central tendency are equal.

In describing a set of data, a measure of central tendency alone does not really tell us enough to make many decisions or inferences. Several distributions can have the same mean yet the shape of the distribution of observations can be quite different. Thus, to describe a set of data we typically use some measures of variation among the observations along with a measure of central tendency. As was true of the measures of central tendency, there are a number of measures of variation, and each is communicating or giving a different kind of information about the data. Our goal is to discuss the measures of dispersion in the data so we can quantify the spread of data. There are three numerical measures for describing the variation, or spread, or dispersion, in data. These measures are:

1. The Range, 2. The Variance, and 3. The Standard Deviation.

The simplest measure of dispersion is the range. Thus, the Range, R , of a variable is the difference between the largest and the smallest values in the ordered data. That is,

Definition 2.7: Range = R = Largest data value – smallest data value = Maximum – minimum.

EXAMPLE 2.7: What is the range for the data in **EXAMPLE 2.1**?

Solution: From the array in **EXAMPLE 2.1**, we see that minimum = 7, and the maximum = 34. Therefore, the range is given by $R = 34 - 7 = 27$.



More specifically, for n observations which are ordered from the smallest $Y_{(1)}$ to the largest $Y_{(n)}$ the range is: $R = Y_{(n)} - Y_{(1)}$. The range does not tell us how the observations are distributed between the smallest and the largest ones. The only information we really have from the range is the distance between the smallest and the largest measurements. As such, the range statistic is not a measure of dispersion of all the observations. It is a measure of the distance between the extremes in the data. Describing the distribution in terms of the range can be useful in cases of casual communications, e.g., “The grades on the first test were evenly distributed over a range of 30 points”. As it was with the measures of central tendency, we usually do not give the range statistic without qualifying it with other information, such as the measure of central tendency and the value of either the lowest and/or the highest observations.

The Variance and the Standard Deviation

These are the most important concepts in a course on elementary statistics. Do not treat these concepts lightly because the rest of the course relies upon your understanding how to compute and use the variance and the standard deviation for a set of data. Some of the applications of these two concepts will be discussed below, but not all, since these measures of variability will be an integral part of every remaining chapter in this book. For a moment, we talk about the population variance and standard deviation using the lower case Greek letter σ , sigma, where σ^2 is the population variance, while σ is the standard deviation. It is needless to say that the relationship between the variance and the standard deviation is given by

The Standard Deviation = The Positive Square Root of the Variance.

Our discussion Starts with the mean of the population μ , and how those observations are distributed around that value of μ . We are interested in the variability of the observations, in other words, we like to see what variation is there by calculating the variance and the standard deviation by using μ as a reference point. For any one observation Y_i we can check how far that observation is from the mean μ , i.e. in the difference $Y_i - \mu$. This difference is called the deviation of the observation Y_i from the mean μ . Based on the definition of

This e-book
is made with
SetaPDF





SETASIGN

PDF components for PHP developers

www.setasign.com



the mean for a finite population, it is easily seen that “the sum of the deviations of all data points from the mean of the finite population” is zero. In case we have a sample on our hands, the deviation from the mean of the sample \bar{X} can be found as $X_i - \bar{X}$. Moreover, based on the definition of the sample mean, we can see that “The sum of the deviations of the sample data points from their mean is also zero” (**This is an exercise for the student**).

We now turn to how to compute the variance and the standard deviation of a finite population.

Definition 2.8: The population variance σ^2 is defined as the average of all the squared deviations of the observations about their mean, i.e.

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

Thus, the standard deviation σ will be given by $\sigma = \sqrt{\sigma^2}$.

The formula for the variance, as defined above can be simplified and more accurate in case there was a rounding in calculating the mean, μ , of the data.

Definition 2.9: In case we have a sample of n points, the sample variance, S^2 , and the sample standard deviation, S , will be respectively given by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Thus, the standard deviation will be given by the positive square root of the variance: $S = \sqrt{S^2}$.

EXAMPLE 2.8: Consider the following finite population that has these observations: 2, 4, 6, 8, and 10. Calculate the variance and the standard deviation for this population.

Solution: As described above we see that $\mu = (2 + 4 + 6 + 8 + 10)/5 = 6 = \bar{X}$, Thus we have $\sigma^2 = [(2 - 6)^2 + (4 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 + (10 - 6)^2]/5 = 8$, and $\sigma = 2\sqrt{2} = 2.8284$.

◆ -----

In case we considered the above data as a sample, we can see that $\bar{X} = 6$, but there is a difference in calculating the variance and the standard deviation for a sample, as given by $S^2 = 10$, instead of 8, since we divide by 4. Hence $S = \sqrt{10}$. Thus, as it shows, S is greater than σ .

NOTE 1: From the example above, you can find that the sum of the deviation around the mean of the population, or of the sample, is zero, i.e. $[(2 - 6) + (4 - 6) + (6 - 6) + (8 - 6) + (10 - 6)] = 0$.

NOTE 2: The above formulas for the variance and the standard deviation, whether we have a population or a sample, are by definition. Other computational formulas are available, and could be more accurate, especially if there were some rounding in finding the means for the population and the sample. The formulas that will give the variances for the population and the sample are, respectively, as follows:

$$\sigma^2 = \frac{\sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N}}{N}, \text{ and } S^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}$$

Similarly, the formulas that will give the standard deviations for the population and the sample are, respectively, as follows:

$$\sigma = \sqrt{\sigma^2}, \text{ and } S = \sqrt{S^2}.$$

NOTE 3: Range Rule of Thumb for Understanding the Variability in a Distribution

The **Range Rule** of thumb is a crude but simple rule for understanding the spread in the data and interpreting the standard deviation. As it will be clear when we apply the **Empirical Rule** (See Figure 8, below), the majority of the data (such as 95%) will be within 2 standard deviations from the mean of the data. Thus, to roughly estimate the standard deviation, the **Range Rule** of thumb states that

$$S \approx \frac{\text{Range}}{4}.$$

In the above formula, we are sacrificing accuracy for the sake of simplicity. We could be more accurate, if the **Range Rule** of thumb is modified to be

$$S \approx \frac{\text{Range}}{6}.$$

(Check the Empirical rule below).

NOTE 4: Coefficient of Variation:

The **Coefficient of Variation (CV)** is a measure that allows for the easy comparison of two or more variables measured on different scales. It gives the relative variability in terms of the mean of the variable. It is a ratio and thus, it has no units of measurement. The lower the Coefficient of Variation is the better. In that it will show a small variability in the data. As it is as for the measures of center and variation, there are two Coefficients of Variation, namely one for a population and another one for a sample. This measure is applicable more often on a sample rather than on a population, and it is expressed as percentage. It is given by

$$CV = [\text{Standard Deviation} / \text{Mean}] \cdot 100\%.$$

NOTE 5: Coefficient of Skewness (CS):

The skewness of a distribution can often tell us about the relative values of the measures of center, see **Figures 1-3** below. For a sample data, the shape of the distribution will become clear after drawing the frequency polygon on the histogram. This **Coefficient of Skewness**



FOSS

Sharp Minds - Bright Ideas!

Employees at FOSS Analytical A/S are living proof of the company value - First - using new inventions to make dedicated solutions for our customers. With sharp minds and cross functional teamwork, we constantly strive to develop new unique products - Would you like to join our team?

FOSS works diligently with innovation and development as basis for its growth. It is reflected in the fact that more than 200 of the 1200 employees in FOSS work with Research & Development in Scandinavia and USA. Engineers at FOSS work in production, development and marketing, within a wide range of different fields, i.e. Chemistry, Electronics, Mechanics, Software, Optics, Microbiology, Chemometrics.

We offer
A challenging job in an international and innovative company that is leading in its field. You will get the opportunity to work with the most advanced technology together with highly skilled colleagues.

Read more about FOSS at www.foss.dk - or go directly to our student site www.foss.dk/sharpminds where you can learn more about your possibilities of working together with us on projects, your thesis etc.

Dedicated Analytical Solutions

FOSS
Slangerupgade 69
3400 Hillerød
Tel. +45 70103370
www.foss.dk

The Family owned FOSS group is the world leader as supplier of dedicated, high-tech analytical solutions which measure and control the quality and production of agricultural, food, pharmaceutical and chemical products. Main activities are initiated from Denmark, Sweden and USA with headquarters domiciled in Hillerød, DK. The products are marketed globally by 23 sales companies and an extensive net of distributors. In line with the corevalue to be 'First', the company intends to expand its market position.





was developed as a measure by **Karl Pearson**, and it is given by the following formula (which is applicable for both a sample and a population given data):

$$CS = 3[(\text{Mean} - \text{Median})/\text{Standard deviation}]$$

The value of this measure usually lies between -3 and +3, i.e. $-3 < CS < 3$. The closer is the value to -3, the more indication that the distribution is negatively skewed. On the other hand, the closer is the value to +3, the more indication that the distribution is positively skewed. Moreover, a value of zero shows a symmetric distribution, as shown in the Figures below.

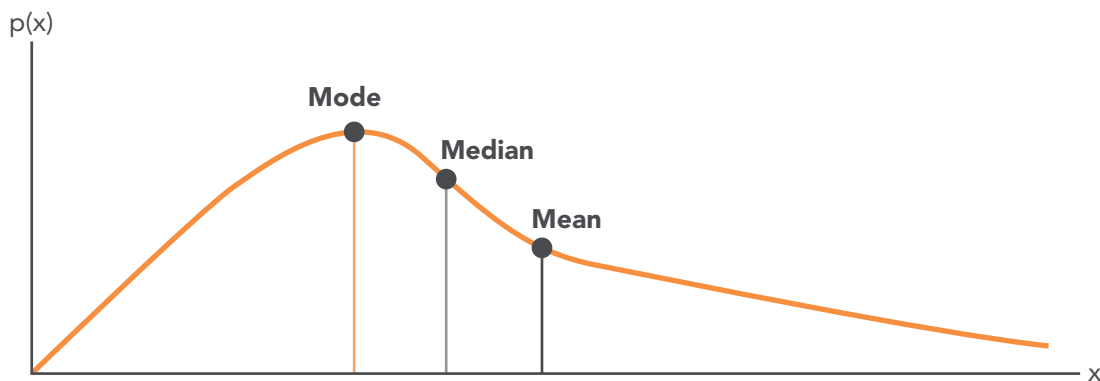


Figure 1. A Positively Skewed Distribution (Reference Internet)

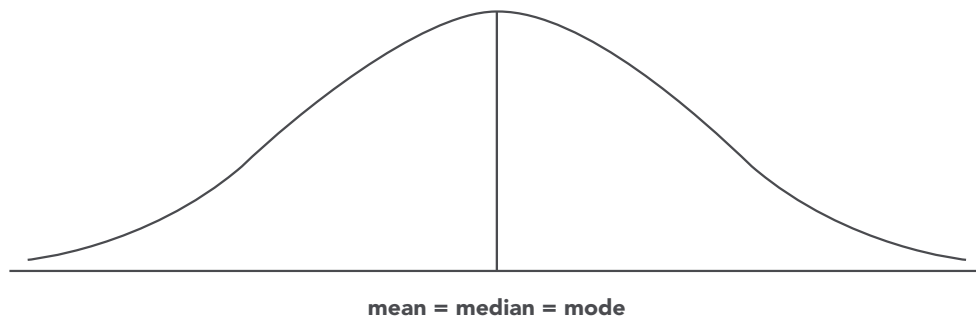


Figure 2. A Symmetric Distribution (Reference Internet)

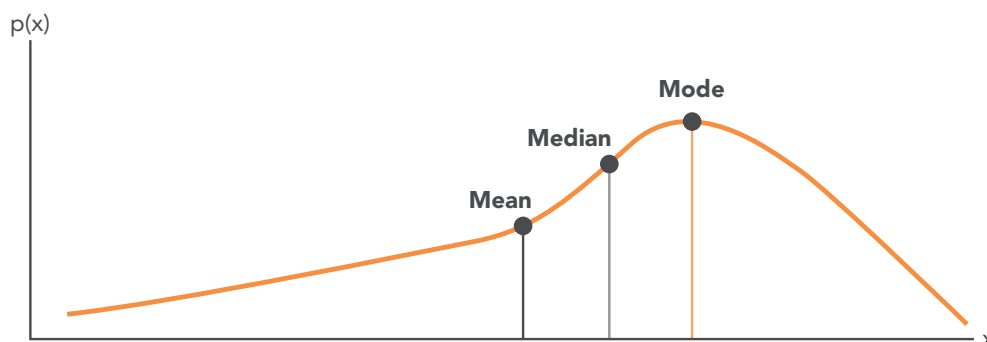


Figure 3. A Negatively Skewed Distribution (Reference Internet)

NOTE 6: Mean of Absolute Deviations (MAD):

It was seen earlier, that the sum of the deviations around the mean, whether we had a population or a sample, that sum is zero. In this case, we take the absolute value of each of those deviations, add them up and divide by the number of points we have. Thus, we see that, when data is presented by $X_i, i = 1, 2, \dots, n$ or $X_i, i = 1, 2, \dots, N$,

$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}, \text{ in the sample case.}$$

While in the finite population case we have

$$\text{MAD} = \frac{\sum_{i=1}^N |X_i - \mu|}{N}.$$

NOTE 7: Coefficient of Kurtosis (CK):

This coefficient is more involved with a population than with a sample. We give the formula for the coefficient of Kurtosis, and leave its calculation and applications for a higher course in mathematical Statistics.

$$\text{CK} = E [(X - \mu)^4] / (E[(X - \mu)^2])^2$$

REMARK: The above three coefficients, namely CV, CS, and CK are of theoretical value, as well as for checking on the quality of the data for later analysis.

EXAMPLE 2.9: Consider the following set of data: 12, 15, 25, 24, 27, 34, 15, 20, 7, 8, 9, 10, 22, 18, 30, 17, 21, 11, 12, 15, 16, 19, 20, 6 and 19. Calculate: a) The Coefficients of Skewness, b) The MAD, and c) The Coefficient of Variation.

Solution: We will treat the data as sample, and use the TI-83-plus calculator to find and display the above mentioned coefficients.

Putting data in L_6 , Stat Calculate 1-Variable Stat L_6 gave the following:

$$\begin{aligned} \bar{X} &= 17.28 & \sum_{i=1}^n X_i &= 432 & \sum_{i=1}^n X_i^2 &= 8716 & S_x &= 7.219879962 \\ \sigma_x &= 7.074008764 & n &= 25 & \text{Minx} &= 6 & Q_1 &= 11.5 & \text{Med} &= 17 & Q_3 &= 21.5 \\ \text{Maxx} &= 34. \end{aligned}$$

Based on the above calculations we have:

a) $CS = 3[(\text{Mean} - \text{Median})/\text{Standard deviation}] = 3 \cdot [17.28 - 17]/7.219879962 = 0.1163.$

As it shows the above sample data is positively skewed since $\text{Mode} < \text{Median} < \text{Mean}$, i.e.,
 $< 17 < 17.28$

- b) Based on the data being taken as a sample, we have $\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n} = 5.7312$.
 In case you consider the data as a population, we find that, in this case as well, that $\text{MAD} = 5.7312$.
- c) $\text{CV} = [\text{Standard Deviation} / \text{Mean}] \cdot 100\% = 100 \cdot (7.219879962/17.28) = 41.78\%$.

2.3 SOME PROPERTIES OF THE NUMERICAL MEASURES OF QUANTITATIVE DATA

1. If a constant is added to each data point, the mean of the new data will be the old mean plus that constant. Let Y_i , $i = 1, 2, \dots, n$ be the original data, and $X_i = Y_i + b$, where b is a constant. It is easily seen that $\bar{X} = \bar{Y} + b$. Similarly, the mode, and the median will be changed by adding the same constant to get the new ones. The mid-range will not change. Proof is left as an exercise for the student.

"I studied English for 16 years but...
 ...I finally learned to speak it in just six lessons"
 Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

2. If data were multiplied by a constant, $X_i = bY_i$, $i = 1, 2, \dots, n$ then the mean of the new data will be $\bar{X} = b\bar{Y}$. Also in this case, the mode, the median and the mid-range will change. Each will be multiplied by that constant. Proof is left as an exercise for the student.
3. If a constant is added to each point, the variance will not change. Based on that, there will be no change in the standard deviation. Proof is left as an exercise for the student.
4. If each point in a data set is multiplied by a constant, then the variance will be multiplied by the square of that constant. The standard deviation will be multiplied by the absolute value of that constant. Why? Proof is left as an exercise for the student.
5. In case the shape of the distribution for the data is roughly bell-shaped, the Empirical Rule states that:

The interval: $(\mu - \sigma, \mu + \sigma)$ will contain approximately 68% of all the measurements

The interval: $(\mu - 2\sigma, \mu + 2\sigma)$ will contain approximately 95% of all the measurements

The interval: $(\mu - 3\sigma, \mu + 3\sigma)$ will contain approximately 99.7% of all the measurements

Recall that the Empirical Rule can be used in case we have a sample with the sample mean \bar{X} replacing μ and the sample standard deviation S replacing σ in the above inequalities. We thus have the Figure 8, below.

6. Tchebyshev's (Chebyshev's) Inequality (pronounced Tcheb-e-shev's): Given a constant $k > 1$, and regardless of the shape of the distribution, for any set of data, at least $(1 - 1/k^2)$ 100% of the observations will lie within k standard deviations of the mean, i.e., at least $(1 - 1/k^2)$ 100% of the data will lie between $\mu - k\sigma$ and $\mu + k\sigma$. We can also use Tchebyshev's Inequality based on a sample data.

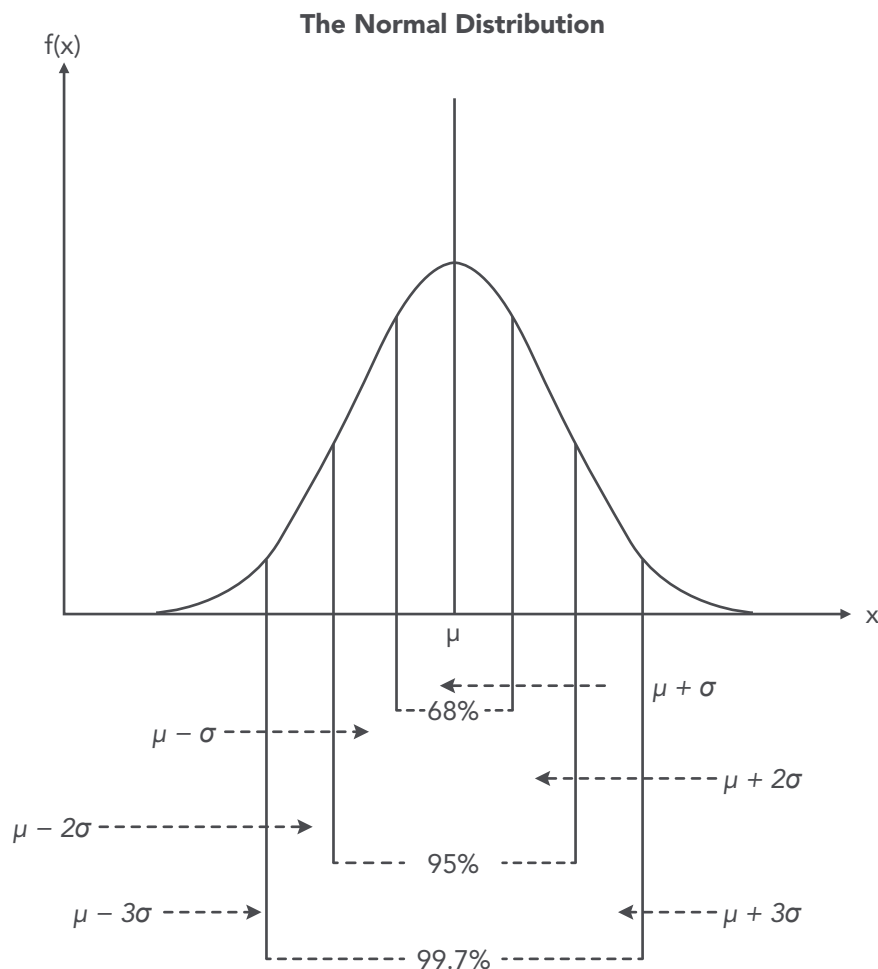


Figure 4 Empirical Rule showing the probabilities on the Normal Curve

2.4 MEASURES OF POSITION FOR QUANTITATIVE DATA

In **Section 2.2** we discussed the measures of central tendency, the measures of variation, and some properties of those measures. In this section, we discuss measures of position. These measures include:

1. **The Z-scores,**
2. **The Percentiles,**
3. **The Deciles,**
4. **The Quartiles,**
5. **Trimean**

The Z-score: It represents the unit-less distance that a data point is away from the mean of all observations in terms of the number of standard deviations. As it can be seen, the

Z-score is a ratio, and it is unit-less, i.e. there are no units of measurement for the Z-scores. There are two Z-scores, a Z-score for the population and another one for the sample. Their formulas are, respectively, given by

Definition 2.10: **Population Z-Score:** $Z_i = \frac{X_i - \mu}{\sigma}$ and
Sample Z-Score: $Z_i = \frac{(Y_i - \bar{Y})}{S}$.

As it is seen from the above formulas, each data point (whether of a sample or of a population) has a Z-Score.

The Z-scores have mean 0 and standard deviation 1. This is an exercise for the student to verify.

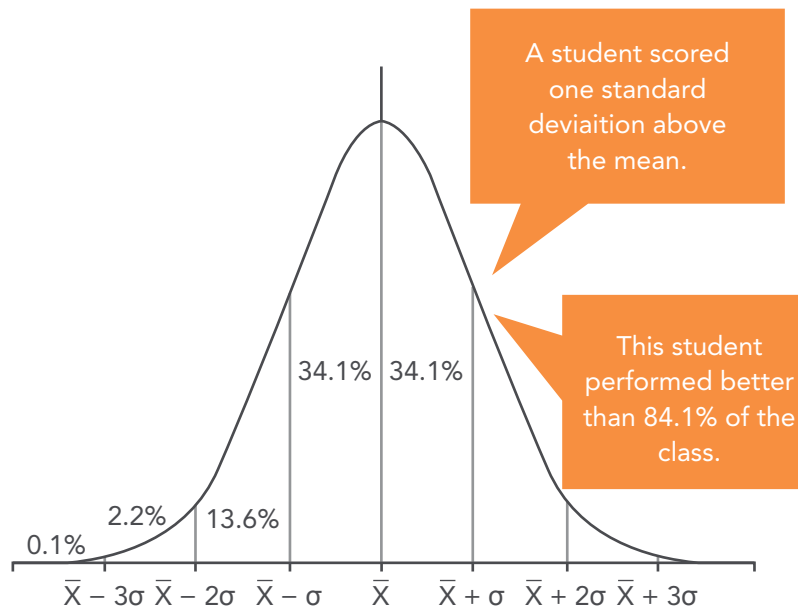


Figure 5 Empirical Rule Displayed on the Normal Curve
(Source: Internet, Normal curve images)

Figure 6 displays the Empirical rule using **Z-scores**.

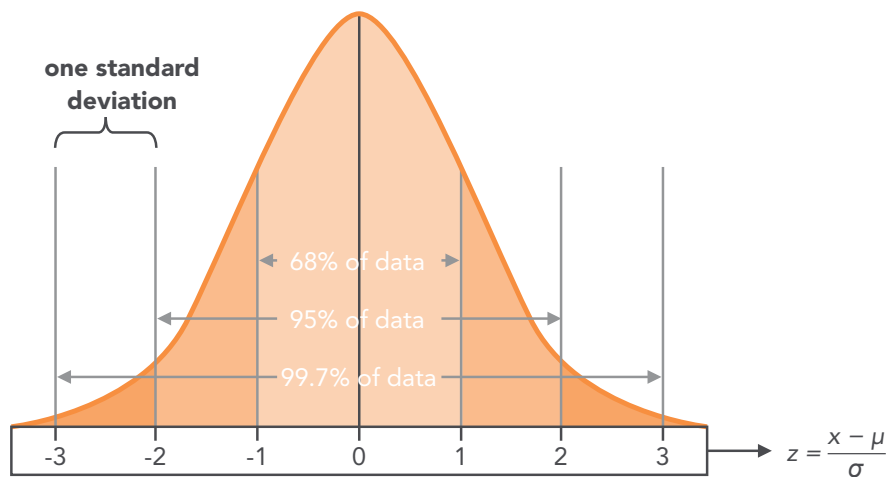


Figure 6 Empirical Rule Displayed on the Normal Curve
(Source: Internet, Normal curve images)

The **Kth Percentile**, denoted by P_k , of a set of data, is a value such that k percent of the data are less than or equal to that value. Thus, the percentiles divide the array, the data set in order of magnitude, into 100 parts; hence 99 percentiles can be determined. Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT, or ACT, College Entrance Exams use percentiles to provide students with understanding

The Wake

the only emission we want to leave behind

Low-speed Engines Medium-speed Engines Turbochargers Propellers **Propulsion Packages** PrimeServ

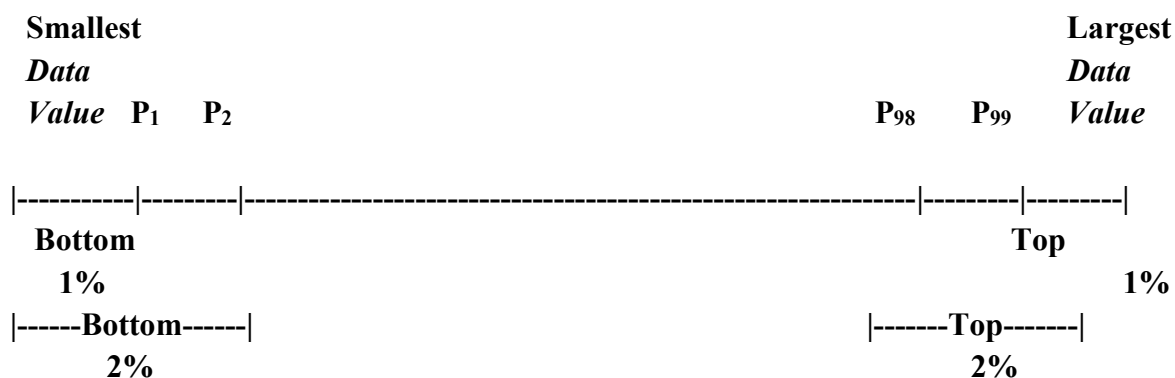
The design of eco-friendly marine power and propulsion solutions is crucial for MAN Diesel & Turbo. Power competencies are offered with the world's largest engine programme – having outputs spanning from 450 to 87,220 kW per engine. Get up front! Find out more at www.mandieselturbo.com

Engineering the Future – since 1758.

MAN Diesel & Turbo

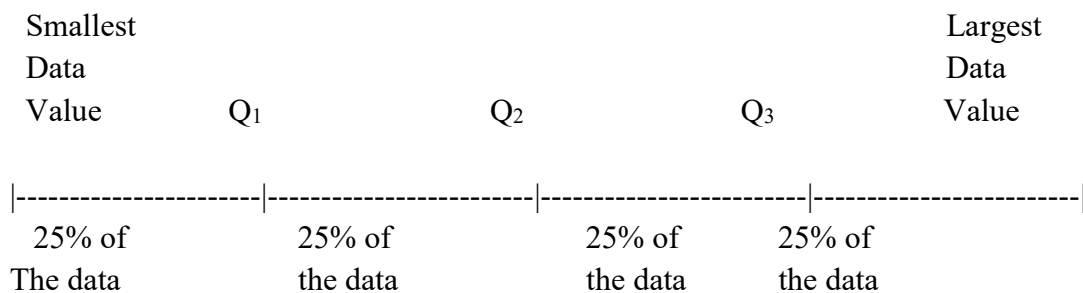


of how they scored in the exam in relation to the other students who participated in the same exam.



The **Deciles** are parts of the percentiles that divide the data array into 10 parts, with each value presenting 10% of the data less than or equal to that value. Clearly the fifth **Decile**, D_5 , is the **Median** and it is equal to P_{50} , the 50th **Percentile**.

The most common used percentiles are the **Quartiles**. Quartiles divide the ordered data, the data array, into fourths, or four equal parts. The second Quartile, Q_2 , is the **Median**, and it divides the bottom 50% of the data from the top 50%. Thus, it is the 50th **Percentile**. Clearly it can be seen that the first Quartile, Q_1 , is the median of the lower half of the data, while the third quartile, Q_3 , is the median of the upper half of the data set.



Within the measures of dispersion, we have introduced the **Range** and the **Standard Deviation**, neither of which is resistant to the extreme values. **Quartiles**, on the other hand, are resistant to the extreme values. Based on this property, we can define a measure of dispersion that is based on quartiles, namely the **Interquartile Range**.

Definition 2.11: The **Interquartile Range, IQR**, is the range of the middle 50% of the observations in the ordered data set. That is, the IQR is the difference between the third quartile and the first quartile and it is found by the following formula:

$$IQR = Q_3 - Q_1.$$

As it was the case with the range and standard deviation, the larger the IQR the more spread a data set has.

Whenever performing any type of data analysis, we should always check for extreme observations in the data set. Extreme observations are referred to as outliers. If outliers were encountered, their origin should be investigated. They can occur by chance, because of error in the measurement of a variable, during data entry, or from errors in sampling. We can use the following steps to check outliers using quartiles. For sure there might be outliers on either end of the data array, i.e. values are too small to be considered acceptable, or there are values that are too large to be taken as true values. Having calculated the IQR, then we determine the fences which serve as the cutoff points for determining outliers. Thus, we have

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR}), \quad \text{Upper Fence} = Q_3 + 1.5(\text{IQR}).$$

Any data value that is less than the lower fence, or greater than the upper fence, will be considered as an outlier.

The **Trimean** is another measure of the center of the data, or the middle of the sample. It is not frequently used in practice, and it serves as a measure of quality since it is found by using the quartiles. It is calculated by the following formula:

$$\text{Trimean} = (Q_1 + Q_2 + Q_3)/4$$

In the previous sections, we have presented the numerical measures for quantitative data. In those sections, we have found that the median is resistant to extreme values and it is the preferred measure of central tendency when data is skewed right or left. Similarly, the IQR is also resistant to the extreme values. However, the median, Q_1 and Q_3 do not provide information about the extreme values in the data. To get this information, we need to know the smallest and largest values in the data set. Thus, we have what we call the five-number summary of a data set that consists of the smallest data value, Q_1 , the median, Q_3 , and the largest data value. The five-number summary can be used to make another graph, called the **Boxplot**.

A reasonable question to ask at this time is “Why all the fuss about having different symbols distinguishing population and sample measures?” The answer is quite simple. Even with good sampling procedures the sample mean and the sample standard deviation will not necessarily be equal to the population mean and standard deviation respectively. In fact, \bar{Y} and S^2 will be typically not equal to μ and σ^2 respectively, even though we wish to make inferences about the characteristics of the population by using those statistics of the sample. Our hope,

of course, is to use computational procedures for the sample's characteristics (e.g. \bar{Y} and S^2) which would provide good estimates for the population's characteristics (μ and σ^2).

Let us add two more definitions to your vocabulary.

Definition 2.12: A Parameter: is a characteristic of the population set of observations; e.g. N, μ, σ and σ^2 .

Definition 2.13: A Statistic: is a characteristic of a sample of observations, e.g., $n, \bar{Y}, S,$ and S^2 .

2.5 DESCRIPTION OF GROUPED DATA

It is often that the data is already summarized in a Frequency Table. When it is given in terms of the classes' limits and/or boundaries, it is difficult to retrieve the actual raw data. In such a case, it is not easy to find an exact value for the mean or the standard deviation. Given the Frequency Table for the data, we will assume that within each class all the data values, in that class, is equal to the class midpoint. We then multiply the class midpoint by

qaiteye[®]
Challenge the way we run

**EXPERIENCE THE POWER OF
FULL ENGAGEMENT...**

**RUN FASTER.
RUN LONGER..
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY
WWW.GAITEYE.COM**

the frequency of that class. This product is expected to be very close to the sum of the data that lie in that class. The process is repeated for all the classes, and the total will be calculated. This sum approximates the total of the data on hand. Thus, based on this procedure we see that the means, for the population and the sample, respectively, are given by

$$\text{Population mean is } \mu = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i}, \text{ and Sample mean is } \bar{x} = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i}.$$

In the above formulas; X_i is the midpoint of the i^{th} class, f_i is the frequency in that class, and n is the number of classes. Similarly, we can find the variance and the standard deviation of grouped data by the following formulas:

$$\text{Population variance is } \sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2 f_i}{\sum_{i=1}^n f_i}, \text{ and Sample variance is } S^2 = \frac{\sum_{i=1}^n (X_i - \bar{x})^2 f_i}{(\sum_{i=1}^n f_i) - 1}.$$

Again, where X_i is the midpoint of the i^{th} class, f_i is the frequency in that class, and n is the number of classes. Alternatively, there is another equivalent formula for the calculations of the population and sample variances that might give more accurate values, in this case, than the above formulas. We are talking about the following formulas for the variances and the standard deviations of the population and the sample.

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i)^2 f_i - \frac{(\sum_{i=1}^n X_i f_i)^2}{\sum_{i=1}^n f_i}}{\sum_{i=1}^n f_i}, \text{ and } S^2 = \frac{\sum_{i=1}^n (X_i)^2 f_i - \frac{(\sum_{i=1}^n X_i f_i)^2}{\sum_{i=1}^n f_i}}{(\sum_{i=1}^n f_i) - 1}.$$

There is no doubt that by taking the square root of the above formulas we can get the standard deviations for the population and the sample, respectively. To clarify all of the above, here is an example.

EXAMPLE 2.10: Let us consider the following data, as summarized below (Check **EXAMPLE 1.8**, and its summary, as given in Table 2, Chapter 1), and let us look at it in a different way. Consider the following presentation for the cited data, and check as the mid points of the classes in

X	25	35	45	55	65	75	85	95
f	1	2	2	3	12	14	12	4

Solution: Thus, we have $\sum_{i=1}^n f_i = 50$, not necessarily all different, and $n = 8$, classes. Applying the above formulas for the sample mean and variance we calculate that $\bar{x} = \frac{\sum_{i=1}^8 X_i f_i}{\sum_{i=1}^8 f_i} = [1 \cdot (25) + 2 \cdot (35) + 2 \cdot (45) + 3 \cdot (55) + 12 \cdot (65) + 14 \cdot (75) + 12 \cdot (85) + 4 \cdot (95)] / 50 = 71.6$

$$S^2 = \frac{\sum_{i=1}^n (X_i)^2 f_i - \frac{(\sum_{i=1}^n X_i f_i)^2}{\sum_{i=1}^n f_i}}{(\sum_{i=1}^n f_i) - 1} = \frac{268450 - \frac{(3580)^2}{50}}{49} = 247.3878.$$

Hence $S = 15.7286$.



Classes	Tally	Frequency	Relative Frequency (%)	Cumulative Rel. Freq (%)
20-29	/	1	2	2
30-39	//	2	4	6
40-49	//	2	4	10
50-59	///	3	6	16
60-69	//// //	12	24	40
70-79	//// //	14	28	68
80-89	//// //	12	24	92
90-99	////	4	8	100
Total		50	100%	

Table 1

EXAMPLE 2.11: Let us have another look at the data in **EXAMPLE 1.8**. We like to calculate the measures of central tendency, the measures of variation, the measures of quality and the Z-scores. We can do all that by entering the data in a graphing calculator, and with some manipulations we have the following picture:

Solution: **1-var Stats**

$$\bar{X} = 71.26$$

$$\sum_{i=1}^n X_i = 3563$$

$$\sum_{i=1}^n X_i^2 = 264781$$

$$S_x = 14.90214339$$

$$\sigma_x = 14.7523693$$

$$n = 50$$

$$\text{Min}_x = 29$$

$$Q_1 = 63$$

$$\text{Med} = 73.5$$

$$Q_3 = 82$$

$$\text{Max}_x = 97$$

It can be seen, from the display above, that we need to do some work to get all the measures we are looking for. There is No mode, no range, no variance, No IQR, and No mid-range.

**Technical training on
WHAT you need, *WHEN* you need it**

At IDC Technologies we can tailor our technical and engineering training workshops to suit your needs. We have extensive experience in training technical and engineering staff and have trained people in organisations such as General Motors, Shell, Siemens, BHP and Honeywell to name a few.

Our onsite training is cost effective, convenient and completely customisable to the technical and engineering areas you want covered. Our workshops are all comprehensive hands-on learning experiences with ample time given to practical sessions and demonstrations. We communicate well to ensure that workshop content and timing match the knowledge, skills, and abilities of the participants.

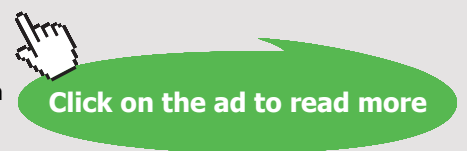
We run onsite training all year round and hold the workshops on your premises or a venue of your choice for your convenience.

For a no obligation proposal, contact us today at training@idc-online.com or visit our website for more information: www.idc-online.com/onsite/

OIL & GAS ENGINEERING
ELECTRONICS
AUTOMATION & PROCESS CONTROL
MECHANICAL ENGINEERING
INDUSTRIAL DATA COMMS
ELECTRICAL POWER

Phone: +61 8 9321 1702
Email: training@idc-online.com
Website: www.idc-online.com

IDC TECHNOLOGIES



There is something extra here, and that is σ_x , as if the data were looked at as a population. That is the standard deviation of the assumed population. Thus

$$S^2 = (14.90214339)^2 = 222.0738776$$

$$\sigma_x = (14.7523693)^2 = 217.6324$$

No doubt that $S_x > \sigma_x$, since the denominator in finding those values was $n-1$ and n respectively.

$$\text{Mid-Range} = (97 + 29)/2 = 63.$$

Clearly, since they are given, the following quantities, $\sum_{i=1}^n X_i = 3563$, and $\sum_{i=1}^n X_i^2 = 264781$, can be used to find more accurate variance for the sample and the population, in case it is needed.

$$\text{Range} = 97 - 29 = 68$$

$$\text{IQR} = 82 - 63 = 19$$

Mode = 63. It appeared 3 times more than any other data point.

$$\text{Lower Fence} = Q_1 - 1.5 \cdot \text{IQR} = 63 - 1.5 \cdot 19 = 34.5$$

$$\text{Upper Fence} = Q_3 + 1.5 \cdot \text{IQR} = 82 + 1.5 \cdot 19 = 110.5$$

It shows that there is an outlier on the lower end. The outlier is the data point 29, the MinX in the data. There are no outliers on the upper end, since all data points are < 110.5 . In addition to the above we have what we call the 5 number summaries: MinX, Q_1 , Med, Q_3 , and MaxX, displayed on the box plot.

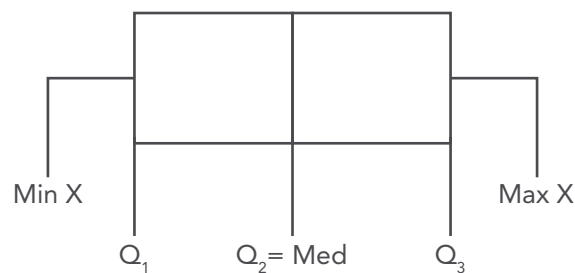


Figure 7 BOXPLOT and the Five Number Summaries



The **Histogram**, **Figure 8** for the data, in **Table 1**, is shown below. It is skewed to the left, or negatively skewed. The class width used is 10, and the frequency is as displayed in Table 2 for the classes, with the lower limit of the first class is 20, the upper limit is 30, which stands as the lower limit for the second class, and so on, as again displayed in

Table 2. Based on the calculations for the lower fence and the upper fence we found that the minimum, in the data, is an outlier; since $29 < 34.5$, as displayed on the box plot.

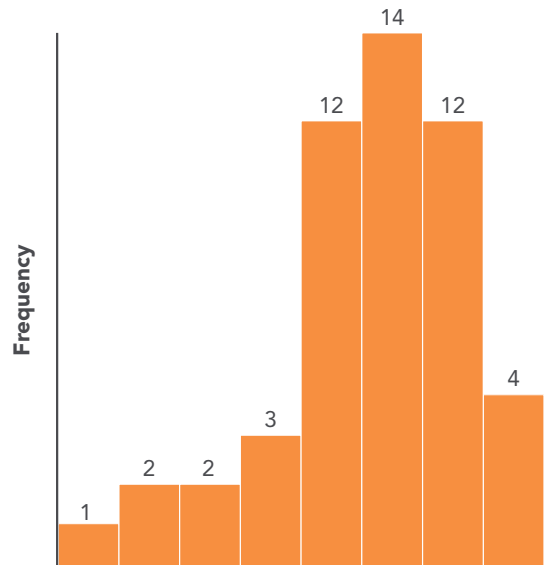


Figure 8 Histogram for Data in Table 1

CHAPTER 2 EXERCISES

2.1 The following are the scores made on an intelligence test by a group of children who participated in the experiment:

114	115	113	112	113	132	130	128	122	121	126	117	115
88	113	90	89	106	104	126	127	115	116	109	108	122
123	149	140	121	137	120	138	111	100	116	101	110	137
119	115	83	109	117	118	110	108	134	118	114	142	120
119	143	133	85	117	147	102	117					

Calculate the Measures of center for the above data (refer to Exercise 1.1).

2.2 75 employees of a general hospital were asked to perform a certain task. The taken to complete the task was recorded. The results (in hours) are as shown below:

1.5	1.3	1.4	1.5	1.7	1.0	1.3	1.7	1.2	1.8	1.1	1.0	1.8
1.6	2.1	2.1	2.1	2.1	2.4	2.9	2.7	2.3	2.8	2.0	2.7	2.2
2.3	2.6	2.8	2.1	2.3	2.4	2.0	2.8	2.2	2.5	2.9	2.0	2.9
2.5	3.6	3.1	3.5	3.7	3.7	3.4	3.1	3.5	3.6	3.5	3.2	3.0
3.4	3.4	3.2	4.5	4.6	4.9	4.1	4.6	4.2	4.0	4.3	4.8	4.5
5.1	5.7	5.1	5.4	5.7	6.7	6.8	6.6	6.0	6.1			

Calculate the Measures of Variation for the above data (refer to Exercise 1.2).

2.3 On the first day of classes, last semester, 25 students were asked for their one-way Travel-Time from home to college (to the nearest 5 minutes). The resulting data were as follows:

20	20	30	25	20	25	30	15	10	40	35	25	15
25	25	40	25	30	5	25	25	30	15	20	45	

Calculate the geometric Mean for the time of travel to college

2.4 On the first day of classes, 20 students were asked the distance they drove to school, and the data was given as shown below:


35	25	10	10	15	20	20	20	25	20	20	20
20	10	20	30	10	25	15	25	20			

Calculate the Harmonic mean for the above data.

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com







Month 16

I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements







2.5 Show that: $\sum_{i=1}^n (X_i - \bar{X}) = 0$, for any n values of X_i , with \bar{X} being their mean.

2.6 The following are the fasting blood glucose levels of a sample of 10 children:

56 62 63 65 65 65 65 68 70 72

Compute: a) The mean b) The median c) The Mode d) The Range.

2.7 The following are the weights (in pounds) of 10 animals:

13.2 15.4 13.0 16.6 16.9 14.4 13.6 15.0 14.6 13.1

Find: a) The mean b) The median

2.8 The following are the scores made on an intelligence test by a group of children who participated in the experiment:

114 115 113 112 113 132 130 128 122 121 126 117 115
 88 113 90 89 106 104 126 127 115 116 109 108 122
 123 149 140 121 137 120 138 111 100 116 101 110 137
 119 115 83 109 117 118 110 108 134 118 114 142 120
 119 143 133 85 117 147 102 117

Find: a) The mean b) The median c) The Mode Class. d) Describe the skewness in the distribution

2.9 If 75 employees of a general hospital were asked to perform a certain task. The taken to complete the task was recorded. The results (in hours) are as shown below:

1.5 1.3 1.4 1.5 1.7 1.0 1.3 1.7 1.2 1.8 1.1 1.0 1.8
 1.6 2.1 2.1 2.1 2.1 2.4 2.9 2.7 2.3 2.8 2.0 2.7 2.2
 2.3 2.6 2.8 2.1 2.3 2.4 2.0 2.8 2.2 2.5 2.9 2.0 2.9
 2.5 3.6 3.1 3.5 3.7 3.7 3.4 3.1 3.5 3.6 3.5 3.2 3.0
 3.4 3.4 3.2 4.5 4.6 4.9 4.1 4.6 4.2 4.0 4.3 4.8 4.5
 5.1 5.7 5.1 5.4 5.7 6.7 6.8 6.6 6.0 6.1

Find: a) The mean b) The median c) The Mode Class d) Describe the skewness in the distribution

- 2.10 Consider the following sample: 2, 4, 7, 8, and 9, find the following:
 a) The mean b) The Mode c) The Midrange
- 2.11 Consider the following sample: 7, 6, 10, 7, 5, 9, 3, 7, 5, and 13, find the following:
 a) The measures of Central Tendency, b) The Range c) The Coefficient of Variation
- 2.12 15 randomly selected college students were asked to state the number of hours they slept last night. The resulting data are: 5, 6, 6, 8, 7, 7, 9, 5, 4, 8, 11, 6, 7, 8, and 7. find
 a) The mean b) The median c) The Mode d) The Coefficient of Variation
- 2.13 Find the mean, after grouping the data in Exercise 2.1, and compare with the results in 2.1.
- 2.14 Find the standard deviation after grouping the data in Exercise 2.2.
- 2.15 Give a detailed proof for the properties of the measures listed in section 2.3.
- 2.16 Mrs. Food wishes to develop a new type of meatloaf to sell at her restaurant. She decides to combine 2 pounds of ground sirloin (cost \$2.70 per pound), 1 pound of ground Turkey (cost \$1.30 per pound), and ½ pound of ground pork (cost \$1.80 per pound). What is the cost per pound of the meatloaf?
- 2.17 Determine the original set of data below. The stem represents tens digit and the leaf represents the ones digit, a) calculate the measures of central tendency, b) Draw the box-plot and check for outliers

Stem	Leaf
7	0 1 4
8	1 4 4 7 9
9	3 5 5 5 7 7 8
10	0 0 1 2 6 6 8 9 9
11	3358
12	12

2.18 Determine the original set of data below. The stem represents ones digit and the leaf represents the tenths digit, and calculate the measures variation:

Stem	Leaf
4	2 4 6
5	4 4 7 7 9
6	3 5 7 7 8
4	1 1 3 6 6 8 9 9
5	3 4 5 8
6	2 4

www.job.oticon.dk

oticon
PEOPLE FIRST



2.19 Determine the original set of data below, when the stem represents ones digit and the leaf represents the tenths digit, and find:

- a) The Coefficient of Variation b) The Coefficient of Skewness

Stem	Leaf
17	3 7 7 9
18	0 4 5 4 7 8 9
19	2 4 4 7 7 8 9
20	1 2 2 5 6 7
21	0 3 4 5 8
22	1 2 4

2.20 Refer to the data in **EXAMPLE 2.6** (Time between successive menstrual Periods), calculate

- a) The measures of variation. b) The measures of position and c) The Five-Number Summary. d) Draw the Box-plot., and give the IQR.

2.21 The following data is presented here as a practice exercise for the measures on data that had been presented in **Chapter 1** and **Chapter 2**, namely the graphical and numerical descriptions of real; Hospital data: The Data is giving weights at birth (oz) of 40 babies born at a Boston Hospital: Rosner, 2011, p. 22.

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113

Understanding the Concepts Exercises CHAPTER 2

1. What is a parameter? What is a statistic?
2. Will there be one mode in any set of data? Will there be one mean for the data set?
3. If the histogram for a set of data is positively skewed, how the mean and median compare?

4. Is the standard deviation resistant?
5. What does it mean when a statistic is biased?
6. What is the simplest measure of variation?
7. What is meant by an outlier? Should it be removed, and why?
8. What are the units of measurement for the z-scores?
9. What is the IQR?
10. When will the boxplot indicate that the data is positively skewed?

TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP Determining the Mean and Median

TI-83/84 Plus

1. Enter the raw data in L1 by pressing **Stat** and selecting **1: Edit**.
2. Press **Stat**, highlight the **CALC** menu, and select **1: 1-Var stats**.
3. With **1-Var Stats** appearing on the **HOME** screen, press **2nd** then **1** to insert **L1** on the **HOME** screen. Press **ENTER**.

Excel

1. Enter the raw data in column A.
2. Select **TOOLS** and **Data Analysis...**
3. In the **Data Analysis** window, highlight **Descriptive Statistics** and click **OK**.
4. With cursor in the **Input Range** window, use the mouse to highlight data in column A.
5. Select the **Summary Statistics** option and click **OK**.

TECHNOLOGY STEP-BY-STEP Determining the Range, Variance, and Standard Deviation

The same steps followed to obtain the measure of central tendency from raw data can be used to obtain the measures of dispersion.

TECHNOLOGY STEP-BY-STEP Determining the Mean and Standard Deviation from grouped Data

TI-83/84 Plus

1. Enter the class midpoints in **L1** and the frequency or relative frequency in **L2** by pressing **Stat** and selecting **1: Edit**.
2. Press **Stat**, highlight the **CALC** menu, and select **1: 1-Var stats**.
3. With **1-Var Stats** appearing on the **HOME** screen, press **2nd** then **1** to insert **L1** on the **HOME** screen. Then press the comma and press **2nd** and **2** to insert **L2** on the **HOME** screen. So, the **HOME** screen should have the following:
1-Var Stats L1, L2
Press **ENTER** to obtain the mean and the standard deviation.

TECHNOLOGY STEP-BY-STEP Determining Quartiles

TI-83/84 Plus

Follow the same steps given to obtain the mean and median from raw data.



Schlumberger

WHY WAIT FOR PROGRESS?

DARE TO DISCOVER

Discovery means many different things at Schlumberger. But it's the spirit that unites every single one of us. It doesn't matter whether they join our business, engineering or technology teams, our trainees push boundaries, break new ground and deliver the exceptional. If that excites you, then we want to hear from you.

careers.slb.com/recentgraduates

Excel

1. Enter the raw data in column A.
2. With the data analysis Tool Pak enabled, select **TOOLS** menu and highlight **Data Analysis...**
3. Select **Rank and Percentile** from the Data Analysis window.
4. With cursor in the **Input Range** cell, use the mouse to highlight the data in column A. Press **OK**.

TECHNOLOGY STEP-BY-STEP Drawing Boxplots Using Technology

TI-83/84 Plus

1. Enter the raw data in L1 by pressing **Stat** and selecting **1: Edit**.
2. Press **2nd Y =** to access Stat-Plots menu. Select **1: plot 1**.
3. Place the cursor on **"ON"** and press **ENTER**, to turn the plots on
4. Use the cursor to highlight the modified boxplot icon.
5. Press **ZOOM**, AND SELECT **9: Zoom Stat**.

Excel

1. Load the DDXL Add-in.
2. Enter the raw data in column A. If you are drawing side-by-side boxplots, enter all data in column A and use index values to identify which group the data belongs to in column B.
3. Select the DDXL menu and highlight Charts and Plots. If you are drawing a single boxplot, select "Boxplot" from the pull-down menu; if you are drawing side-by-side boxplot, select "Boxplots by Groups" from the pull-down menu.
4. Put the cursor in the "Quantitative Variable" window. From the names and Columns window, select the column of the data and click the < arrow. If you are drawing side-by-side boxplots, place the cursor in the "Group Variable" window. From the Names and Columns window, select the column of the indices and click the < arrow. If the first row contains the variable name, check the "First row is variable names" box. Click OK.

3 PROBABILITY

3.1 INTRODUCTION

Knowledge of the properties of theoretical probability distributions is an important part of the decision-making process in the various areas of the applied and basic sciences.

Let us look at an example in an applied area. A certain change in the technique of producing tranquilizer pills is predicted to decrease the average number of defective units per lot of 1,000 tranquilizers. In order to check out, or test, this prediction we need to know the expected number of defective units per lot, and the expected variability of defective units per day, under the present technique of production, in order to compare them with their companions under the new technique.

To make such a comparison, a number of samples are taken under both production techniques. Suppose that the quality control engineer, charged with this investigation, has provided the following data for the number of defectives:

	Production Technique	
	Old	New
Mean (defectives/lot)	80	60
Variance	400	225
Range	10 to 150 = 140	20 to 100 = 80

The visual inspection, of the descriptive statistics, suggests that the new technique is better, but the results are not really clear cut. While the mean and variance of the number of defectives, in the two samples, showed the predicted decrease, there was a certain amount of overlap in the two sample distributions. The new technique may not have been a chance of occurrence, and that, if repeated sampling were performed, the direction of the difference observed in the first sampling might only occur half of the time.

Again, we raise the question: What might occur in the long run, i.e. over repeated sampling? Furthermore, we need some rules in order to make a final decision as whether the new production method is better, or not, than the old one. An approach, which is commonly used, is to describe the characteristics of a distribution of the possible outcomes which assumes that no difference (between the two methods) exists. This distribution is often called

the **Null** (no difference) distribution. **The Null distribution is developed as a probability distribution which states the probability of obtaining every possible outcome of sampling assuming no difference.** If such a distribution is known, we could then state the probability of obtaining our particular sample if it were selected from the parent population which generated the null probability distribution. If our sample is a highly improbable outcome, given the null probability distribution, then we might conclude that the sample was drawn from some distribution other than the null distribution. In short, you would conclude that the new method very likely produces results different from the old method.

To make a decision like the one described above, we need to know the nature of the theoretical probability distribution which describes the probability of each possible event. In Chapter 4, we will discuss the basic rationale and characteristics of two very common probability distributions: The **Binomial** and the **Normal Probability Distributions**. Before initiating these discussions, we need to review some basic definitions and concepts of probability theory. What follows is a summary of the probability concepts and basic notion on set theory. We recommend that you start this summary as if you were seeing the material for the first time, and you need a lot of memorization.



PREPARE FOR A LEADING ROLE.

English-taught MSc programmes in engineering: Aeronautical, Biomedical, Electronics, Mechanical, Communication systems and Transport systems. No tuition fees.

→ liu.se/master

li.u LINKÖPING UNIVERSITY



3.2 PROBABILITY AXIOMS

We start this section with some definitions.

Definition 3.1: An Experiment is a process by which an observation (or a measurement) is obtained.

Definition 3.2: A Sample Space for an experiment, denoted by S [or Ω : the capital Greek letter Omega], is the set of all possible outcomes of that experiment.

EXAMPLE 3.1: Determine the sample space of the following experiments:

- i. Tossing a normal coin once.
- ii. Taking a test, as a student in any course.

Solution:

- i) The Sample Space, S , is given by $S = \{\text{Head, Tail}\} = \{H, T\}$
- ii) The Sample Space, S , consists of the following grades, $S = \{A, B, C, D, F, I, W, P\}$.
(Note on ii. Above. The sample space for that experiment could be displayed as {pass, fail})

Definition 3.3: An Event is any collection of outcomes from, or a subset of, the sample space of a probability experiment.

Events that contain one outcome are called simple events, while those with more than one outcome are called compound events. In general, events are denoted using capital letters such as E, F, G, H , etc.

EXAMPLE 3.2: A probability experiment consists of rolling a single fair die.

- i) Identify the outcomes of this experiment.
- ii) Determine the sample space.
- iii) Define the event $E = \text{“roll an even number.”}$

Solution:

- i) The outcomes from rolling a single fair die are $x_1 = \text{“rolling a one”} = \{1\}$, $x_2 = \text{“rolling a two”} = \{2\}$, $x_3 = \text{“rolling a three”} = \{3\}$, $x_4 = \text{“rolling a four”} = \{4\}$, $x_5 = \text{“rolling a five”} = \{5\}$, $x_6 = \text{“rolling a six”} = \{6\}$.

a five" = {5}, and x_6 = "rolling a six" = {6}. These events are all simple events, since none of them can be broken down any further.

- ii) The sample space, S, has six outcomes, as it appeared from part i). Thus $S = \{1, 2, 3, 4, 5, 6\}$. Another representation for S is given by: $S = \{\text{even, odd}\}$
- iii) The event E = "roll an even number" = {2, 4, 6}.



EXAMPLE 3.3:

- i) In tossing a fair normal coin, in Example 3.1, identify the simple events.
- ii) In rolling a fair die, in Example 3.2, identify the compound event.

Solution:

- i) The sample space here is the set $S = \{H, T\}$. Each outcome is a simple one. It cannot be broken any further.
- ii) The event E = "roll an even number" is a compound event.



Definition 3.4: The **Probability** of an outcome, in a sample space, is that chance or relative frequency, or the mathematical measure of the likelihood for that outcome (given a particular experiment) to occur.

If a sample space, of an experiment, consists of the following n sample points: x_1, x_2, \dots, x_n then we can assign the number p_i for the probability of the outcome x_i , and we write $P(x_i) = p_i, i = 1, 2, \dots, n$, on the condition that

- i) The probability of an event E, P(E), must be greater than or equal to 0 and less than or equal to 1, $0 \leq P(E) \leq 1$, or by the above notation we have $0 \leq p_i \leq 1$,
- ii) $P(\Omega) = 1$, or $P(S) = 1$
- iii) The sum of the probabilities of all the outcomes, in a probability experiment must be equal to 1. That is, if the sample space is given as $S = \{x_1, x_2, \dots, x_n\}$ then $P(S) = \sum_1^n p_i = 1$.

Definition 3.5: A Probability Model lists the possible outcomes of the experiment and the associated probability of each outcome.

It is displayed in a **Table**, by a **Graph**, or by a **Formula**. The probabilities, in any probability model, should satisfy the above three conditions.

As it was shown above, if E is an event, then the probability that E has occurred is the sum of the probabilities of the simple sample points that make the event E. Hence if $E = \{Y_1, Y_2, \dots, Y_n\}$, then $P(E) = P(y_1) + P(y_2) + \dots + P(y_n)$. The void set [the null, or the empty

set], is an event with no sample points in it. The empty set is a subset of every set even of itself. The symbol for an empty set is the Greek letter ϕ , (Phi).

If an event is impossible, (does not occur), the probability of that event is 0. If an event is a certainty, then its probability should be 1. By our setting, we can see that $P(\phi) = 0$. The sample space is a certainty. Any outcome will be there in S , and S , as an event, is always occurring.

3.2.1 TYPES OF PROBABILITY

From the definition of probability, we see that it deals with the long-term proportions with which a particular outcome will occur. Based on that, how can we determine probabilities of outcomes? For this purpose, we have the following types of probabilities:

1. **Empirical Probability:** It is the probability, of an event E , that can be approximated by the ratio of the number of times that E had occurred to the number of times that the experiment has been carried out. Thus

$$P(E) \approx \text{Relative Frequency of } E = (\text{frequency of } E) / (\text{number of Trials made}).$$

Click here to learn more

TAKE THE
RIGHT TRACK

Give your career a head start
by studying with us. Experience the advantages
of our collaboration with major companies like
ABB, Volvo and Ericsson!

Apply by
15 January

World class
research

www.mdh.se

MÄLARDALEN UNIVERSITY
SWEDEN

The probability obtained using the empirical approach is an approximate value for $P(E)$. If the experiment is repeated, more (or less), times the relative frequency will change.

2. **Classical Probability:** If an experiment has n equally likely outcomes and if the number of ways of an event, E , to occur is m , then

$$P(E) = (\text{Number of Ways that } E \text{ occurs}) / (\text{Number of possible outcomes in the experiment}) = m/n.$$

When computing probabilities by the classical method, the experiment is not needed to be actually performed. Applying the classical method for calculating probabilities requires that all the outcomes are equally likely to occur. In this case the outcomes of the experiment are equiprobable.

3. **Subjective Probability:** It is a probability obtained on the basis of a personal judgment, or earlier knowledge.

It is to be realized that subjective probabilities are completely acceptable and legitimate, and possibly the only method of assigning likelihood to an outcome, check the stock market prices, and their forecast.

3.3 OPERATIONS AND PROBABILITY CALCULATION ON EVENTS

Let A and B be any two events defined on the sample space S , or Ω , or U .

Definition 3.6: The Union of the two events, or sets, A and B , written $A \cup B$, is defined to be the set

$$A \cup B = A \text{ or } B = \{x \mid x \in A \text{ or } x \in B\},$$

where $x \in B$ means that x is an element of B , or x belongs to the set B . The shaded area in **Figure 1** represents the **Union** of the two sets A and B . In other words, the union of two sets is the collection of all the elements in the two sets without repeating the common elements between them.

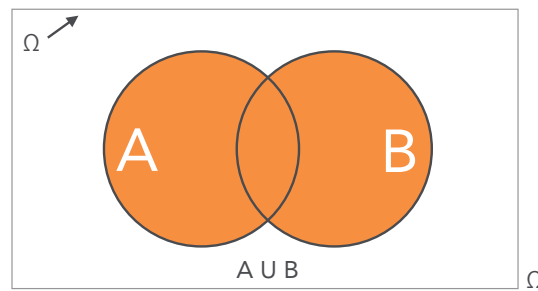


Figure 1 The Union of two sets

Definition 3.7: The Intersection of two events, or sets, A and B, written as $A \cap B$, is defined to be the set

$$A \cap B = A \text{ and } B = \{x \mid x \in A \text{ and } x \in B\}.$$

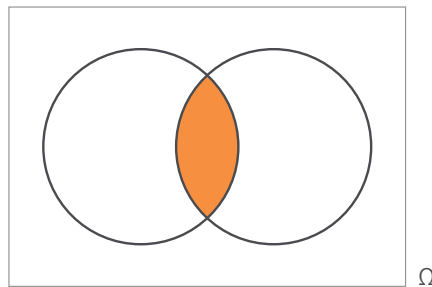


Figure 2 The Intersection of two sets

Thus, the intersection of two sets is the set of all the elements that are common between the two given sets. The shaded area in **Figure 2** represents the **Intersection** of the two sets A and B that were cited in Figure 1, above.

In case the intersection of the two sets is void, or it is the empty set, as shown in **Figure 3**, then the two sets are termed as disjoint, or mutually exclusive.

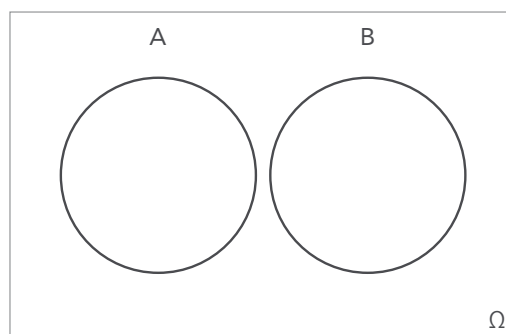


Figure 3 Two Mutually Disjoint Sets

For any two disjoint sets A and B, we have the **Addition Rule**, (think of the probability as the ratio of the area in the event to the area of the sample space, See **Figure 3**, when **Venn Diagrams** are being used to represent the sets, when A and B have nothing in common), then

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

The **General Addition Rule** states that:

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

In other words, we can write it in the following form

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Applying the same concept, as before, you see that the part between A and B has been taken into consideration twice, hence it is needed to take it out once.



How will people travel in the future, and how will goods be transported? What resources will we use, and how many will we need? The passenger and freight traffic sector is developing rapidly, and we provide the impetus for innovation and movement. We develop components and systems for internal combustion engines that operate more cleanly and more efficiently than ever before. We are also pushing forward technologies that are bringing hybrid vehicles and alternative drives into a new dimension – for private, corporate, and public use. The challenges are great. We deliver the solutions and offer challenging jobs.

www.schaeffler.com/careers

SCHAEFFLER



Definition 3.8: The Complement of a set A , written \bar{A} (A' , or A^c), is defined to be, check **Figure 4** below,

$$\bar{A} = \{x \mid x \notin A\},$$

where $x \notin A$ means that x does not belong to A , or x is not an element of the set A . In other words $x \notin A$ is the negation of $x \in A$.

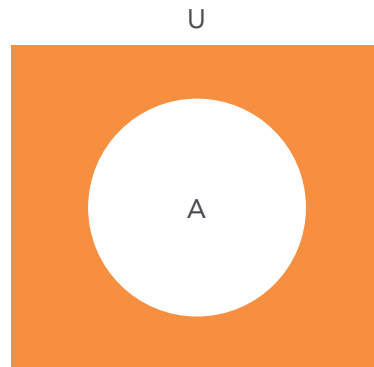


Figure 4 The Complement of A in the Universal Set U , S , or Ω

The Complement Rule: For any set, or event, E since $E \cup \bar{E} = S$, (or $E \cup \bar{E} = \Omega$), and $E \cap \bar{E} = \phi$, then $P(\bar{E}) = 1 - P(E)$.

This is based on $P(S) = 1$, (since S is the certain event, and always occurring) and with the **Addition Rule**, we have $1 = P(S) = P(E \text{ or } \bar{E}) = P(E) + P(\bar{E})$.

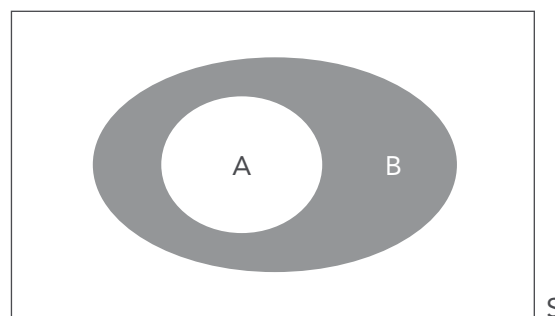


Figure 5 Sets and Subsets relation

Subsets introduce the notion of membership or containment, when one set is contained, or contains, another set. If every element of a set A is an element of another set B , then A is a subset of B , which is written as

$$A \subseteq B.$$

When A is a subset of B , and B is a subset of A then $A = B$, i.e.

$$A \subseteq B \text{ and } B \subseteq A, \text{ then } A = B$$

EXAMPLE 3.4: Consider the following sample space S , or as it is called, sometimes, the universal set, where $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Let $A = \{1, 2, 3, 4, 5, 6\}$, $B = \{3, 4, 5, 6, 7, 8, 9\}$, $E = \{2, 4, 6, 8, 10, 12\}$, $F = \{3, 6, 9, 12\}$, and $G = \{5, 7, 11\}$. Furthermore, assume that the elements in S are all equally likely to occur. Find

- i) $A \cup B$, and $P(A \cup B)$, ii) $A \cap B$, and $P(A \cap B)$, iii) $E \cap F$, and $P(E \cap F)$
 iv) E' , $P(E')$, v) $P(F \cup G)$.

Solution: Thus we can see, based on the definitions above, that

- i) $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Hence $P(A \cup B) = 9/12 = 0.75$
 ii) $A \cap B = \{3, 4, 5, 6\}$, and $P(A \cap B) = 4/12 = 1/3$.
 iii) $E \cap F = \{6, 12\}$, with $P(E \cap F) = 2/12 = 1/6$.
 iv) While $E' = \{1, 3, 5, 7, 9, 11\}$, we have $P(E') = 0.5$.
 v) $P(F \cup G) = P(F) + P(G) - P(F \cap G) = 4/12 + 3/12 - 0/12 = 7/12$.



We have already introduced the rules for some probability calculations, when we gave the definitions for the types of probabilities. Now we introduce some more rules for computing probabilities. It should not be a surprise if we say, in addition to the **General Addition Rule**, there are: **Multiplication or Product and Quotient Rules**. Really, do you believe that? Let us find out.

The **Addition Rule** above, as it has been seen, is involved with finding the probability of E or F . The emphasis here is on “or”, i.e. when one or the other will occur, or both will occur. Each probability rule is related to some kind of events. Before introducing the **Product Rule**, let us define what is meant by **Independent Events**.

Two events A and B are **Independent** if the occurrence of either one of them does not affect the occurrence of the other in a probability experiment. Thus, two events are termed dependent if the occurrence of one, in a probability experiment, affects the occurrence of the other. If you toss a fair coin twice, clearly the outcome on the first toss has nothing to do with the outcome on the second toss. Moreover, roll a die and you get a 1, has nothing to do with the outcome of the next roll. Recall the sample space of flipping a fair coin twice which consists of $S = \{HH, HT, TH, TT\}$, we find that the probability of 2 heads, i.e. $P(\{HH\}) = 1/4$, one out of four in the sample space. Remember also, we are tossing a

fair coin in such a way that $P(\{H\}) = \frac{1}{2}$, and $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ that implies $P(\{HH\}) = P(\{H\}) \cdot P(\{H\})$, and hence the **Product Rule**:

The **Product Rule** of probabilities: Two events E and F are **Independent** if and only if

$$P(E \text{ and } F) = P(E) \cdot P(F).$$

Otherwise E and F are **Dependent** events.

The above rule can be generalized to more than two events. It is to be noted that, in the above rule, if neither E nor F is the impossible event, then E and F cannot be mutually exclusive. Thus, **Independent** and **Mutually Exclusive** are two different notions, and thus they should not be taken as anonymous.

EXAMPLE 3.5: Let $P(E) = 0.6$ and $P(E|F) = 0.34$. Are E and F independent?

Solution: Clearly not, since $P(E|F)$ is not equal to 0.6. Thus, E has been affected by F after F has occurred.



**STUDY FOR YOUR MASTER'S DEGREE
IN THE CRADLE OF SWEDISH ENGINEERING**

Chalmers University of Technology conducts research and education in engineering and natural sciences, architecture, technology-related mathematical sciences and nautical sciences. Behind all that Chalmers accomplishes, the aim persists for contributing to a sustainable future – both nationally and globally.

Visit us on **Chalmers.se** or **Next Stop Chalmers** on facebook.

CHALMERS
UNIVERSITY OF TECHNOLOGY



For finding the “**quotient**” or the general product rule, let us ask a student this question: What is the chance of passing this class, with a high grade, if you do not study? Or, if you study what is the chance you pass this course, with a high grade? No doubt one event is “conditioned” on the other to occur. If it is not cloudy, at some time, the chance of rain, at that time is much smaller than when it is cloudy.

The symbol $P(E|F)$, read as the probability of E given F (or **the Conditional Probability of E given F**), gives the probability that the event E will occur provided F has occurred already. We conclude with this “**Quotient**” rule as described below.

If A and B are any two events, then

$$P(A|B) = P(A \text{ and } B) / P(B), \text{ (with } P(B) \neq 0), \text{ or}$$

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

Similarly, $P(B|A) = P(A \text{ and } B) / P(A)$, (with $P(A) \neq 0$), or

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

since (A and B) is the same as (B and A). In the above two versions for the rule, we should have $P(A) \neq 0$, and $P(B) \neq 0$, as shown. This is so, since the event has occurred already. It is NOT the impossible event. Hence the **General Product Rule**:

$$P(A \cap B) = P(A \text{ and } B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Let us give some examples on the above definitions.

EXAMPLE 3.6: Suppose that A and B are two events such that $P(A \text{ and } B) = 0.6$, and $P(A) = 0.8$. What is $P(B|A)$?

Solution: By applying the above formula, we find that $P(B|A) = 0.6/0.8 = 0.75$.



REMARK: Let the two events E and F be independent, then so are

- a) E^c and F b) E and F^c c) E^c and F^c . (Hint: $[E^c \cap F^c] = [E \cup F]^c$).

{The proofs of the above statements are left for the reader as exercises.}

3.4 BAYES FORMULA AND TOTAL PROBABILITY

The definition for the conditional probability, given in the above section, forms the basis for the **Bayes' Formula**. Suppose that we have the following sets (or events) A_1, A_2, \dots, A_n , of an experiment in such a way that the following conditions are satisfied:

- i) $P(A_i)$ is known for all $i=1, 2, \dots, n$;
- ii) $\cup A_i = S$, the sample space of the experiment; and
- iii) $A_i \cap A_j = \phi$ for all $i \neq j$.

Then these sets A_1, A_2, \dots, A_n will form what we call a **Partition** for the sample space S . It is worth noting that the sets A_1, A_2, \dots, A_n are mutually exclusive and exhaustive events, based on the properties listed above. Now given any other set (or event) B with the conditional probabilities, $P(B|A_i)$, (these are known to be the **Prior Probabilities**) are known for all $i = 1, 2, \dots, n$. Then we have the **Law of Total Probability**, presented as

$P(B) = \sum_{i=1}^n P(A_i) \cdot P(B|A_i)$, then the conditional probability $P(A_i|B)$, or The **Bayes' Formula**, is given by

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{P(B)} = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{i=1}^n P(A_i) \cdot P(B|A_i)}; \text{ for } i = 1, 2, \dots, n.$$

To show the Total Probability, we have: Since the sets A_1, A_2, \dots, A_n are forming a partition for the sample space, S , we can see that

$$B = B \cap S = \bigcup_{i=1}^n (B \cap A_i).$$

Thus $P(B) = P[\bigcup_{i=1}^n (B \cap A_i)] = \sum_{i=1}^n P(B \cap A_i)$, since the sets A_1, A_2, \dots, A_n form a partition of S . Hence, we have

$$P(B) = P[\bigcup_{i=1}^n (B \cap A_i)] = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(A_i) \cdot P(B|A_i).$$

The conditional probabilities, $P(A_i|B)$, for $i = 1, 2, \dots, n$, are called the **Posterior Probabilities**.

EXAMPLE 3.7: A certain factory has 3 lines of production; A, B, and C, with the proportions of production on the lines are: 0.40, 0.35, and 0.25 respectively. The percentages of defective, D, items produced on those lines are given by: 0.10, 0.05, and 0.01. An item was checked, and it is found to be defective. What is the probability that the defective item came from Line A; i.e. What is $P(A|D)$?

Solution: There is no doubt that the lines of production, A, B, and C for a partition for the factory. Thus, the items are produced with the following probabilities: therefore, probability that the item was produced by line A is $P(A) = 0.40$, and hence $P(B) = 0.35$, and $P(C) = 0.25$. Moreover, we have $P(D|A) = 0.10$, $P(D|B) = 0.05$, and $P(D|C) = 0.01$.

A defective item from this factory will come either from Line A, or Line B, or Line C, and as a factory set up, it will not come from two lines. Thus, the defective item will come either from $A \cap D$, or $B \cap D$, or $C \cap D$. Thus, Total probability formula applies, and we have

$$D = (A \cap D) \text{ or } (B \cap D) \text{ or } (C \cap D),$$

$$P(D) = P[(A \cap D) \text{ or } (B \cap D) \text{ or } (C \cap D)] = P(A \cap D) + P(B \cap D) + P(C \cap D),$$

$$P(D) = P(A) \cdot P(D|A) + P(B) \cdot P(D|B) + P(C) \cdot P(D|C)$$

$$= 0.4 \cdot (0.1) + 0.35 \cdot (0.05) + 0.25 \cdot (0.01) = 0.060.$$

$$\text{Thus } P(A|D) = P(AD)/P(D) = P(A) \cdot P(D|A) / P(D) = 0.4 \cdot (0.1) / 0.06 = 2/3.$$





Scholarships

Open your mind to new opportunities

With 31,000 students, Linnaeus University is one of the larger universities in Sweden. We are a modern university, known for our strong international profile. Every year more than 1,600 international students from all over the world choose to enjoy the friendly atmosphere and active student life at Linnaeus University. Welcome to join us!

Linnaeus University
Sweden

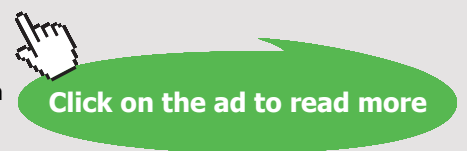


Ln.u.se

Bachelor programmes in
Business & Economics | Computer Science/IT | Design | Mathematics

Master programmes in
Business & Economics | Behavioural Sciences | Computer Science/IT | Cultural Studies & Social Sciences | Design | Mathematics | Natural Sciences | Technology & Engineering

Summer Academy courses



CHAPTER 3 EXERCISES

3.1 Match the proposed probability of the event E, column I with the appropriate description in column II

Probability	Description
a) 0.95	i) No Chance to occur
b) 0.02	ii) very likely to occur
c) 3.00	iii) Fifty fifty chance
d) -0.10	iv) very little chance to occur
e) 0.30	v) It may occur but not certain
f) 0.0	vi) An incorrect statement
g) 0.25	vii) the odds are 1 to 3
h) 0.35	viii) Occurs moderately often
i) 0.04	ix) E rarely occurs
j) 1.2	x) cannot not be a probability
k) 1.0	xi) sure to occur
l) 0.005	xii) very unlikely to occur

3.2 Consider an ordinary deck of 52 cards, and let S be the sample space of selecting a card from the deck. Assume that the probability set function assigns $1/52$ to each of the 52 outcomes. Let

$$E = \{x: x \text{ is a jack, a queen, or a king}\},$$

$$F = \{x: x \text{ is a 9, 10, or jack, and } x \text{ is red}\},$$

$$G = \{x: x \text{ is a club}\},$$

$$H = \{x: x \text{ is a diamond, a heart, or a spade}\}.$$

Find:

- | | | |
|------------------------------------|-------------------------------------|----------------------|
| i) $P(E)$, | ii) $P(F)$, | iii) $P(E \cap F)$, |
| iv) $P(E \cup F)$, | v) $P(G \cup H)$, | vi) $P(G \cap H)$, |
| vii) $P(E \cup F \cup G \cup H)$, | viii) $P(E \cap F \cap G \cap H)$. | |

3.3 There are 4 elementary outcomes in a sample space. If $P(E_1) = 0.2$, $P(E_2) = 0.5$, and $P(E_3) = 0.1$, what is the probability of E_4 ?

- 3.4 Four applicants will be interviewed for a certain position at the Big University. They have the following characteristics
1. Psychology major, male, GPA 3.5
 2. Chemistry major, female, GPA 3.3
 3. Sociology major, female, GPA 3.7
 4. Statistics major, male, GPA 3.8

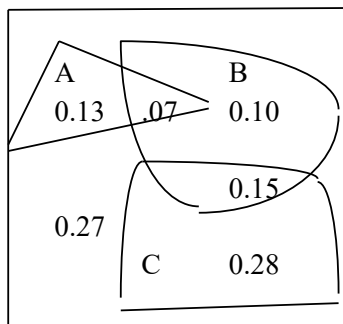
One of the candidates will be hired.

- a) Draw a Venn diagram and exhibit these events:
 - A: A social Science major is hired;
 - B: The GPA of the selected candidate is higher than 3.6.
 - C: A male candidate is hired.
- b) Give the composition of the events $A \cup B$, $A \cap B$, and $B \cap C$,

- 3.5 A letter is chosen at random from the word "TEXAS." What is the probability that it is a vowel?

- 3.6 If events A and B are such that $A \subset B$, then $P(A) \leq P(B)$.

3.7



The Above Venn Diagram shows three events A, B, and C and the probabilities of the various intersections. [For example, $P(AB) = 0.07$. Determine

- a) $P(A)$, and $P(AB^c)$
- b) $P(B)$
- c) $P(C)$
- d) $P(A \cup B)$
- e) $P(B \cap C^c)$

- 3.8 A white die and a red one are rolled. Identify the following events:
- $A = [\text{sum} = 6]$, $B = [\text{sum} = 7]$, $C = [\text{sum is even}]$, $D = [\text{same number on each die}]$
 - If both dice were fair, assign probabilities to each elementary outcome.
 - Find the probability for each of the events: A , B , C , and D ; based on how they were classified above.
- 3.9 An on-campus organization will select one of the week for an end-of-year picnic. Assume the days Monday–Friday are equally likely, and that each weekend day, Saturday and Sunday, is twice as likely as a weekday to be selected.
- Assign probabilities to the seven outcomes.
 - Find the probability a weekday will be selected.
- 3.10 A fair coin is tossed 4 times, and the outcomes in the sample space have been listed.
- List the outcomes of the sample space
 - Let A be the outcome of at least one head, find $P(A)$.
 - Let B be the outcome of at least one tail, find $P(B)$.
 - Let C be the outcome of 2 tails, find $P(C)$.



e-learning for kids

About e-Learning for Kids Established in 2004, e-Learning for Kids is a global nonprofit foundation dedicated to fun and free learning on the Internet for children ages 5 - 12 with courses in math, science, language arts, computers, health and environmental skills. Since 2005, more than 15 million children in over 190 countries have benefitted from eLessons provided by EFKI. An all-volunteer staff consists of education and e-learning experts and business professionals from around the world committed to making difference. eLearning for Kids is actively seeking funding, volunteers, sponsors and courseware developers; get involved! For more information, please visit www.e-learningforkids.org.

- The number 1 MOOC for Primary Education
- Free Digital Learning for Children 5-12
- 15 Million Children Reached

- e) Find $P(A \cup B)$
 f) Find $P(A \cap B)$
 g) Find $P(A \cup B \cup C)$
 h) Find $P(A \cap B \cap C)$
- 3.11 A fair eight-sided die is rolled once. Let $A = \{2, 4, 6, 8\}$, $B = \{3, 6\}$, $C = \{2, 5, 7\}$, and $D = \{1, 3, 5, 7\}$. Assume that all the faces are equiprobable.
- a. Calculate: i) $P(A)$, ii) $P(B)$, iii) $P(C)$, iv) $P(D)$.
 b. Give the values for i) $P(A \cap B)$, ii) $P(B \cap C)$, iii) $P(C \cap D)$.
 c. Give the values for i) $P(A \cup B)$, ii) $P(B \cup C)$, iii) $P(C \cup D)$, iv) $P(A \cup D)$.
 d. Give the values for i) $P(A \cap B \cap C)$, ii) $P(B \cap C \cap D)$, iii) $P(A \cap B \cap C \cap D)$.
 e. Give the values for i) $P(A \cup B \cup C)$, ii) $P(B \cup C \cup D)$, iii) $P(C \cup D \cup A)$.
 f. Give the value for $P(A \cup B \cup C \cup D)$.
- 3.12 If $P(A) = 0.4$, $P(B) = 0.6$, and $P(A \cap B) = 0.3$, find a) $P(A \cup B)$, b) $P(A \cap B')$, and c) $P(A' \cup B')$.
- 3.13 Given that $P(A \cup B) = 0.77$, and $P(A \cup B') = 0.87$, find $P(A)$.
- 3.14 Suppose that $P(A) = 0.68$, $P(B) = 0.55$, and $P(AB) = 0.35$. Find:
 a) $P(B|A)$; b) $P(B^c|A)$; c) $P(B|A^c)$.
- 3.15 Suppose that $P(A) = 0.40$, $P(B) = 0.50$, and $P(A \text{ or } B) = 0.70$. Find:
 a) $P(A|B)$; b) $P(A^c|B)$; c) $P(B|A^c)$.
- 3.16 Let $P(A) = 0.3$ and $P(B) = 0.6$. Find:
1. $P(A \cup B)$ when A and B are independent.
 2. $P(A|B)$ when A and B are mutually exclusive.
 3. $P(B|A)$ when A and B are mutually exclusive
- 3.17 A red die and a white die are rolled. Let event $A = \{4 \text{ on the red die}\}$ and event $B = \{\text{sum of dice is odd}\}$. Find: i) $P(A)$; ii) $P(B)$; iii) $P(A \cap B)$. iv) Are A and B independent?
- 3.18 Let $P(A) = 0.40$; $P(B) = 0.40$ and $P(A \cup B) = 0.70$. Answer the following questions:
 a) Are A and B independent? Why or why not?
 b) Can A and B be mutually exclusive? Why or why not?

- 3.19 Of the three events, A, B, and C, suppose that the events A and B are independent, while the events B and C are mutually exclusive, and their probabilities are $P(A) = 0.70$, $P(B) = 0.20$, and $P(C) = 0.30$, respectively. Express the following events in set notation and calculate their probabilities:
- Both B and C occur.
 - At least one of A or B occurs.
 - B does not occur.
 - All three events occur.
- 3.20 Suppose that $P(A) = 0.60$, $P(B) = 0.22$.
- Determine $P(A \cup B)$ if A and B are independent.
 - Determine $P(A \cup B)$ if A and B are mutually exclusive.
 - Find $P(A|B^c)$ if A and B are mutually exclusive.
- 3.21 In a certain region, 12% of the adult population is smokers, 0.80% are smokers with emphysema, and 0.20% are nonsmokers with emphysema.
- What is the probability that a person, selected at random, has emphysema?
 - Given that the selected person is a smoker, what is the probability that this person has emphysema?
 - Given that the selected person is not a smoker, what is the probability that this person has emphysema?
- 3.22 Let x equal a number that is selected randomly from the closed interval [0, 1]. Use your intuition and assign values to the following probabilities:
- $P(\{x: 0 \leq x \leq 1/3\})$;
 - $P(\{x: 1/3 \leq x \leq 1\})$;
 - $P(\{x: 1/2 < x < 5\})$.
- 3.23 If A, B, and C are any three events, prove that
- $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C).$$

3.24 For two events G and H, the following probabilities are given: $P(G) = 0.52$, $P(H) = 0.36$, $P(G \cap H) = 0.2$.

	G	G^c
H		
H^c		

- Enter these probabilities in the adjacent Table.
- Determine the probabilities of: GH^c , G^cH , and G^cH^c .
- Fill in the table.
- Express the following events in set notation and find their Probabilities:
 - G occurs and H does not occur,
 - Neither G nor H occurs,
 - Either G occurs or H does not occur

3.26 The following table classifies 1500 students by their gender and by whether or not they favor a gun law.

.....Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



3.27

	Male (S_1)	Female (S_2)	Total
Favor (F_1)	390	650	1040
Oppose (F_2)	280	180	460
Totals	670	830	1500

Compute the following probabilities if one of these 1500 students is selected randomly:

- a) $P(S_1)$
- b) $P(F_1|S_1)$
- c) $P(F_1|S_2)$
- d) Give an interpretation to your answer in b) and c).

3.28 Two cards are drawn successively and without replacement from an ordinary deck of cards. Compute the probability of drawing

- a) Two hearts.
- b) A heart on the first draw and an ace on the second draw.
- c) A heart on the first draw and a club on the second draw.

3.29 Suppose that $P(E) = 0.7$, $P(F) = 0.5$, and $P([E \cup F]') = 0.1$. Find

- a) Find $P(E \cap F)$.
- b) Calculate $P(E|F)$.
- c) Calculate $P(F|E)$.

3.30 In reference to Example 3.7, calculate, separately: i) $P(B|D)$, ii) $P(C|D)$; and check if $P(A|D) + P(B|D) + P(C|D) = 1$.

3.31 Concerning three events A, B, and C, the probabilities of the various intersections are given in the table below

	B		B^c	
	C	C^c	C	C^c
A	0.05	0.10	0.05	0.17
A^c	0.20	0.15	0.18	0.10

- a) Draw a Venn diagram to identify the intersections and mark their probabilities.
- b) Determine the following probabilities: $P(AB)$, $P(AC^c)$, $P(C)$.
- c) Fill in the table below



- 3.32 A doctor is concerned about the relationship between blood pressure and irregular heartbeats. Among her patients, she classifies blood pressures as high, Normal, or low and heartbeats as regular and irregular and finds that:
- a) 16% have High blood pressure;
 - b) 19% have low blood pressure;
 - c) 17% have an irregular heartbeat;
 - d) Of those with an irregular heartbeat, 35% have high blood pressure; and
 - e) Of those with normal blood pressure, 11% have an irregular heartbeat. What percentage of her patients has a regular heartbeat and low blood pressure?
- 3.33 Box B_1 contains two White chips; box B_2 contains two red chips; box B_3 contains two white and two red chips; and box B_4 contains three white chips and one red chip. The probabilities of selecting box B_1 , B_2 , B_3 , or B_4 are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{8}$ respectively. One of the boxes is selected using these probabilities and a chip is then drawn at random. Find:
- a) The probability of drawing a white chip, $P(W)$.
 - b) The conditional probability that Box B_1 had been selected, given that a white chip was drawn, $P(B_1|W)$.
 - c) The probability of drawing a red chip, $P(R)$.
 - d) The conditional probability that Box B_1 had been selected, given that a white chip was drawn, $P(B_2|R)$.

4 DISCRETE PROBABILITY DISTRIBUTIONS

4.1 INTRODUCTION

Let us begin with some definitions and terminology.

Definition 4.1: A Random Experiment is any process of measurements or observations, in which the outcome cannot be completely determined in advance.

Definition 4.2: The Sample Space, S , of any random experiment, is the total collection of all possible outcomes of the random experiment.

Each outcome has a certain probability (or chance) to occur. Thus, tossing a coin once is a random experiment. The sample space for this experiment is {Head, Tail}, or {H, T}. Observing the value of a certain stock in the market is a random experiment. The sample space is the set of all possible values that the stock might take.



Nido

Luxurious accommodation

Central zone 1 & 2 locations

Meet hundreds of international students

BOOK NOW and get a £100 voucher from voucherexpress

Nido Student Living - London

Visit www.NidoStudentLiving.com/Bookboon for more info.

+44 (0)20 3102 1060

Definition 4.3: A Random Variable is any real-valued quantity, or numerical measure, whose value depends on the outcomes of a random experiment. In other words; **A Random Variable** is a function from the sample space of the experiment to the set of all real numbers.

Given a random experiment with a sample space S , a function that assigns one and only one real number, $X(s) = x$ to each element s in S , is called a random variable. **Random Variables** are typically denoted, or identified, by capital **Roman** letters such as X , Y , and Z . The values that a random variable will take will be denoted by lower case letters x , y , and z , respectively. Thus, in tossing a coin once, the number of heads, X , that may appear is a random variable that will take the values $x = 0$ or $x = 1$. In **This Chapter**, we will discuss the first type of the random variables, namely **The Discrete Random Variable**. **Chapter 5** will have the presentation of the other type, **The Continuous Random Variable**.

Definition 4.4: A Discrete Random Variable is that random variable that has either a finite or a countable number of values. The values of a **Discrete Random Variable** can be plotted on the number-line with spaces between them.

For instance; the number of cousins you have, or the number of friends you know, the number of doors in your house, the number of cylinders in the engine of your car, are examples of discrete random variables.

In the discrete case, any random variable will take a certain value with a certain probability. The function that controls these probabilities is called the **Probability Mass Function**, or **PMF**. The value of the **PMF**, at a certain value of the discrete random variable, will give the probability that the random variable will take such a value, for example $P(X = x) = P(x)$.

Definition 4.5: For the $P(x)$ to be a **PMF**, for a discrete random variable, the following conditions should be satisfied:

1. $P(x) \geq 0$, for all values of the discrete random variable X ,
2. $\sum_{\text{all values of } x} P(x) = 1$.

In other words, (as it is shown in 2. above) the sum of the probabilities of all the outcomes, in a probability experiment must be equal to 1. In addition to the **PMF**, there is another important concept, or function, called the **Cumulative Distribution Function**, or **CDF**.

Definition 4.6: The CDF is defined all over the real line, and for the discrete random variable it is given by: $F(k) = P(X \leq k) = \sum_{\text{all values of } x}^k P(x)$, where k is an integer.

In addition to the above definitions of the **PMF** and **CDF**, there is another concept that needs to be given here. That concept is for the **Probability Distribution**.

Definition 4.7: A **Probability Model**, or a **Probability Distribution**, is a table or a formula, in the case of a discrete random variable, that depicts the values of the random variable and their associated probabilities.

EXAMPLE 4.1: Consider the experiment of counting the number of heads when tossing a fair coin twice.

Solution: As we have assumed earlier, the sample space has the following outcomes: {HH, HT, TH, TT}. Counting the heads in each outcome we find that our discrete random variable will take the values $x = 0, 1, 2$. Hence we have the following discrete probability distribution

X	0	1	2
P(x)	0.25	0.50	0.25

It is seen that, in a probability distribution, the associated probabilities are satisfying the conditions stated above, for each being non-negative, and they add up to one.



4.2 THE EXPECTATION AND VARIANCE OF A RANDOM VARIABLE

Given a set of measurements, we are interested in the average of the data, and in how much variation there is. For example, the average income and the variation in the incomes of a group of employees at a certain university, the average of a set of measurements on the content of iron in any compound, and the variability in the readings of a certain apparatus.

The concept of “**Average**” is referred to as the **Expected Value**, or the **Mean** value of the random variable. The concept of **Variation** is referred to as the **Variance**, or the **Standard Deviation** of a random variable. In some cases, when data is available, as it will be seen later, the **Range** could be used to indicate the variation in the data.

Definition 4.8: The **Mean**, or **The Expected Value**, of a discrete random variable is given as:

$$E(X) = \mu_X = \sum_{\text{all values of } x} [x.P(x)],$$

where x is the value that can be assumed by the random variable and $P(x)$ is the associated probability with that value of X .

It is seen that $E(X)$ can be considered as a **Weighted Average** of the values of the random variable X with the associated probabilities, of those values, being the weights, especially in the discrete case.

NOTE 1: $E(X)$ is called the first moment, or the mean, of the random variable X .


Definition 4.9: The Variance of a random variable, denoted by $V(X)$ or $\text{Var}(X)$ or σ_X^2 , and based on the expected value of a random variable that we defined above, is given by

$$\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2,$$


with μ_X is the mean, as defined above, and $E(X^2)$ is given by

$$E(X^2) = \sum_{\text{all values of } x} [x^2 \cdot P(x)].$$

[The mathematically equipped student can show that $E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$.]


SIMPLY CLEVER


WE WILL TURN YOUR CV INTO AN OPPORTUNITY OF A LIFETIME



Do you like cars? Would you like to be a part of a successful brand?
As a constructor at ŠKODA AUTO you will put great things in motion. Things that will ease everyday lives of people all around Send us your CV. We will give it an entirely new new dimension.

Send us your CV on
www.employerforlife.com



NOTE 2: $E(X^2)$ is called the second moment of the random variable X .

NOTE 3: $E(X^r)$ is called the r^{th} moment of the random variable X , when r is a positive integer.

It can be clearly seen that the variance is the average of the squared deviations of the values of X from their mean μ_X . The variance is a quantity that reflects the extent to which the random variable is close to its mean. The variance, as sometimes called, is the second central mean of the random variable. However, the variance, as it could be checked, is expressed in square units of the measurements that the data is expressed in. Based on that, it looks as if it is needed to express the variation in terms of the units that the data are measured in. Thus, we have the following definition.

Definition 4.10: The Standard Deviation, denoted by σ_X , or **STD(X)**, is the positive square root of the variance, i.e., the **Standard Deviation** of a random variable = $\sigma_X = +\sqrt{\sigma_X^2}$.

Let X be a random variable of the discrete type with **PMF** $p(x)$. If there is a positive number h and each of the following expectations exist, and being finite for $-h < t < h$, i.e.

$$E(e^{tX}) = \sum_{\text{all values of } x} [e^{tX} \cdot P(x)],$$

Then, the following definition is needed.

Definition 4.11: The function, **M(t)**, defined by $M(t) = E(e^{tX})$, is called the **Moment-Generating Function, MGF**, of X . This function will be used to find all moments of the random variable X , as it will be seen later.

EXAMPLE 4.2: Which is of the following Tables represent a discrete probability distribution?

a)	X	P(x)	b)	X	P(x)	c)	X	P(x)
	1	0.20		1	0.20		1	0.20
	2	0.35		2	0.25		2	0.25
	3	0.12		3	0.10		3	0.10
	4	0.40		4	0.14		4	0.15
	5	-0.07		5	0.49		5	0.30

Solution:

a) The Table does not represent a discrete probability distribution since $P(5) = -0.07$, which is less than 0.

b) The Table does not represent a discrete probability distribution since

$$\sum_{\text{all values of } x} P(x) = 0.2 + 0.25 + 0.10 + 0.14 + 0.49 = 1.18 \neq 1.$$

c) In this case, the Table describes a discrete probability distribution because the sum of the probabilities equals 1, and each probability is greater than or equal 0 and less than or equal 1.



EXAMPLE 4.3: In **EXAMPLE 4.2**, part (c) was shown to be a discrete probability distribution, find:

- a) The mean,
- b) The variance, and
- c) The standard deviation.

Solution:

a) $E(X) = \mu_X = \sum_{\text{all values of } x} [x \cdot P(x)] = 1(0.20) + 2(0.25) + 3(0.10) + 4(0.15) + 5(0.30) = 3.10$

b) $\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$. In this case, we will use the computational form of the variance rather than the definition form, i.e.,

$\sigma_X^2 = E(X^2) - \mu_X^2$, rather than $\sigma_X^2 = E[(X - \mu_X)^2]$. Based on that we see

$$E(X^2) = \sum_{\text{all values of } x} [x^2 \cdot p(x)] = 1(0.2) + 4(0.25) + 9(0.10) + 16(0.15) + 25(0.30) = 12.$$

$$\sigma_X^2 = E(X^2) - \mu_X^2 = 12 - (3.10)^2 = 2.39.$$

c) $\sigma_X = \sqrt{2.39} = 1.54596$.



4.2.1 SOME PROPERTIES OF THE EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

In this subsection, based on the definitions of the expected value and the variance, of a discrete random variable that were given above, we have the following properties, or formulas, on the expected value of a random variable.

Theorem 4.1: In all of the following properties, a , b , and c will be constant real numbers.

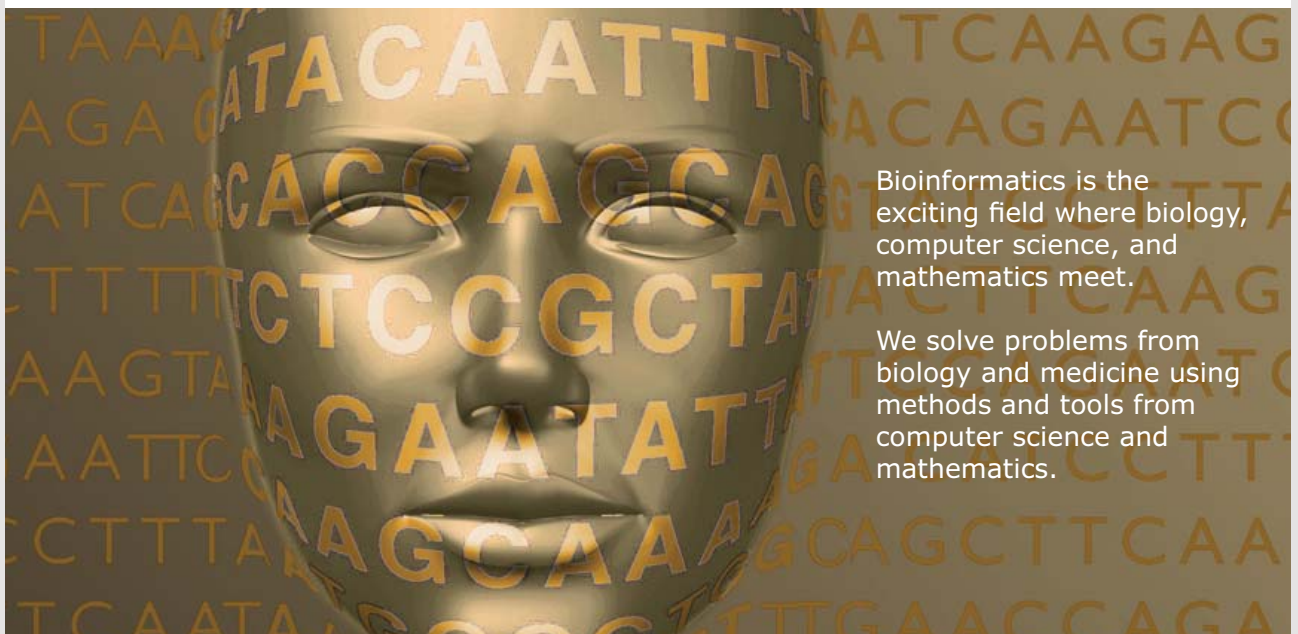
1. $E(c) = c$.
2. $E(X + b) = E(X) + b$
3. $E(cX) = cE(X)$.
4. $\text{Var}(X + c) = \text{Var}(X)$.
5. $\text{Var}(cX) = c^2 \cdot \text{Var}(X)$.

(The proofs, for the above **Theorem 4.1**, are left as exercises for the interested reader)



UPPSALA
UNIVERSITET

Develop the tools we need for Life Science Masters Degree in Bioinformatics



Bioinformatics is the exciting field where biology, computer science, and mathematics meet.

We solve problems from biology and medicine using methods and tools from computer science and mathematics.

Read more about this and our other international masters degree programmes at www.uu.se/master



4.3 DISCRETE PROBABILITY DISTRIBUTIONS

As it was shown above, there are two types of random variables, and thus there are two types of probability distributions; namely the discrete and continuous probability distributions. Probability distributions, in statistics, are used as models which are simplified versions of some real-life phenomena. In the discrete case, there are six probability distributions that are quite referred to very often. Those distributions are: The **Bernoulli**, the **Binomial**, the **Poisson**, the **Geometric**, the **Hyper Geometric**, and the **Negative Binomial** distributions.

We start with the **Bernoulli** distribution.

4.3.1 BERNOULLI PROBABILITY DISTRIBUTION

It is quite often that an experiment will have only two outcomes. Such an experiment is called a **Bernoulli trial**, after **Jacob Bernoulli**. (See **Hogg and Tanis (2010)** for the history of the Bernoulli family). Situations where the elements of a population can be classified into a dichotomy exist in all aspects of science and nature. To check on that, let us list some examples:

Inspect an item of production coming from a factory, and you can classify that item as defective or not.

On the first day of any semester, on your campus, ask the students in that class:

Did you vote in the last presidential election or not?

Did you have milk this morning?

Do you have a calculator for this statistics class?

Do you like Statistics?

Is your car domestic made?

Checking on the above questions we find ourselves having yes or no, most of the time, as an answer. In addition to that situation, if one student had milk in the morning for breakfast, this has nothing to do with any other student that will have milk for breakfast that morning, thus having independent outcomes. Hence selecting a single element of a population will be considered as an experiment, and thus each trial will produce one of two

possible outcomes. For the sake of clarity and use, those two outcomes for any **Bernoulli trial** will be labeled as Success, S, or failure, F. The simplest example of a Bernoulli trial is tossing a coin, where, again, the occurrences, head or tail can be labeled as S or F, respectively. Thus, for a fair, and normal, coin we assign the probability of $p = \frac{1}{2}$ for a success and $q = \frac{1}{2}$ for a failure. Let us summarize the conditions under which an experiment will be labeled as a **Bernoulli trial**.

For an experiment to be labeled as a **Bernoulli trial**, the following conditions should be satisfied:

1. Each trial gives one of two outcomes, practically (or theoretically) called Success (S) and Failure (F).
2. For each trial, the probability of success, $P(S)$, is the same and it is denoted by $p = P(S)$. The probability of failure is then $P(F) = 1 - P(S) = 1 - p$ for each trial. This probability will be denoted by q . Hence $p + q = 1$.
3. Trials will be independent. In essence, the probability of success does not change throughout the experiment even if given any information about the outcomes of the other trials.

Let X be the random variable that is associated with a **Bernoulli trial**. Thus, X as a function from the sample space to the set of real numbers, can be defined as follows:

$$X(\text{Success}) = 1, \text{ and } X(\text{failure}) = 0.$$

Based on this definition, we see that the **PMF** of X can be presented by:

$$P(x) = p^x(1 - p)^{1-x}, x = 0, 1.$$

Thus, X will have a Bernoulli distribution. As any other discrete random variable, we have, for the expected value:

$$E(X) = \mu_X = \sum_{\text{all values of } x} [x \cdot P(x)] = 0(1-p) + 1(p) = p.$$

For the variance, we have: $\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$, with

$$E(X^2) = \sum_{\text{all values of } x} [x^2 \cdot p(x)] = 0(1-p) + 1(p) = p.$$

Hence

$$\sigma_X^2 = E(X^2) - \mu_X^2 = p - p^2 = p(1-p) = pq.$$

Clearly, from the above value of the variance, we have the standard deviation to be given by

$$\sigma_x = \sqrt{pq}.$$

THEOREM 4.2: The **MGF** of a Bernoulli random variable X is given by: $M(t) = 1 - p + pe^t$, where t is any real number, and p is the probability of success.

PROOF: Applying the definition of the **MGF** on the Bernoulli random variable, we see that

$$M(t) = E(e^{tx}) = \sum e^{tx} \cdot P(x), \text{ for } x = 0, 1. \text{ Carrying on the summation we see that } M(t) = 1 - p + pe^t.$$

As promised earlier, using the above **MGF** we can find the moments of X , for any order, just by calculating the derivative of the **MGF** and find its value at $t = 0$. Doing that we have $M'(t) = pe^t$, where $M'(t)$ is the derivative of $M(t)$ with respect to t . Thus, when $t = 0$, we will get $M'(0) = p = E(X) = \mu_x$. Similarly, $M''(t) = pe^t$, with $M''(t)$ being the second derivative of $M(t)$, and $M''(0) = p = E(X^2)$. Hence,

UNIVERSITY OF COPENHAGEN



Copenhagen Master of Excellence

Copenhagen Master of Excellence are two-year master degrees taught in English at one of Europe's leading universities

Come to Copenhagen - *and aspire!*

Apply now at
www.come.ku.dk

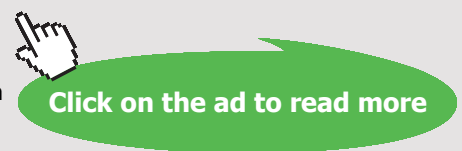




cultural studies

religious studies

science



$\sigma_x^2 = E(X^2) - \mu_x^2 = M''(0) - [M'(0)]^2 = p - p^2 = p(1-p) = pq$, as was previously found.

REMARK: The aforementioned technique, can, and will be used to find the moments of other random variables as well.

4.3.2 BINOMIAL PROBABILITY DISTRIBUTION

A **Bernoulli trial** is a random experiment that has only two outcomes. Those two outcomes are mutually exclusive, i.e. cannot happen at the same time, and they are labeled as S, for a success, or F, for a failure, with constant probability of success as p , i.e. $P(S) = p$, and that of a failure as q , or $P(F) = q$, with $p + q = 1$. A sequence of such an experiment is said to have a **Binomial Probability Distribution** if the following conditions are satisfied:

1. There is a fixed number, n , of **Bernoulli** independent trials. This means that the outcome of one trial will not affect the outcomes of the other trials.
2. The probability of success, p , is constant throughout the experiment, as well as that of a failure, q , thus we have $p + q = 1$.
3. We are interested in the number of successes, X , as a result of those n **Bernoulli Trials**, regardless how they will occur. Hence, we see that X will take on the values of $x = 0, 1, 2, \dots, n$.
4. The probability of obtaining x successes, in those n independent **Bernoulli Trials**, is given by the following pmf

$$P(X = x) = P(x) = {}_n C_x p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n.$$

With ${}_n C_x$ denoting the number of combinations of x items taken from n distinct items without paying attention to order, or it can be set like $\binom{n}{x}$, and p is the probability of success. In such a case, we write $X \sim \text{Bin}(n, p)$; i.e., the random variable X has a **Binomial Distribution** with the number of trials being n , and the probability of success is p . The notation will be denoted by **$X \sim \text{Bin}(n, p)$** , in other words, **X has a Binomial Distribution with the parameters n and p .**

The values for the above **PMF**, based on different values for n and p , are tabulated as a **CDF** values in **Table III, Appendix A, The Binomial Distribution Table**. The **CDF** values for the Binomial **$X \sim \text{Bin}(n, p)$** are given by

$$F(k) = P(X \leq k) = \sum_{\substack{\text{all values of } x=0 \\ \text{to } k}} P(x).$$

EXAMPLE 4.4: From clinical trials, it is known that 20% of mice inoculated with a serum will not develop protection against a certain disease. If 10 mice were inoculated, find

- a) The probability of at most 3 mice will contract the disease.
- b) The probability that exactly 5 mice will contract the disease.

Solution: Let X be the number of mice contracting the disease. Then $X \sim \text{Bin}(10, 0.2)$, that is, X has a Binomial Distribution with $n = 10$ and $p = 0.2$. Thus, we have

- a) $P(X \leq 3) = F(3) = 0.8791$, from **Table III** for cdf of the Binomial Distribution, with $n = 10$, $p = 0.2$, and $x = 3$.

Using Technology, namely, $P(X \leq 3) = \text{Binomcdf}(n, p, x) = \text{Binomcdf}(10, 0.2, 3) = 0.8791$.

- b) $P(X = 5) = P(5) = P(X \leq 5) - P(X \leq 4) = 0.9936 - 0.9672 = 0.0264$, by using **Table III**.

Using Technology, namely, $P(X = 5) = P(5) = \text{Binompdf}(n, p, x) = \text{Binompdf}(10, 0.2, 5) = 0.0264$



4.3.2.1 The Mean and Variance of a Binomial Random Variable

We discussed finding the mean (or expected value) and the standard deviation of a discrete random variable in Section 4.2. Those formulas can be used to find the mean (or expected value) and the standard deviation for the binomial random variable as well due to their importance and frequent use.

A binomial experiment with n independent Bernoulli trials and probability of success p has a mean and standard deviation given by the formulas:

$$E(X) = \mu_X = \sum_{\text{all values of } x} [x.P(x)] = np, \text{ and}$$

$$\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2 = np(1-p), \text{ with}$$

$$\sigma_x = +\sqrt{\sigma_x^2} = \sqrt{np(1-p)}.$$

(The proofs of the above expressions are left as exercises for the mathematically interested student. Look up the formula for how to expand a binomial term, and you can derive the above formulas.)

EXAMPLE 4.5: According to the Federal Communications Commission, 75% of all U.S. households have cable television in 2004. In a simple random sample of 300 households, determine the mean and standard deviation for the number of households that will have cable television.

Solution: This is a binomial experiment since the conditions set above are satisfied. So, in this case we have $n = 300$, and $p = 0.75$. We can use the formulas for the mean and the standard deviation for a binomial random variable to reach at

$$\mu_x = np = 300(0.75) = 225, \text{ and}$$

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{300(0.75)(0.25)} = 7.5$$



Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF



Constructing a binomial probability histogram is no different from constructing other probability histograms. The values of the random variable X will be, definitely, the classes, and the probabilities of those values, as we listed them as relative frequencies, are the heights of the bars in the histogram. In order to see the effect of n and p , on the distribution, check the Example 4.6.

EXAMPLE 4.6: Given the following binomial probability distributions, construct a probability histogram for each case.

- a) $n = 10$, and $p = 0.2$
- b) $n = 10$, and $p = 0.5$
- c) $n = 10$, and $p = 0.8$

Solution: To construct the binomial histogram, sure we need to have the classes defined and the associated probabilities calculated for each value. No doubt, the values that the random variable will take in each of the cases above are: $x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, and 10 . Based on these values and the associated probabilities with them we have the following tables:

a) X	0	1	2	3	4	5	6	7	8	9	10
P(x)	0.1074	0.2684	0.3020	0.2013	0.0881	0.0264	0.0055	0.0008	0.0001	0.0000	0.0000
b) X	0	1	2	3	4	5	6	7	8	9	10
P(x)	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010
c) X	0	1	2	3	4	5	6	7	8	9	10
P(x)	0.0	0.0	0.0001	0.0008	0.0055	0.0264	0.0881	0.2013	0.3020	0.2684	0.1074

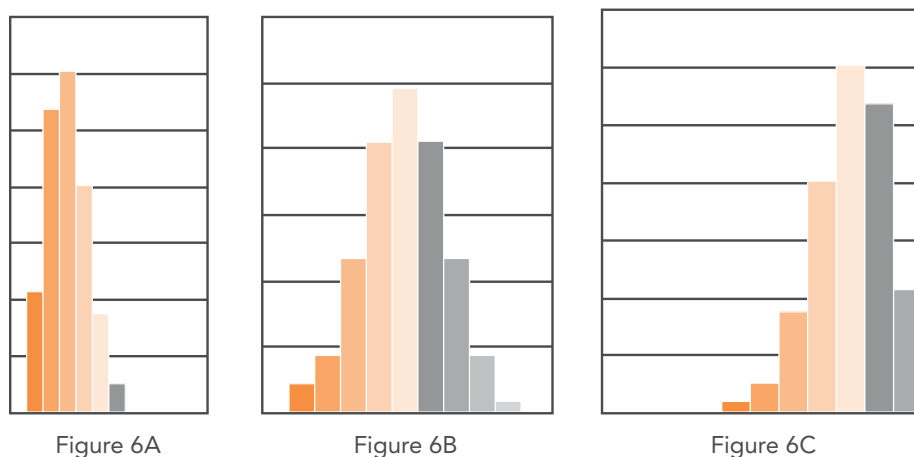
REMARK 1: It is to be noted that the probabilities in Part a) and part c) in the tables are reversed, and that is expected. This is true since calculating the probability of 6 successes equals the probability of calculating 4 failures, when p and q have been interchanged.

REMARK 2: For part b) the probabilities are symmetric about the center value of 5, since $p = 0.5$, check the table for this part.

The graphs below show the binomial histograms for the data in a), b) and c) respectively with the probabilities expressed as a percentage to make the graph little bigger and look better.

a) Histogram $n = 10$, $p = 0.2$ b) Histogram $n = 10$, $p = 0.5$ c) Histogram $n = 10$, $p = 0.8$

Using the results in **Example 4.6**, we have the tendency to conclude that the binomial probability distribution is skewed to the right if $p < 0.5$, symmetric and approximately bell-shaped if $p = 0.5$, and skewed to the left if $p > 0.5$, with n being held constant for those variable values of p , the probability of success in a **Bernoulli** trial.



The binomial probability distribution depends on two parameters, namely the number of trials, n , in the binomial experiment and the probability of success, p . Holding n constant, we have seen how the changes in the p value affect the shape of the distribution, and determine which side is skewed to. Now, what will happen if we kept p constant, and let n varies? In other words, what role does n play in the shape of the distribution? To answer this question, we compare the following sets of distributions for the fixed value of $p = 0.2$, and $n = 10, 30$, and 70 .

For the case $n = 10$ and $p = 0.2$, we will have **Figure 6A**. For $n = 30$, and $p = 0.2$ (by using technology to graph the binomial histogram we can see that the graph is slightly skewed to the right, while for $n = 70$, and $p = 0.2$, we will have what appears to be a bell-shaped histogram. We can come to this conclusion:

As the number of trials, n , in a binomial experiment, increases the histogram for the probability distribution of the random variable X becomes approximately bell-shaped. As a rule of thumb, if $np(1-p) \geq 10$, the probability distribution for a binomial random variable will be approximately bell-shaped.

THEOREM 4.3: If $X \sim \text{Bin}(n, p)$, i.e. X has a binomial distribution with n Bernoulli trials with p being the probability of success, then the **MGF** is given by: $M(t) = (1 - p + p e^t)^n$.

COROLLARY 4.3.1: By referring to the **MGF** for a binomial random variable X , and using the techniques depicted earlier, show that $\mu_X = np$, and $\sigma_X = \sqrt{np(1-p)}$.

ADDENDUM: The **Bernoulli** experiment that has only two outcomes, and once it is repeated n independent times, it generates the Binomial distribution. This experiment can be extended to be a trinomial distribution. In this case, we have three mutually exclusive and exhaustive outcomes for the experiment. This is clearly possible, and been in action, especially in industry. Any product can be classified as: “**Perfect**”, “**Seconds**”, and “**Defective.**” We assume that the experiment can be repeated n independent times, and the probabilities for Perfect, Second, and Defective, as p_1, p_2, p_2 , and $1 - p_1 - p_2$ remain constant from trial to trial. If we denote by X_1, X_2 , and $n - X_1 - X_2$, the number of perfect items, the number of seconds and the defectives in n trials, with x_1, x_2 as nonnegative integers such that $x_1 + x_2 \leq n$, then the probability of having x_1 perfect items, x_2 seconds, and $n - x_1 - x_2$ defectives, in this order, is given by

$$p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}.$$

Now, the question is raised in how many ways can we achieve that? Clearly, then $P(X_1 = x_1, X_2 = x_2) = P(x_1, x_2)$ is given by

Trust and responsibility

NNE and Pharmaplan have joined forces to create NNE Pharmaplan, the world’s leading engineering and consultancy company focused entirely on the pharma and biotech industries.

Inés Aréizaga Esteva (Spain), 25 years old
Education: Chemical Engineer

– You have to be proactive and open-minded as a newcomer and make it clear to your colleagues what you are able to cope. The pharmaceutical field is new to me. But busy as they are, most of my colleagues find the time to teach me, and they also trust me. Even though it was a bit hard at first, I can feel over time that I am beginning to be taken seriously and that my contribution is appreciated.



NNE Pharmaplan is the world’s leading engineering and consultancy company focused entirely on the pharma and biotech industries. We employ more than 1500 people worldwide and offer global reach and local knowledge along with our all-encompassing list of services.
nnepharmaplan.com

nne pharmaplan®



$$P(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} \cdot p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2},$$

with $x_1 + x_2 \leq n$. It can be easily seen that $X_1 \sim \text{Bin}(x_1, p_1)$ and similarly for $X_2 \sim \text{Bin}(x_2, p_2)$. In addition to the claim that was just made, we see that $P(x_1) \cdot P(x_2) \neq P(x_1, x_2)$, and this shows that X_1 and X_2 are dependent.

4.3.3 POISSON DISTRIBUTION

Consider a sequence of random events such as: radioactive disintegration during a unit time interval, incoming phone calls at a telephone switchboard during lunch hour, the number of traffic accidents at a certain intersection during a week, the number of messages you sent to friends per week, and the number of misprints per page in a book. Each of the above events displays an experiment that results in counting the number of times the particular events occur at a given time or with a given physical object. That count or measure, X , can be looked at as a random variable of a random phenomenon over a continuous medium, and this counts to define a **Poisson Distribution**. The **Poisson Distribution** is named after **Simeon Denis Poisson**. This distribution has the following **PMF**.

$$P(X = x) = P(x) = \lambda^x e^{-\lambda} / x!, \quad x = 0, 1, 2, \dots,$$

= 0; otherwise,

where the constant, $\lambda > 0$ (the Greek letter lambda) represents the average number (or the density of events) per one unit of measurement (unit of time, unit of length, or unit of area, or unit of volume). Table IV has the probabilities for the **Poisson distribution**. In addition to the above conditions on the Poisson distribution, the following conditions need to be satisfied:

1. The numbers of changes, or events, occurring in non-overlapping intervals are independent.
2. The probability of exactly one change occurring in a sufficiently short, or small, interval of length h is approximately λh .
3. The probability of two or more changes occurring in a sufficiently short interval is essentially zero.

(It is left, as an exercise, for the student to show that the above $P(x)$ satisfies the conditions for a **PMF**.)

REMARK: If X is the number of successes in an interval of length t , the pmf for X is given by

$$P(X = x) = P(x) = (\lambda t)^x e^{-\lambda t} / x!, x = 0, 1, 2, 3, \dots$$

Moreover, applying the formulas for the mean and the variance of a discrete random variable, it is easily verified that the mean and the variance, in this case, are equal, and each is equal to λ . (See something EXTRA below.)

EXAMPLE 4.7: Consider the number of accidents between 8 and 9 am on an intersection on Saturday. From data recorded let the mean of accidents on that intersection have a mean of 4. Hence this follows what we call a Poisson distribution with $\lambda = 4$. Find the probability that on a given Saturday, between 8 and 9 am, there will be:

- a) No accident,
- b) At least one accident,
- c) Exactly 4 accidents.

Solution:

- a) From Table IV, we have, with $\lambda = 4$ and $x = 0$, and using the above formula for $P(x)$ that gives the value $P(0) = 0.018$.
- b) $P(\text{at least 1 accident}) = P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0) = 1 - P(X = 0) = 1 - 0.018 = 0.982$.
- c) $P(X = 4) = 4^4 e^{-4} / 4! = 0.1954$. This can be done using Table IV, Poisson distribution, as $P(X = 4) = P(X \leq 4) - P(X \leq 3) = 0.6289 - 0.4335 = 0.1954$.



Something EXTRA

Since $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots$, we see that

$$e^{-\lambda} = \sum_{i=0}^{\infty} \frac{(-\lambda)^i}{i!} = 1 - \frac{\lambda}{1!} + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \dots$$

$$\mu = E(X) = \sum_{x=0}^{\infty} x \cdot p(x) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{x-1}}{(x-1)!} = \lambda.$$

$$E(X^2) = E[X(X-1)] + E(X)$$

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) \cdot \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{x-2}}{(x-2)!} = \lambda^2$$

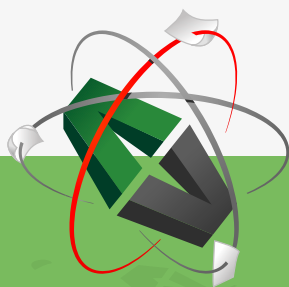
$$\sigma^2 = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

THEOREM 4.4: Let X have a Poisson probability distribution with parameter $\lambda > 0$, show that the **MGF** of X is given by $M(t) = e^{\lambda} \cdot (e^t - 1)$, and find the mean and variance of X by the aforementioned technique.

4.4 MORE DISCRETE RANDOM VARIABLES

In the subsections below we will be presenting, and discussing more discrete random variables that appear in some experiments in real life. It is quite worth it, for the reader to acquaint himself with the foundations and more important the applications of such random variables.

This e-book
is made with
SetaPDF



PDF components for PHP developers

www.setasign.com



4.4.1 GEOMETRIC DISTRIBUTION

Let us consider an experiment in which the conditions are like those for a Binomial one, and based on independent **Bernoulli** trials. In other words, let us have an experiment where there are two outcomes only, namely success and failure, with $P(S) = p$, and that of a failure $P(F) = q$, with $p + q = 1$. Consider playing the lotto, by buying one ticket, and you keep playing by one ticket until you win the jackpot. No doubt once you become a millionaire you would not be greedy any more, and you quit. There is an analogy between the Binomial probability distribution and the new coming one. The number of trials for the new distribution is not fixed ahead of time as it was the case in the **Binomial distribution**. We are interested in the number, X , of trials needed to get the first success, with clear understanding that the trials are independent. Based on these assumptions, we see that the following function is a **PMF** for a discrete probability distribution called the **Geometric Distribution**;

$$P(X = x) = p \cdot q^{x-1} \quad x = 1, 2, 3, \dots$$

$$= 0, \text{ otherwise.}$$

As it was with the **Poisson distribution**, the above function satisfies the conditions for a **PMF**.

(The proof is left to the reader, with remembering the sum of an infinite geometric series with ratio < 1 .)

EXAMPLE 4.8: In a certain producing process, it is known that, on the average, 1 in every 100 items is defective. What is the probability that the fifth item inspected is the first defective item found?

Solution: Using the above formula for the Geometric distribution with $x = 5$ and $p = 0.01$, we have $P(5) = (0.01) (0.99)^4 = 0.0096$.



EXAMPLE 4.9: During the busy time of the day, a telephone exchange is near capacity, so people cannot find a line to use. It may be of interest to know the number of attempts necessary in order to gain a connection. Suppose that we let $p = 0.05$ to be the probability of a connection during the busy time period. We are interested in knowing the probability that 5 attempts are needed for a successful call.

Solution: As above, by using the formula for the geometric distribution with $x = 5$ and $p = 0.05$ we find that $P(5) = (0.05) (0.95)^4 = 0.041$



THEOREM 4.5: a) Due to their importance and use, the mean and variance of a random variable following the **Geometric distribution**, are given by:

$$\mu = 1/p \quad \text{and} \quad \sigma^2 = (1 - p)/p.$$

b) The **MGF** for a random variable with **Geometric distribution** is $M(t) = pe^t/(1 - e^t)$.

(The proofs for **Theorems 4.4** and **4.5** are left, as an exercise, to the interested reader.)

4.4.2 HYPER GEOMETRIC DISTRIBUTION

There is a distinct difference in sampling for the **Binomial** and **Hyper Geometric** distributions. In the binomial case, the sampling is done with replacement in order to keep the probability of a success as a constant throughout the experiment. On the other hand, the sampling for the **Hyper Geometric** is done without replacement, and thus the repeated trials are not independent. This kind of sampling will affect the probability of a success.

Applications for the **Hyper Geometric** distribution are found in many areas and fields, with heavy uses in acceptance sampling, Clinical Trials, electronic testing, and quality assurance, as examples. Clearly, in these instances the item chosen is destroyed and cannot be put back in the sample. For the **Hyper Geometric** experiment, we consider a finite population of size N , which is composed of two categories, good and defective, for instance, with the number of good items denoted by R , and thus the number $N - R$, of the remaining items, will make the number of defectives. In general, we are interested in the number of successes, X , selected from those R items and with $n - x$ failures selected from $N - R$, with n being our sample size selected from the population of size N . This is known as **Hyper Geometric** experiment that has the following two properties:

1. A random sample of size n is selected without replacement from N distinct items.
2. R of the N items are classified as successes and $N - R$ are classified as failures.

The number X of successes that we are interested in, in this experiment, is labeled as a **Hyper Geometric** random variable. Based on the setting, we had so far, we find that the **PMF** for X is given by

$$P(X = x) = P(x) = \frac{\binom{R}{x} \binom{N-R}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2 \dots n;$$

with the following conditions: $0 \leq x \leq \min(n, R)$, $0 \leq n - x \leq N - R$.

Checking the above formula, we find that it is based on the multiplication rule of finding the number of ways of doing things when we have more than one option. That number clearly appears in the numerator and denominator of the above formula, respectively. Hence, the probability is the ratio of those two numbers based on the definition of the **Classical Probability Rule**.

EXAMPLE 4.10: A class, in Statistics 1315, has 25 students, 15 males and 10 females. A committee of 5 students is needed to be appointed to handle the class decisions, what is the **PMF** for the number, X , of females on the committee?

Solution: Per our settings, we see that $N = 25$, $n = 5$, $R = 10$, and $N - R = 15$. Hence the **pmf** is given by

$$P(X=x) = P(x) = \frac{\binom{10}{x} \binom{15}{5-x}}{\binom{25}{5}}, \quad x = 0, 1, 2, \dots, 5.$$

The above PMF for X can be expressed in a tabular form as follows:

X	0	1	2	3	4	5
---	---	---	---	---	---	---



Sharp Minds - Bright Ideas!

Employees at FOSS Analytical A/S are living proof of the company value - First - using new inventions to make dedicated solutions for our customers. With sharp minds and cross functional teamwork, we constantly strive to develop new unique products - Would you like to join our team?

FOSS works diligently with innovation and development as basis for its growth. It is reflected in the fact that more than 200 of the 1200 employees in FOSS work with Research & Development in Scandinavia and USA. Engineers at FOSS work in production, development and marketing, within a wide range of different fields, i.e. Chemistry, Electronics, Mechanics, Software, Optics, Microbiology, Chemometrics.

We offer
A challenging job in an international and innovative company that is leading in its field. You will get the opportunity to work with the most advanced technology together with highly skilled colleagues.

Read more about FOSS at www.foss.dk - or go directly to our student site www.foss.dk/sharpminds where you can learn more about your possibilities of working together with us on projects, your thesis etc.

The Family owned FOSS group is the world leader as supplier of dedicated, high-tech analytical solutions which measure and control the quality and production of agricultural, food, pharmaceutical and chemical products. Main activities are initiated from Denmark, Sweden and USA with headquarters domiciled in Hillerød, DK. The products are marketed globally by 23 sales companies and an extensive net of distributors. In line with the corevalue to be 'First', the company intends to expand its market position.



Dedicated Analytical Solutions

FOSS
Slangerupgade 69
3400 Hillerød
Tel. +45 70103370
www.foss.dk





P(x)	0.0565	0.2569	0.3854	0.2372	0.0593	0.0047
------	--------	--------	--------	--------	--------	--------



THEOREM 4.6: The mean and the variance for the **Hyper Geometric Random Variable** X are given by

$$\mu = nR / N, \text{ and } \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{R}{N} \cdot \left(1 - \frac{R}{N}\right).$$

(Try to compare the above mean and variance for the **Hyper Geometric Distribution** to those of the Binomial when n is very small compared to N, i.e., $(N - n) / (N - 1) \approx 1$, and R/N is taken as the probability of Success.)

4.4.3 NEGATIVE BINOMIAL DISTRIBUTION

As it was with the **Binomial** distribution, we are again appealing to the **Bernoulli Trial**, that is famous for its two only outcomes, a success with probability p, and a failure with probability q = 1 - p. Our interest, in this case, lies in the number of trials, X, as our **Negative Binomial Random Variable**, to produce r successes, when the order in which they had occurred does not matter here. If we were very lucky, we can get r successes in the first r trials, and that suggests that the number of trials needed, to get r successes, is at least r. Once we got the rth success, on the xth trial, we see that there had been r - 1 successes, and x - r failures in the first x - 1 trials. Since the trials are independent, we can multiply all the probabilities corresponding to each desired outcome. Therefore, the probability for the specified order, ending in the rth success, is

$$P^{r-1} q^{x-r} p = p^r q^{x-r}, \quad x = r, r+1, r+2, \dots$$

The total number of sample points in the experiment is found to be $\binom{x-1}{r-1}$, with each having the above probability. Thus, we have the general formula for the **PMF** of the **Negative Binomial distribution** to be given by

$$P(X = x) = P(x) = \binom{x-1}{r-1} \cdot p^r q^{x-r}, \quad x = r, r+1, r+2, \dots$$

THEOREM 4.7:

- a) The mean and variance of a random variable, following the **Negative Binomial distribution**, are given by:

$$\mu = r/p \text{ and } \sigma^2 = r(1-p)/p.$$

- b) The mgf is $M(t) = \frac{(pe^t)^r}{(1-qe^t)^r}$,

(The proofs are left as an exercise to the interested reader. The interested student can set a comparison between the **Geometric** and the **Negative Binomial distribution**, to see that one is a special case of the other.)

EXAMPLE 4.11: Find the probability that a person, flipping a fair coin, gets

- a) The third head on the seventh trial;
b) The first head on the fourth trial.

Solution:

- a) Using the **Negative Binomial Distribution** with $x = 7$, $r = 3$, and $p = 0.5$, we find that

$$P(7) = \binom{6}{2} (0.5)^7 = 0.1172.$$

- b) Again, by applying the **Negative Binomial Distribution** with $x = 4$, $r = 1$, and $p = 0.5$, we find that

$$P(4) = \binom{3}{0} (0.5)^4 = 1/16.$$



4.5 BIVARIATE RANDOM VARIABLES

We have discussed, in this chapter and the ones before it, as it has been noticed, the organization and the summary, graphically and numerically, of data, of a population, concerning a single variable. Observations, or distributions, on two or more quantitative variables are often recorded and found in real life, and in many applications of sciences. It is not rare for an investigator to face a situation when the need arises for finding data on

ordered pairs, or matched data, on two, or more random variables that are tight up with a pmf or a pdf. It is of at most interest to check on the high school rank and the score on the ACT or SAT tests in order to “predict” the GPA of a prospective athletic student before being given a scholarship. In such a case, we are trying to have a relationship among the three variables involved: namely, the high school rank, X ; the ACT (or SAT) score, Y ; and the GPA, Z . In other words is there a relationship among those variables that can be put as $z = u(x, y)$, for some range values on x and y ? In **Chapter 10**, we will discuss the dependence of one random variable, the predicted, on another variable, the predictor that has been assigned few values. In that presentation, the linear relationship between an input and output quantities will be addressed in full details under the topic of **Simple Linear Regression**. Moreover, in **Chapter 11**, the topic on multivariate regression will be introduced to study the relationship among more than two variables. In **Chapter 12**, we will study some cases on qualitative data that relate two or more random variables. In this **Chapter**; the following subsection will introduce, and present the case on how two discrete random variables can be related and get to study the behavior based on their probability mass functions. The continuous case, based on two continuous random variables will be presented in **Chapter 5**. We refer the interested reader to check on the volumes of univariate and multivariate distributions written by **Johnson, N.L. and Kotz, S., (1969–1972)**.

“I studied English for 16 years but...
...I finally learned to speak it in just six lessons”
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

The other subsection, of Section 4.5, will be devoted to study the conditional probability distributions of two discrete random variables.

4.5.1 BIVARIATE DISCRETE RANDOM VARIABLES

As it has been described before, random variables can be classified as discrete or continuous. In this section we will address the case when the two random variables are discrete and have a joint probability mass function, $p(x, y)$. Thus, we have to adopt and introduce some concepts and terminology in order to explore the various aspects of those two random variables.

Let X and Y denote the two discrete random variables that are of interest to us, and let S be the set of all ordered pairs (x, y) in such a way when $X = x$ and $Y = y$, i.e., let $S = \{(x, y) \mid X = x, \text{ and } Y = y\}$ to represent their common sample space. Based on this notation, we write the probability of $X = x$ and $Y = y$ as $P(x, y) = P(X = x, Y = y)$. Thus the function $P(x, y)$ will be used to denote the probability mass function, or the joint pmf, of X and Y , with the following two properties conditioned that for all (x, y) values in the sample space S , we have:

1. $0 \leq P(x, y) \leq 1$ and
2. $\sum_x \sum_y P(x, y) = 1$.

We will let the set of values that X will assume be S_1 , and similarly S_2 will represent the set of values that Y will take. Using this notation, we have the following defined marginal probability mass functions for X and Y , respectively, as presented by

- A. $P_x(X = x) = P_x(x) = \sum_y P(x, y)$, for all values of X in S_1 , and
- B. $P_y(Y = y) = P_y(y) = \sum_x P(x, y)$, for all values of Y in S_2 .

$P_x(x)$ and $P_y(y)$ are called the marginal pmf of X and Y , respectively. They are called “marginal” since they will appear in the margins for the table displaying the joint probabilities for the pairs (x, y) , see **Example 4.23** below. In addition to the above notation for the pmf for X and Y , if $P(x, y) = P_x(x) \cdot P_y(y)$, for all values of X and Y , then X and Y are independent; otherwise X and Y are said to be dependent. This is in line when we defined two independent events A and B . Recall the joint probability of A and B , written $P(A \text{ and } B)$ or $P(A \cap B)$, and we write $P(A \cap B) = P(A) \cdot P(B)$ if and only if A and B are independent.

As it was the case in **Section 4.2**, above, once the pmf of the random variable X, (or Y), has been defined, the mean μ_x , and the variance σ_x^2 can be calculated as before. Thus, let us make the definitions, cited herein more meaningful and clearer by giving an example.

EXAMPLE 4.23: Let the joint pmf of X and Y be defined by: $P(x, y) = (x + y)/21$, for $x = 1, 2, 3$ and $y = 1, 2$; and 0 otherwise. a) Give the joint table of the probabilities for the pairs (x, y). b) Find the marginal pmf for each of X and Y.

Solution:

		X			
		1	2	3	P_y
Y	1	2/21	3/21	4/21	9/21
	2	3/21	4/21	5/21	12/21
P_x		5/21	7/21	9/21	1.000

The Table above displays the joint probabilities of X and Y on their corresponding values. For Example: $2/21 = P(1, 1) = P(X = 1 \text{ and } Y = 1)$, and so on. The values of P_y , appearing in the margin of the above table, and along the values of Y, represent the probabilities to the corresponding values of Y. Thus $P(Y = 1) = P_y(1) = 9/21 = 3/7$, and $P(Y = 2) = P_y(2) = 12/21 = 4/7$. Therefore, the pmf of Y, and X can be displayed as:

Y	1	2
P_y	3/7	4/7

X	1	2	3
P_x	5/21	7/21	9/21

As it is very clear, from the above, we see that P_y and P_x are the pmf for the random variable Y, and X respectively.



NOTE: The above pmf for either X, or Y can be derived by applying either the property A. or B. above, and we have: $P_x(X = x) = P_x(x) = \sum_y P(x, y) = (x+1)/21 + (x + 2)/21 = (2x + 3)/21$, for $x = 1, 2, 3$; and 0 otherwise. Similarly, $P_y(Y = y) = P_y(y) = \sum_x P(x, y) = (1 + y)/21 + (2 + y)/21 + (3 + y)/21 = (6 + 3y)/21$, for $y = 1, 2$; and 0 otherwise.

Moreover, we have

$\mu_x = 1 \cdot 5/21 + 2 \cdot 7/21 + 3 \cdot 9/21 = 46/21$, and $E(X^2) = 1^2 \cdot 5/21 + 2^2 \cdot 7/21 + 3^2 \cdot 9/21 = 114/21$,
hence $\sigma_x^2 = E(X^2) - \mu_x^2 = 114/21 - (46/21)^2 = (21 \cdot 114 - 46^2)/441 = 278/441$.

REMARK: In **Section 4.3**, we have dealt with discrete random variables, and we derived the means and variances of those random variables, as well as the mgfs for some of them. Can we do the same here when we have two random variables that are jointly having a pmf? The answer is yes we can, and in the discussion below we will illustrate how that can be done. So, let us start with the expected value of two discrete random variables X and Y with $P(x, y)$ representing their pmf. Thus, we can write

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y) \cdot P(x, y) = \sum_x \sum_y x \cdot P(x, y) + \sum_x \sum_y y \cdot P(x, y) \\ &= \sum_x x \cdot [\sum_y P(x, y)] + \sum_y y \cdot [\sum_x P(x, y)] \\ &= \sum_x x \cdot P_x(x) + \sum_y y \cdot P_y(y) \\ &= E(X) + E(Y) \end{aligned}$$

In the case when we have one random variable, we talked about the expected value, or the mean, and the variance of that random variable. The situation is different when we have

The Wake
the only emission we want to leave behind

Low-speed Engines Medium-speed Engines Turbochargers Propellers Propulsion Packages PrimeServ

The design of eco-friendly marine power and propulsion solutions is crucial for MAN Diesel & Turbo. Power competencies are offered with the world's largest engine programme – having outputs spanning from 450 to 87,220 kW per engine. Get up front! Find out more at www.mandieselturbo.com

Engineering the Future – since 1758.
MAN Diesel & Turbo



two random variables. There is something extra connecting the two random variables. This extra concept that relates the two random variables X and Y is the Covariance of X and Y , written as **Cov** (X , Y), and it is defined as

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

It can be clearly (The proofs are left to the interested reader) seen that

$$\text{Cov}(X, Y) = \text{Cov}(Y, X) = E(XY) - E(X)E(Y) = E(XY) - \mu_x \cdot \mu_y, \text{ and } \text{Cov}(X, X) = \sigma_x^2.$$

EXAMPLE 4.24: Calculate Cov (X , Y), based on **Example 4.23**.

Solution: From Example 4.23, we have $\mu_x = 46/21$. To find Cov (X , Y), we need $E(Y)$, $E(XY)$. Thus we have: $E(Y) = 11/7$, $E(XY) = 1 \cdot 1 \cdot 2/21 + 2 \cdot 1 \cdot 3/21 + 3 \cdot 1 \cdot 4/21 + 1 \cdot 2 \cdot 3/21 + 2 \cdot 2 \cdot 4/21 + 3 \cdot 2 \cdot 5/21 = 24/7$. Hence Cov (X , Y) = $(504 - 506)/147 = -2/147$.



THEOREM 4.8:

- a) Cov (X , Y) = 0, if and only if X and Y are independent random variables.
- b) Var ($aX + bY$) = $a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab \text{Cov} (X, Y)$.
- c) If the **Correlation Coefficient** $\rho(X, Y)$, of X and Y , is defined by $\rho(X, Y) = \text{Cov}(X, Y)/(\sigma_x \cdot \sigma_y)$, then $-1 \leq \rho(X, Y) \leq 1$.

4.5.2 CONDITIONAL PROBABILITY OF TWO RANDOM VARIABLES

We have discussed conditional probability of two events earlier in **Chapter 3**. We did not need to specify the type of the event as discrete or continuous, since events were subsets, or parts, of the common sample space of the experiment. In this section, we like to explore the conditional probabilities for two discrete random variables that are jointly related by a **probability mass function, PMF**. As was the setup in the above two subsections of this chapter for the discrete, we need to introduce the notion for the conditional probability for two random variables. Recalling the earlier notation,

$$P(A|B) = P(A \text{ and } B)/P(B), \text{ provided } P(B) \neq 0.$$

For our discussion here we will adopt the following notation: Let A be the event such that $A = \{X = x\}$, and event $B = \{Y = y\}$, where (x, y) is a point in the joint space of the discrete random variables X and Y . Building on this convention, we see

$A \cap B = \{X = x, Y = y\}$, and $P(A \cap B) = P(X = x, Y = y) = P(x, y)$, with $P(B) = P_y(y) > 0$.

Therefore, the above conditional probability can be expressed as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(x, y)}{P_y(y)}, \text{ with } P_y(y) > 0.$$

Thus, we have the following definition:

The conditional probability mass function of the discrete random variables X , given $Y = y$, is defined as

$$P(x|y) = P(x, y) / P_y(y), \text{ when } P_y(y) > 0.$$

By the same convention, we see that the conditional probability mass function of Y , given $X = x$, is

$$P(y|x) = P(x, y) / P_x(x), \text{ when } P_x(x) > 0.$$

Let us check, using **EXAMPLE 4.23**, on how we can derive the conditional pmf for X or Y . Since X and Y , being discrete random variables, each can take a finite number of values. Thus there will be as many conditional pmf's of X on Y as there are values for Y . By the same token, there will be as many conditional pmf's of Y as there are values for X .

EXAMPLE 4.25: Let the joint pmf of X and Y be defined by: $P(x, y) = (x + y)/21$, for $x = 1, 2, 3$ and $y = 1, 2$; and 0 otherwise. a) Find the conditional pmf of X on $Y = y$, b) Find the conditional pmf of Y on $X = x$.

Solution: In the **NOTE**, above, we have: $P_x(X = x) = P_x(x) = \sum_y P(x, y) = (2x + 3)/21$, for $x = 1, 2, 3$; and 0 otherwise. Similarly, $P_y(Y = y) = P_y(y) = \sum_x P(x, y) = (6 + 3y)/21$, for $y = 1, 2$; and 0 otherwise. Using this part we can write $P(x|Y = y) = P(x, y) / P_y(y) = (x + y)/(6 + 3y)$, for $x = 1, 2, 3$ and $y = 1, 2$; and 0 otherwise. Since there are two values that Y can assume, therefore there are two conditional pmfs of X on $Y = y$. These pmfs are as follows:

For $y = 1$, $P(x | Y = 1) = (x + 1)/9$, $x = 1, 2, 3$.

For $y = 2$, $P(x | Y = 2) = (x + 2)/12$, $x = 1, 2, 3$.

Similarly, since X can have 3 values, that there are 3 conditional pmfs of Y on $X = x$. Following what was done, the conditional pmfs of Y on $X = x$, can be found from the following:

$$P(y | X = x) = P(x, y) / P_x(x) = (x + y) / (2x + 3), \text{ for } x = 1, 2, 3 \text{ and } y = 1, 2.$$

Thus we can write; for $x = 1$, $P(y | X = 1) = (1 + y) / 5$, for $y = 1, 2$, and 0 otherwise.

For $x = 2$, $P(y | X = 2) = (2 + y) / 7$, for $y = 1, 2$; and 0 otherwise.

For $x = 3$, $P(y | X = 3) = (3 + y) / 9$, for $y = 1, 2$; and 0 otherwise.

Guess how many $P(y | X = x)$ are there?



CHAPTER 4 EXERCISES

- 4.1 An oil exploration firm finds that 5% of the test wells it drills yield deposit of natural gas. If it drills 6 wells, find the probability that at least one well will yield gas.

gaiteye[®]
Challenge the way we run

**EXPERIENCE THE POWER OF
FULL ENGAGEMENT...**

**RUN FASTER.
RUN LONGER..
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY
WWW.GAITEYE.COM**

4.2 A medical research suggests that 20% of the general population suffer adverse side effects from a new drug. If a doctor prescribes the drug for 4 patients, what is the probability that: a) None will have side effects, b) All will have side effects. c) At least one will have side effects. d) Exactly 2 will have side effects. e) Find the expected number of patients that will have side effects.

4.3 Determine whether the distribution is a discrete probability distribution. If not, state why.

a)

X	0	1	2	3	4
$P(x)$	0.2	0.2	0.2	0.2	0.2

b)

X	100	200	300	400	500
$P(x)$	0.25	0.25	0.25	0.25	0.25

c)

X	1	2	3	4	5
$P(x)$	0	0	0	0	1

4.4 Determine the required value of the missing probability to make the distribution a discrete probability distribution

a)

X	3	4	5	6
$P(x)$	0.4	?	0.1	0.2

b)

X	0	1	2	3	4	5
$P(x)$	0.3	0.15	?	0.2	0.15	0.05

4.5 Consider the **Exercise 4.4**, after finding the missing probability for part a), find
a) the mean,
b) The variance, and
c) The standard deviation.

4.6 A random variable X has a binomial distribution with mean 6 and variance 3.6. Find $P(X=4)$.

4.7 A box contains 18 balls, of which there are 7 red and 11 white. A ball is drawn at random from the box. Let $X = 1$ if a white ball is drawn, and let $X = 0$ if a red ball is drawn. Give the **PMF**, the mean, and the variance of X .

4.8 Suppose that in **Exercise 4.7**, $X = 1$ if a red ball is drawn and $X = -1$ if a white ball is drawn. Give the **PMF**, the mean, and the variance of X .

4.9 An insurance company finds that 0.005% of the population dies from a certain kind of accident each year. What is the probability that the company must pay off no more than 3 of 1000 insured risks against such accidents in a given year? (Hint use the Poisson approximation to the binomial distribution with $np = \lambda$.) This

approximation is satisfactory whenever n is large and p value is near 0 or 1. If p is near 0.50 and n is large then the normal distribution is used to approximate the binomial distribution with mean = np , and variance = npq .

- 4.10 Find the probability of the indicated event if $P(E) = 0.25$ and $P(F) = 0.45$. Find
- $P(E \text{ or } F)$ if $P(E \text{ and } F) = 0.15$.
 - $P(E \text{ and } F)$ if $P(E \text{ or } F) = 0.6$.
 - $P(E \text{ and } F)$ if E and F are mutually exclusive.
 - $P(E \text{ or } F)$ if E and F are mutually exclusive.

- 4.11 Which is of the following tables represent a discrete probability distribution?

a)	<u>X</u>	<u>P(x)</u>	b)	<u>X</u>	<u>P(x)</u>	c)	<u>X</u>	<u>P(x)</u>
	1	0.2		1	0.3		1	0.2
	2	0.35		2	0.25		2	0.25
	3	0.12		3	-0.18		3	0.10
	4	0.40		4	0.14		4	0.15
	5	0.07		5	0.49		5	0.30

- 4.12 Consider the Tables in 4.11, and pick up the one that represents a discrete probability distribution. Then find, for that distribution, a) The mean, b) The Variance, c) The Standard deviation.

- 4.13 Suppose in families with 4 children, with only single birth that the probability of having 0, 1, 2, 3, or 4 boys are respectively: $1/16$, $4/16$, $6/16$, $4/16$, and $1/16$. Find

- The expected number of boys in a family of four children,
- The variability in the number of boys in a family of 4 children.

- 4.14 Consider a binomial probability distribution with parameters $n = 5$ and $p = 0.2$.

- Construct a binomial probability distribution with these parameters,
- Compute the mean and standard deviation of the distribution,
- Draw a probability histogram and comment on the shape, and label the mean on the histogram.

- 4.15 It is believed that approximately 65% of Americans under the age of 65 have private health insurance. Suppose this is true, and let X be the number of Americans under age 65 in a random sample of $n = 15$ with private health insurance.
- How X is distributed?
 - Find the probability that X is at least 10,
 - Find the probability that X is at most 10,
 - Find the probability that X is 10,
 - Find the mean, variance, and standard deviation of X .
- 4.16 A boiler has four relief valves, the probability that each opens properly is 0.90. Find:
- the probability that at least one opens properly,
 - the probability that all four open properly.
- 4.17 It is claimed that for a particular lottery, 1/10 of the 50 million tickets will win a prize. What is the probability of winning at least one prize if you purchase, a) 10 tickets, b) 15 tickets?

**Technical training on
WHAT you need, *WHEN* you need it**

At IDC Technologies we can tailor our technical and engineering training workshops to suit your needs. We have extensive experience in training technical and engineering staff and have trained people in organisations such as General Motors, Shell, Siemens, BHP and Honeywell to name a few.

Our onsite training is cost effective, convenient and completely customisable to the technical and engineering areas you want covered. Our workshops are all comprehensive hands-on learning experiences with ample time given to practical sessions and demonstrations. We communicate well to ensure that workshop content and timing match the knowledge, skills, and abilities of the participants.

We run onsite training all year round and hold the workshops on your premises or a venue of your choice for your convenience.

**For a no obligation proposal, contact us today
at training@idc-online.com or visit our website
for more information: www.idc-online.com/onsite/**

**OIL & GAS
ENGINEERING**

ELECTRONICS

**AUTOMATION &
PROCESS CONTROL**

**MECHANICAL
ENGINEERING**

**INDUSTRIAL
DATA COMMS**

**ELECTRICAL
POWER**

Phone: +61 8 9321 1702
Email: training@idc-online.com
Website: www.idc-online.com

**IDC
TECHNOLOGIES**

- 4.18 The random variable X follows a **Poisson** process with $\lambda = 4$. Find each of the following:
- $P(3)$,
 - $P(X < 3)$,
 - $P(X \leq 3)$,
 - $P(X \geq 4)$, e) $P(3 \leq X \leq 5)$, f) The mean, the variance, and the standard deviation of X ?
- 4.19 Explain the role of λ and t in the **Poisson** probability formula. (Remember that λ is the average hits per unit of measurement for the Poisson process.)
- 4.20 The flaws in a piece of timber occur at the rate of 8% per a linear foot. The random variable X is the number of flaws in the next 20 linear feet of timber. What is λ ? What is t ? How many flaws will be there?

In exercises 4.21 and 4.22, the random variable X follows a Poisson process with the given mean.

- 4.21 Assuming $\lambda = 5$, compute:
- $P(6)$,
 - $P(X < 6)$,
 - $P(X \geq 6)$,
 - $P(2 \leq X \leq 4)$.
- 4.22 Assuming $\lambda = 7$, compute:
- $P(10)$,
 - $P(X < 10)$,
 - $P(X \geq 10)$,
 - $P(7 \leq X \leq 9)$.

In exercises 4.23 and 4.24, the random variable X follows a Poisson process with the given value of λ and t .

- 4.23 Assuming $\lambda = 0.07$ and $t = 10$, compute:
- $P(4)$, b) $P(X < 4)$, c) $P(X \geq 4)$, d) $P(4 \leq X \leq 6)$.
- 4.24 Assuming $\lambda = 0.02$ and $t = 50$, compute:
- $P(2)$, b) $P(X < 2)$, c) $P(X \geq 2)$, d) $P(1 \leq X \leq 3)$.

- 4.25 The phone calls to a computer software help desk occur at the rate of 2.1 per minute between 3:00 pm and 4:00 pm, Compute the probability, and interpret the result, that number of phone calls between 3:10 pm and 3:15 pm is:
a) Exactly 8, b) Fewer than 8; c) At least 8
- 4.26 The holes on a major highway, in a large city, occur at the rate of 3.4 per mile. Compute the probability that the number of holes over 3 miles of randomly selected highway is:
a) Exactly 7, b) Fewer than 7, c) At least 7, d) Would it be unusual for a randomly selected 3-mile stretch of highway in this city to contain more than 15 potholes?
- 4.27 A builder ordered 200 8-foot-grade-A 2-by-4 pieces for a construction project. To qualify as a grade-A, each 2-by-4 piece will have no knots and will average no more than 0.05 imperfections per linear foot. The following table lists the number of imperfections per 2-by-4 in the 200 ordered.

Number of imperfections X	0	1	2	3
Frequency	124	51	20	5

- a) Construct a probability distribution for the random variable X, the number of imperfections per 8 feet of board, assuming it follows a Poisson process with $\lambda = 0.05$ and $t = 8$.
- b) Compute the expected number of 2-by-4 that will have 0 imperfections, 1 imperfection, and so on.
- c) Compare these results to with the number of actual imperfections. Does it appear that the 2-by-4 are of grade A quality? Why or why not?
- 4.28 Let X and Y are two discrete random variables, with the ordered values and joint probabilities as given below:

(x, y)	(1, 0)	(1, 1)	(1, 2)	(2, 0)	(2, 1)	(2, 2)	(3, 0)	(3, 1)	(3, 2)
$P(x, y)$	0.25	0.06	0.06	0.06	0.09	0.18	0.03	0.05	0.22

- a) Display the pmf for X and for Y,
b) Calculate Cov (X, Y).
- 4.29 Let X and Y are two discrete random variables, with the ordered values and joint probabilities as given below:

(x, y)	(0, 0)	(0, 2)	(0, 4)	(1, 0)	(1, 2)	(1, 4)	(2, 0)	(2, 2)	(2, 4)
$P(x, y)$	0.20	0.05	0.11	0.11	0.10	0.12	0.03	0.05	0.23

- a) Are X and Y independent?

- b) Calculate $\text{Var}(X)$ and $\text{Var}(Y)$.
- 4.30 Let the joint pmf of X and Y , as two discrete random variables, be defined by: $P(x, y) = (x + y)/32$, $x = 1, 2$, $y = 1, 2, 3$; and 0 otherwise.
- a) Find the marginal pmf's of X and Y .
b) Calculate: i) $P(X > Y)$, ii) $P(Y = 2X)$, iii) $P(X + Y = 3)$, iv) $P(X \leq 3 - y)$.
c) Are X and Y independent?
- 4.31 Let the joint pmf of X and Y , as two discrete random variables, be defined by: $P(x, y) = xy^2/30$, $x = 1, 2, 3$; $y = 1, 2$; and 0 otherwise.
- a) Are X and Y independent?
b) Calculate: i) $P(X \geq Y)$, ii) $P(Y = 2X)$, iii) $P(X - Y = 0)$, iv) $P(X + Y \leq 3)$.
- 4.32 Let X and Y have the joint pmf: $P(x, y) = (x + 2y)/18$, for $(x, y) = (1, 1), (1, 2), (2, 1), (2, 2)$. Calculate $\rho(X, Y)$.

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements



Understanding the Concepts Exercises CHAPTER 4

1. What is a random variable?
2. What are the two requirements for a discrete probability distribution?
3. What is meant by a binomial experiment? Give an example.
4. What are the two requirements for a binomial probability distribution?
5. Explain what success means in a binomial experiment.
6. What is the difference between the pmf and cdf of a discrete random variable?
7. Two characteristics describe the graph of the probability density function, can you name them?
8. The Poisson Probability Distribution has a unique characteristic, can you name it.
9. Do you need to calculate the mean and the variance for a normal distribution?
10. Name some discrete random variables other than the Binomial and the Poisson.

TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP Finding the Mean and Standard Deviation of a Discrete Random Variable

TI-83/84 Plus

1. Enter the values of the random variable in L1 and their corresponding probabilities in L2.
2. Press **STAT**, highlight **CALC**, and select **1: 1-Var Stats**.
3. With 1-VarStats on the HOME screen, type L1 followed by a comma, followed by L2 as follows: 1-Var Stats L1, L2. Hit **ENTER**.

TECHNOLOGY STEP-BY-STEP Computing Binomial Probabilities via Technology

TI-83/84 Plus

Computing $P(x)$

1. Press **2nd VARS** to access the probability distribution menu.
2. Highlight **0: Binompdf** (and hit **ENTER**).

3. With **Binompdf** (on the HOME screen, type the number of trials n , the probability of success p , and the number of successes, x , for example, with $n = 15$, $p = 0.3$, and $x = 8$, type **Binompdf** (15, 0.3, 8) Then hit **ENTER**.

Computing $P(X \leq x)$

1. Press **2nd VARS** to access the probability distribution menu.
2. Highlight A: **Binomcdf** (and hit **ENTER**).
3. With **Binomcdf** (on the HOME screen, type the number of trials n , the probability of success p , and the number of successes, x , for example, with $n = 15$, $p = 0.3$, and $x \leq 8$, type **Binomcdf** (15, 0.3, 8) Then hit **ENTER**.

Excel

Computing $P(x)$

1. Click on the **fx** icon. Highlight **Statistical** in the Function category window. Highlight **BINOMDIST** in the Function name window
2. Fill in the window with the appropriate values. For example, if $x = 5$, $n = 10$, and $p = 0.2$, fill in the window. Click **OK**

Computing $P(X \leq x)$

Follow the same steps as those presented for computing $P(x)$. In the **BINOMDIST** window, type **TRUE** in the cumulative cell.

TECHNOLOGY STEP-BY-STEP Computing Poisson Probabilities via Technology

TI-83/84 Plus

Computing $P(x)$

1. Press **2nd VARS** to access the probability distribution menu.

2. Highlight: `Poissonpdf` (and hit ENTER).
3. With `Poissonpdf` (on the HOME screen, type the value of Lambda, λ (the mean of the distribution), followed by the number of successes x , for example, type
`Poissonpdf (10, 8)`
Then hit ENTER.

Computing $P(X \leq x)$

4. Press 2nd VARS to access the probability distribution menu.
5. Highlight: `Poissoncdf` (and hit ENTER).
6. With `Poissoncdf` (on the HOME screen, type the value of Lambda, λ (the mean of the distribution), followed by the number of successes x , for example, type
`Poissoncdf (10, 8)` Then hit ENTER.



www.job.oticon.dk

oticon
PEOPLE FIRST



Excel

Computing $P(x)$

- a) Enter the desired values of the random variable X in column A.
- b) With cursor in cell B1, select the Formulas tab, Select more Formulas. Highlight Statistical, and then highlight POISSON in the function name menu.
- c) In the cell labeled X , enter A1 In the cell labeled mean, enter the mean (λ , Lambda). In the cell labeled cumulative, type FALSE. Click OK.

Computing $P(X \leq x)$

Follow the same steps as those presented for computing $P(x)$. In the **POISSON** window, type TRUE in the cumulative cell.

5 CONTINUOUS PROBABILITY DISTRIBUTIONS

5.1 INTRODUCTION

As it was the case in **Chapter 4**, let us begin with some definitions and terminology this time concerning continuous random variables.

The Random Experiment and its Sample Space: A random experiment is any process of measurements or observations, in which the outcome cannot be completely determined in advance. The **Sample Space**, S , of any random experiment, is the total collection of all possible outcomes of the random experiment. Each outcome has a certain probability (or chance) to occur. Thus, observing the value of a certain stock in the market is a random experiment. The sample space is the set of all possible values that the stock might take. Analyzing a certain chemical, to determine the iron content, is also a random experiment.



Schlumberger

WHY WAIT FOR PROGRESS?

DARE TO DISCOVER

Discovery means many different things at Schlumberger. But it's the spirit that unites every single one of us. It doesn't matter whether they join our business, engineering or technology teams, our trainees push boundaries, break new ground and deliver the exceptional. If that excites you, then we want to hear from you.

careers.slb.com/recentgraduates

The sample space for such an experiment will be presented as $S = \{X \mid x \geq 0\}$. Moreover, measuring the height, or weight, of an object is again a random experiment.

Random Variables: A random variable is any real-valued quantity, or numerical measure, whose value depends on the outcomes of a random experiment. A **random variable** is a function from the sample space of the experiment to the set of all real numbers. Given a random experiment with a sample space S , a function that assigns one and only one real number, $X(s) = x$ to each element s in S , is called a random variable. **Random Variables** are typically denoted, or identified, by capital **Roman** letters such as X , Y , and Z . The values that a random variable will take will be denoted by lower case letters x , y , and z , respectively. Thus, in observing a certain stock in the market, the value of the stock, X , is a random variable which may take any value between \$1 and \$10, for example, i.e. $1 \leq x \leq 10$. In measuring the weight of a person, the weight X is a random variable that may take any value between 105.45 lbs. and 300.15 lbs., say, i.e., $105.45 \leq x \leq 300.15$. Based on what had been explained, we see that the case is about a measure, not count as was the case in **Chapter 4**. So, in what is coming next, we like to introduce the continuous random variables.

Definition 5.1 A Continuous Random Variable is that variable which has an infinitely many values. The values of a continuous random variable can be plotted on a line in an uninterrupted fashion.

A **Continuous Random Variable** X takes all the values in an interval on the real line. For example, the distance you drive to work, or to school, every day; the distance between where you are, in the classroom, and the blackboard. Any random variable will take a certain value with a certain probability, as was seen in **Chapter 4** for **Discrete Random Variables**. The function that controls these probabilities is called the **Probability Mass Function**, or **PMF**, in the discrete case. It is the **Probability Density Function**, or **PDF**, in the continuous case of a random Variable. The value of the **PDF**, $f(x)$, at a certain value of the random variable will give a point on the graph of the **PDF** for that variable, and it is not a probability. For finding the probability in the continuous case, we find the area under the graph of $f(x)$ and above the values of x that make up that interval.

Definition 5.2: For the function $f(x)$ to be a **probability density function, PDF**, for the continuous random variable X , the following conditions should hold true for all values of x :

1. $f(x) \geq 0$, for all values of the continuous random variable X ,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$, and

$$3. P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f(x) dx.$$

In other words, this means that including a point, two points, excluding a point, or two points at the end of the interval does not change the probability. This is based on the **Fundamental Theorem of Calculus**:

$$\int_a^a f(x) dx = 0, \quad \text{i.e., there is no area above a point.}$$

In addition to the **PDF**, there is another important concept, or function, called the **Cumulative Distribution Function**, or **CDF**.

Definition 5.3: The **CDF** is defined all over the real line regardless whether the random variable is discrete or continuous, and it is given by in the continuous case:

for any real number t , we have the following formula for the **CDF**

$$F(t) = P(X \leq t) = \int_{-\infty}^t f(x) dx.$$

Aside from the above two definitions for the pdf, and cdf, there is another concept that needs to be given here. That concept is for the **Probability Distribution**. For a continuous random variable, the **Probability Model, or probability Distribution**, is a formula that gives the probability of that random variable over a defined set of values on the real line; i.e. over an interval of the real line, using the above definitions.

5.2 CONTINUOUS PROBABILITY DISTRIBUTIONS

We have seen, in the above sections that a discrete random variable can take on a countable number of values. Those values can be plotted on the real line with gaps between them. In addition to that, there are no decimal places when representing the values of a discrete random variable. In this section, we will deal with the other type of the random variables, namely the continuous type. A continuous random variable will take an uncountable, or an infinite, number of values. In essence a continuous random variable will take any value in an interval on the real line. Keeping this notion in mind, we see that the range of a continuous random variable will be a part, or all, of the real line. In the subsections of Section 5.4, we will address some of those continuous random variables that are defined on an interval of the real line, or all over the real line. Some probability distributions will be addressed in more details than others, based on their use and applications in real life.

5.2.1 UNIFORM DISTRIBUTION

It can be clearly understood from the name of this random variable being **Uniform**, on a part of the real line, that the **pdf** of this random variable, $f(x)$, will be a constant over a part of the real line. Thus, we find ourselves in a position to start with an experiment in order to set the grounds for the definition of a **Uniform probability distribution**. Given the interval $[a, b]$ on the real line, clearly with the normal setting, of an interval, as given by $a < b$. Let X be the random variable that will denote the outcome of an experiment of choosing at random from the interval $[a, b]$. If this process is repeated fairly, we find ourselves to assume that all the outcomes are equally likely to occur, and thus the probability of a point selected from the interval $[a, x]$, $x < b$, will be given by $(x - a) / (b - a)$. In other words, it looks clear that the probability is proportional to the length of the interval from which the point has been selected. Hence, we find that the cumulative distribution function for the random variable X is given by

$$F(x) = P(X \leq x) = \begin{cases} F(x) = 0, & x < a \\ F(x) = \frac{x-a}{b-a}, & a \leq x < b \\ F(x) = 1, & b \leq x \end{cases}$$

Since X is a continuous random variable, $F'(x) = f(x)$, the pdf of x whenever $F'(x)$ exists. Thus, for the pdf of X we have

$$f(x) = 1/(b - a), \quad a \leq x \leq b; \text{ and } 0, \text{ otherwise.}$$

In this text, when referring to the random variable X , that is uniformly distributed over the interval $[a, b]$, the following notation will be used: $X \sim U(a, b)$. This random variable is referred to as the **Rectangular distribution**. See **Figure 1** for the graph of $f(x)$ below.

THEOREM: 5.1 a) The mean and the variance of a Uniform random variable, $X \sim U(a, b)$, $-\infty < a < b < \infty$, are given by:

$$\mu = \frac{a+b}{2}, \text{ and } \sigma^2 = \frac{(b-a)^2}{12}.$$

a) The mgf of X is $M(t) = [e^{tb} - e^{ta}] / \{t(b - a)\}$, $t \neq 0$; and $M(0) = 1$.

(The proof of the above theorem is left as an exercise to the interested reader.)

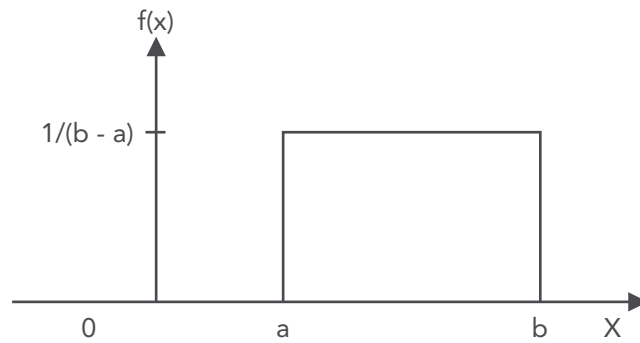


Figure 1 Uniform pdf

5.2.2 THE NORMAL DISTRIBUTION

The **Normal distribution** was first discovered in the eighteenth century. Astronomers and other scientists observed that repeated measurements of the same quantity (like the distance or the mass of an object) tended to vary, and when large of these measurements are taken and collected into a frequency distribution, one shape, similar to the normal curve kept repeating. **The Normal distribution** is often referred to as the **Gaussian distribution**, in

PREPARE FOR A LEADING ROLE.

English-taught MSc programmes in engineering: Aeronautical, Biomedical, Electronics, Mechanical, Communication systems and Transport systems. No tuition fees.

→ liu.se/master

li.u LINKÖPING UNIVERSITY



honor of **Karl Friedrich Gauss** (1777–1855), who also derived its equation from a study of errors in repeated measurements of the same quantity. Check **Figure 2**.

The **Normal distribution**, as a probability distribution for a continuous type of a random variable, is without any doubt, the most important distribution in statistics, and the most widely used continuous probability distribution. Recalling the discrete case of random variables, the **Binomial** is the most used, and referred to, in many applications. There are 4 basic reasons why the **Normal distribution** occupies a prominent place in Statistics. Those reasons are:

1. The **Normal distribution** comes close to fitting the actual observed frequency distributions of many phenomena:
 - a) Human characteristics such as weights, heights, and IQ's.
 - b) Outputs from physical processes; dimensions, and yield.
 - c) Repeated measurements of the same quantity, as described above, and errors made in measuring physical and economical phenomena, all follow a normal distribution.
2. The **Normal distribution** provides an accurate approximation to a large number of probability laws, e.g. the **Binomial distribution**.
3. The **Normal distribution** plays an important role in the theory of inferential statistics. This property is clear in the field because the sampling distributions of the mean and sample proportion, and many other statistics of large samples, tend to be normally distributed.
4. If the data do not follow a normal distribution, a certain transformation could be used in many cases to change it to a normal data.

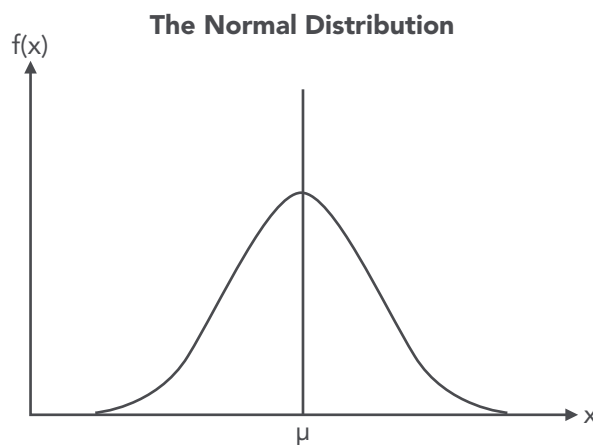


Figure 2 the Bell-Shaped Curve (Source: Interne)

The Probability Density Function (PDF) for a normally distributed random variable X , with mean μ and variance σ^2 , in short; $X \sim N(\mu, \sigma^2)$, is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x - \mu)^2 / (2\sigma^2)], \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \text{and } 0 < \sigma < \infty.$$

In the above notation, the **pdf** of the normal random variable X , μ is the location parameter, while σ is the shape, or scale parameter. The **Normal Distribution**, as given by its **pdf**, is the only continuous distribution that is characterized by its mean and variance. Thus, those parameters are not to be calculated. Due to the extensive use of the above **pdf**, for finding probabilities, and to the exhaustive, and wide range of values for the mean, μ , and the standard deviation, σ , of X , a unique table has been introduced based on transforming the above general random variable to the standard normal random variable Z , where $Z = (X - \mu) / \sigma$, and thus $Z \sim N(0, 1)$, or Z has what is called the **Standard Normal Distribution**. Check **Figure 3**.

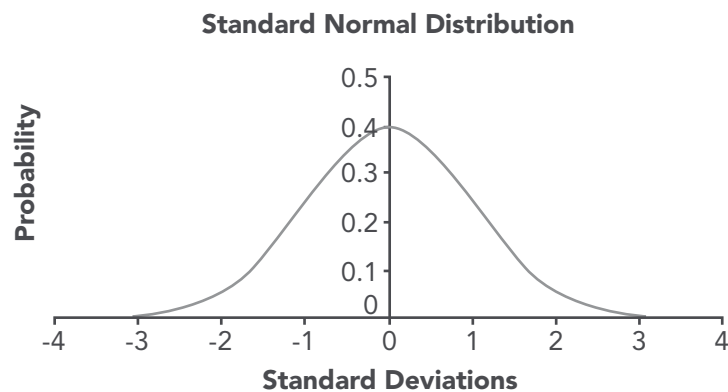


Figure 3 The Standard Normal Curve (Source: Internet)

The above cited transformation, namely $Z = (X - \mu) / \sigma$, has tremendously reduced the volumes of the tables that will correspond to the different values of μ & σ , into one single table, known as the **Standard Normal Table**.

Some basic properties of the probability density function for the normal random variable X are:

1. $f(x)$ is nonnegative all over the real line, i.e., $f(x) \geq 0$, for $x \in (-\infty, \infty)$.
2. The graph of $f(x)$ is symmetric around the value μ , and it is bell-shaped. Check **Figure 4**.

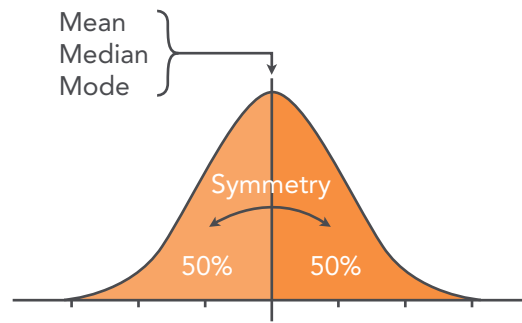


Figure 4 Bell-Shaped Curve (Source: Internet)

3. The integral of $f(x)$ over the entire real line is 1, i.e. $\int_{-\infty}^{\infty} f(x)dx=1$. In other words, the total area under the curve, and above the horizontal axis, is 1.
4. The horizontal axis acts as a horizontal asymptote to the curve of the normal pdf.
5. The areas under the graph of the normal density function represent probabilities. The value of the integral of $f(x)$ over the interval (a, b) represents the probability that $a \leq x \leq b$, in other words,

$$P(a \leq x \leq b) = \int_a^b f(x)dx ,$$

as shown, by the shaded area, in **Figure 5**.

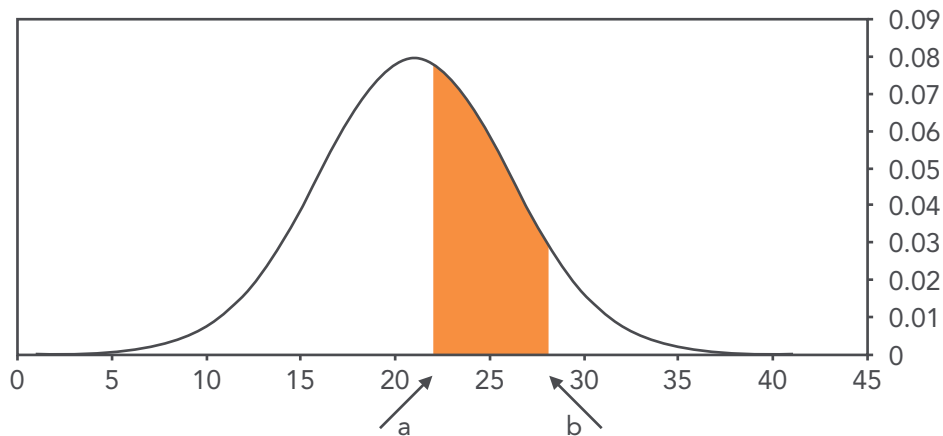


Figure 5 The Shaded area is $P(a < x < b)$ (Source: Internet)

6. It is to be noted that

$$P(a < x < b) = P(a < x \leq b) = P(a \leq x < b) = P(a \leq x \leq b).$$

This is due to the fact that

$$\int_a^a f(x) dx = 0,$$

based on the **Fundamental Theorem of Calculus**, as it was pointed out earlier. In other words, there is no area above a point.

7. **The Empirical Rule:** or the **68%–95%–99.7% Rule**, is the statistical rule for a normal distribution determined by the mean and the standard deviation, of that distribution. Approximately 68% of the area, under the normal curve, is between $X = \mu - \sigma$ and $X = \mu + \sigma$, and 95% of the area is between $X = \mu - 2\sigma$ and $X = \mu + 2\sigma$, while 99.7% of the area is between $X = \mu - 3\sigma$ and $X = \mu + 3\sigma$. Check **Figure 6A**.

In terms of probability, and using the mathematical notation, the above facts can be expressed as follows:

$$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6827,$$

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545, \text{ and}$$

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973.$$

EXAMPLE 5.1 The scores for all high school seniors taking the verbal section of the Scholastic Aptitude Test (SAT) in a particular year had a mean of 490 and a standard deviation of 100. The distribution of SAT scores is bell-shaped.

- What percentage of seniors scored between 390 and 590 on this SAT test?
- One student scored 795 on this test. How did this student do compare to the rest of the scores?
- A rather exclusive university only admits students who were among the highest 16% of the scores on this test. What score would a student need on this test to be qualified for admittance to this university?

For the example above we have $X \sim N(490, 100^2)$. **Figure 6A** displays the areas noted above.

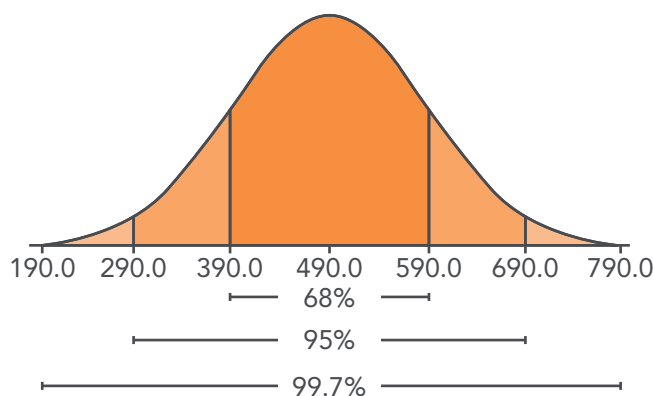


Figure 6A. The Empirical Rule Displayed on the data, Example 5.1(Source: Internet)

Solution: The data being described are the verbal SAT scores for all seniors taking the test in one year. Since this is describing a population, we will denote the mean and standard deviation as $\mu = 490$ and $\sigma = 100$, respectively. A bell-shaped curve summarizing the percentages given by the empirical rule is shown in **Figure 6A**.

- From **Figure 6B** about 68% of seniors scored between 390 and 590 on this SAT test.
- Since about 99.7% of the scores are between 190 and 790, a score of 795 is excellent. This is one of the highest scores on this test.
- Since about 68% of the scores are between 390 and 590, this leaves 32% of the scores outside this interval. Since a bell-shaped curve is symmetric, one-half of the scores, or 16%, are on each end of the distribution. **Figure 6B**, below, shows these percentages.

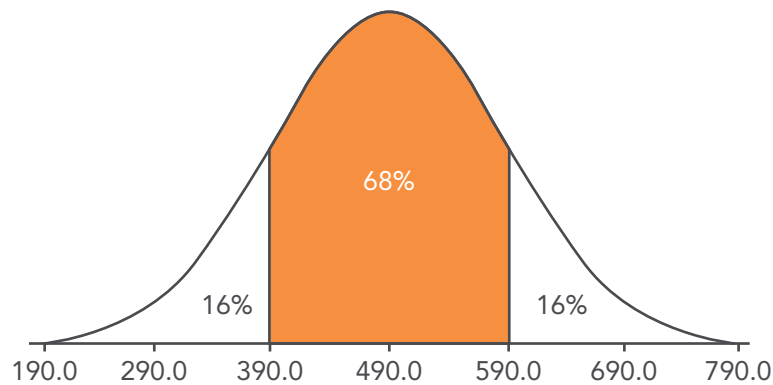


Figure 6 B. (Internet) the Empirical Rule Displayed on the Data

Since about 16% of the students scored above 590 on this SAT test, to be qualified for admittance to this university, a student would need to score 590 or above on this test.



EXAMPLE 5.2: The weight of a certain type of chicken, at a certain age, follows a normal distribution with mean 1.0 kg and a standard deviation of 0.20 kg. Find

- a) The probability that a chicken weighs less than 1.5 kg.



How will people travel in the future, and how will goods be transported? What resources will we use, and how many will we need? The passenger and freight traffic sector is developing rapidly, and we provide the impetus for innovation and movement. We develop components and systems for internal combustion engines that operate more cleanly and more efficiently than ever before. We are also pushing forward technologies that are bringing hybrid vehicles and alternative drives into a new dimension – for private, corporate, and public use. The challenges are great. We deliver the solutions and offer challenging jobs.

www.schaeffler.com/careers

SCHAEFFLER



- b) The probability that a chicken weighs between 0.9 kg and 1.2 kg.
- c) The probability that a chicken weighs more than 1.6 kg.
- d) The percentage of the chickens that weigh between 0.89 kg and 1.5 kg.
- e) Among a group of 300 chickens, how many will weigh between 0.8 and 1.5 kg?

Solution: Let X be the weight of a chicken, then X has a normal distribution with $\mu = 1.0$, and $\sigma = 0.2$, i.e. $X \sim N(1.0, 0.04)$. By using $Z = (X - \mu) / \sigma$, we have

- a) $P(X < 1.5) = P[(X - \mu) / \sigma < (1.5 - \mu) / \sigma] = P(Z < 2.5) = 0.9938$
- b) $P(0.9 < X < 1.2) = P(X < 1.2) - P(X < 0.9) = P(Z < 1) - P(Z < -0.5) = 0.8413 - 0.3085 = 0.5328$
- c) $P(X > 1.6) = 1 - P(X \leq 1.6) = 1 - P(Z \leq 3.0) = 1 - 0.9987 = 0.0013$
- d) $P(0.8 < X < 1.5) = P(Z < 2.5) - P(Z < -1.0) = 0.9938 - 0.1587 = 0.8351 = 83.51\%$.
- e) $0.8351 \cdot 300 = 250.53 \approx 251$.



Having done what we did so far for the normal distribution, let us discuss the procedure for finding the area under the normal curve. For the general normal random variable, we have: $X \sim N(\mu, \sigma^2)$ There are three cases that arise, and these are:

1. Finding the area under the normal curve, above the x-axis and between two values for the random variable. In other words, find the following probability

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b).$$

$$P(a \leq x \leq b) = \int_a^b f(x) dx, \text{ as shown in Figure 7.}$$

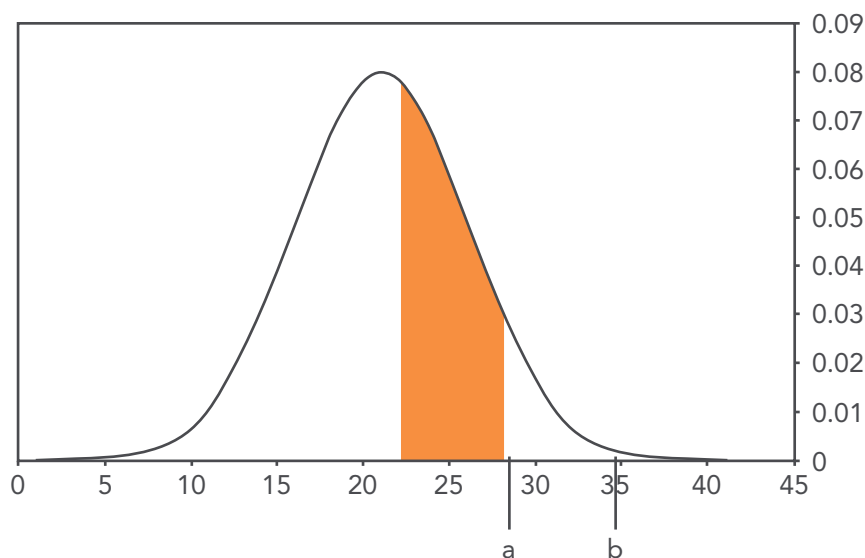


Figure 7 The Shaded area is $P(a < x < b)$ (Source: Internet)

Is it one probability or four different ones? All of them are equal, whether we include the end points, or exclude them, or include one and exclude the other.

2. Finding the area to the left of a value for the random variable: $P(x < c)$, **check Figure 8.**

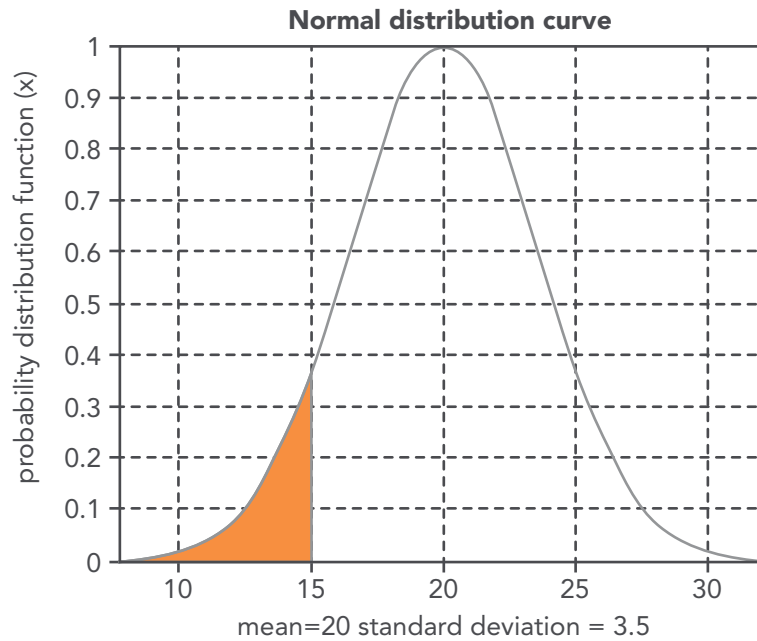


Figure 8 Shaded Area = $P(X < C = 15)$ (Source: Internet)

3. Finding the area to the right of a value for the random variable: $P(x > d)$, this is the un-shaded area in **Figure 9.**

With no doubt, we can find the required probabilities for any value of the variable X , any value for the mean, and any value for the standard deviation. Calculus techniques had been used just to do that. This saves a lot of time and resources, and reduced the tremendous number of tables into just ONE, the standard normal Table.

Therefore, if we use the transformation

$$Z = (X - \mu) / \sigma ,$$

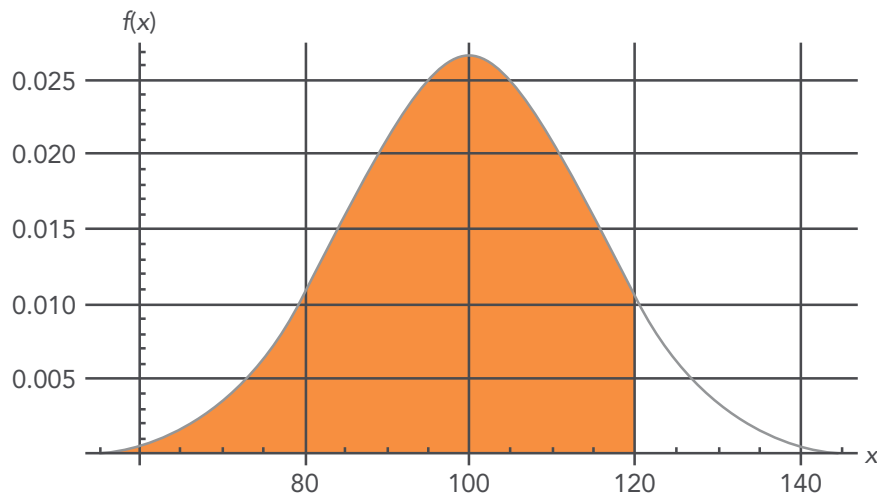


Figure 9 The Unshaded Area = $P(X > d = 120)$ (Source: Internet)

for the above three cases to find the probabilities can be calculated by using the **Standard Normal Table**. The equivalent case, in terms of Z will look like the following

1. $P(a \leq x \leq b) = P(z_1 \leq Z \leq z_2)$, with $z_1 = \frac{a - \mu}{\sigma}$, $z_2 = \frac{b - \mu}{\sigma}$,
2. $P(x < c) = P(Z \leq z_3)$, with $z_3 = \frac{c - \mu}{\sigma}$,

STUDY FOR YOUR MASTER'S DEGREE IN THE CRADLE OF SWEDISH ENGINEERING

Chalmers University of Technology conducts research and education in engineering and natural sciences, architecture, technology-related mathematical sciences and nautical sciences. Behind all that Chalmers accomplishes, the aim persists for contributing to a sustainable future – both nationally and globally.

Visit us on Chalmers.se or **Next Stop Chalmers** on facebook.



$$3. P(x > d) = P(Z \geq z_4), \text{ and } z_4 = \frac{d - \mu}{\sigma}.$$

The **Standard Normal Table** gives the area, under the normal curve, and to the left of any point, a , as long as it is given in terms of $Z = \frac{a - \mu}{\sigma}$, where $Z \sim N(0, 1)$. To find the probability for Part 1, we have

$$4. P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1), \text{ see Figure 7.}$$

To find the probability for Part 2, it is straight forward from the Z-Table,

$$5. P(Z \leq z_3), \text{ see Figure 8.}$$

For Part 3, since the total area under the normal curve is 1, and the **Standard Normal Table** lists the area to the left, we find ourselves doing the following for Part 3.

$$6. P(Z \geq z_4) = 1 - P(Z \leq z_4), \text{ see Figure 9.}$$

CAUTION and WARNING:

It looks as if it has been conventionally, and universally, agreed upon to have:

1. The value of $Z = (X - \mu) / \sigma$, to be calculated to two decimal places, starting at -3.49 to 3.49 by increments of 0.01 ,
2. In most books, two sections for the standard normal table,
3. Probabilities reported to 4 decimal places, and they represent the area on the left side of the cutting point,
4. The first column of the Standard Normal Table to display the Z-value with one decimal place, while the second decimal place is displayed as the top-heading of the Table.

We will display, in **Example 5.3**, how to read the **Standard Normal Table**, based on different values.

EXAMPLE 5.3: Let $X \sim N(70, 100)$, i.e. X is normally distributed with $\mu = 70$, and $\sigma = 10$. Find:

- a) $(56.5 < X < 90.1)$,
- b) $P(X < 73.2)$, and
- c) $P(X > 86.8)$

Solution: Transforming the values by standardizing we see that, the above probabilities can be found by using the standard normal table for the corresponding values of z as follows:

$$\begin{aligned} \text{a) } P(56.5 < X < 90.1) &= P[(56.5-70)/10 < Z < (90.1-70)/10] = P(-1.35 < Z < 2.01) \\ &= P(Z < 2.01) - P(Z < -1.35) = 0.9778 - 0.0885 = 0.8893. \end{aligned}$$

From the **Standard Normal Table**, we have:

Second decimal place for z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0		.9778								

We read .9778 along 2.0 under z and under .01, to get the probability of $z < 2.01$

Second decimal place for z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.3						.0885				

Hence, we read along -1.3 under z and under .05, to get 0.0885. The difference is the answer, as it is seen above.

$$\text{d. } P(X < 73.2) = P[Z < (73.2-70)/10] = P(Z < 0.32) = 0.6255.$$

Second decimal place for z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.3			.6255							

Thus, we read 0.6255 along 0.3 under z and under .02, to get 0.6255

$$e. P(X > 86.8) = 1 - P(X < 86.8) = 1 - P[(86.8 - 70)/10] = 1 - P(Z < 1.68) = 1 - 0.9535 = 0.0465.$$

Second decimal place for z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6									.9535	

Thus, we read 0.9535 along 1.6 under z and under .08, to get 0.9535.



For **Figure 10**, the way is going backwards. It is a two-way street. Now we are given the area that is standing for the probability of the random variable X being less than or equal to, or greater than, a certain cutting point, and we need to find the point, whether on the X-axis or the Z-axis. Finding the cutting point on one of the axes and using the transformation,

$$Z = (X - \mu) / \sigma,$$



Scholarships



Lnu.se

Open your mind to new opportunities

With 31,000 students, Linnaeus University is one of the larger universities in Sweden. We are a modern university, known for our strong international profile. Every year more than 1,600 international students from all over the world choose to enjoy the friendly atmosphere and active student life at Linnaeus University. Welcome to join us!

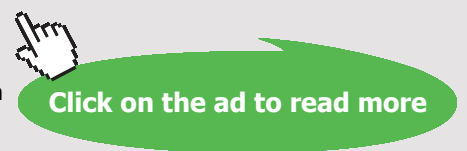
Linnaeus University

Sweden

Bachelor programmes in
Business & Economics | Computer Science/IT | Design | Mathematics

Master programmes in
Business & Economics | Behavioural Sciences | Computer Science/IT | Cultural Studies & Social Sciences | Design | Mathematics | Natural Sciences | Technology & Engineering

Summer Academy courses

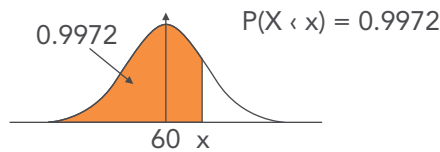


will get us the other cutting point on the other axis. Based on the limitations of the **Standard Normal Table** and the space provided, in case the given area is not in the Table, we will pick up the closer value to the given one and read that corresponding Z -value as our answer.

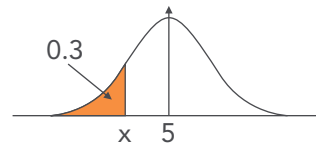
EXAMPLE 5.4 (Source: Internet)

Find the value of x in each of the following diagrams:

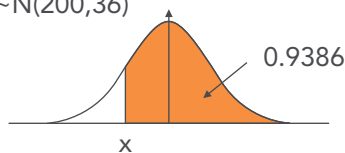
(a) $X \sim N(60,25)$



(b) $X \sim N(5,4/9)$



(a) $X \sim N(200,36)$



(b) $X \sim N(0,4)$

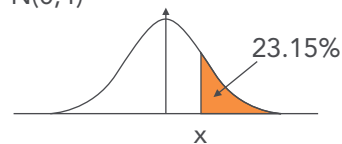


Figure 10 Finding Cutting point when the area, under normal curve, is given

Solution: I) by reading from the Standard Normal Table, after standardizing the variable, we have:

Second decimal place for z

a)	Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	2.7								.9972		

Since $X \sim N(60, 25)$, we have $\mu = 60$, and $\sigma = 5$. Reading the value in the **Standard Normal Table**, we found 0.9972 along 2.7 under Z and under 0.07, for the second place. Hence $P(Z < 2.77) = 0.9972$. From the above transformation, we see that $x = 5(2.77) + 60 = 73.85$.

Second decimal place for z

b)	Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	-0.5		.3015	0.3	.2981						

Since $X \sim N(5, 4/9)$, we have $\mu = 5$, and $\sigma = 2/3$. Since the area is $0.3 < 0.5$, the Z-value will be negative. From the **Standard Normal Table**, we found that $0.3015 < 0.3000 < 0.2981$, and it is between the two values cited along -0.5 under Z and under 0.02 and 0.03 for the second place. Since 0.3015 is closer to 0.3 than .2981, we can take z to be -0.52 . Using the transformation based on the distribution of X, we have $x = (2/3)(-0.52) + 5 = 4.65$.

Second decimal place for z

c)	Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	-1.5					.0618	.0614	.0606			

Based on $X \sim N(200, 36)$, we have $\mu = 200$, and $\sigma = 6$. Since the given area is to the right of the required value, for reading the value in the **Standard Normal Table**, we need to subtract this number from 1, i.e., $1 - 0.9386 = 0.0614 < 0.5$. Again, since the area is less than 0.5, the corresponding z-value will be negative. So, we now read $0.0606 < 0.0614 < 0.0618$, which is cited along -1.5 under Z and under 0.04, and under 0.05, for the second decimal places. We can take z to be -1.54 . Using the transformation based on the distribution of X, we have $x = 6(-1.54) + 200 = 190.76$.

Second decimal place for z

d)	Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	0.70				0.7673	0.7685	0.7704				



Now since $X \sim N(0, 4)$, we have $\mu = 0$, and $\sigma = 2$. As it was in part (c) above, and since the given area, as a percentage, is to the right of the required value, for reading the value from the **Standard Normal Table**, we need to subtract this number from 100, i.e., $100 - 23.15 = 76.85$. Based on that, we now read $0.7673 < 0.7685 < 0.7703$ which is cited along 0.70 under Z and under 0.03, and 0.04 for the second decimal places. We can take Z to be 0.73, the closer value. Using the transformation based on the distribution of X, we have $x = 2(0.73) + 0 = 1.46$.

The above Example could have been solved using **Technology-Step-by-Step** by applying the command: **INVNorm (Area to the Left, Mean, Standard Deviation)**, and press enter to get the value of x to any decimal places you like, and no rounding for the area in order to use the **Standard Normal Table**.

Solution II): Using technology, and in particular the command: **InvNorm (Area to the Left, Mean, Standard Deviation)**, for the four parts of the example we have:

- When $X \sim N(60, 25)$, using $\text{InvNorm}(0.9972, 60, 5)$ will give 73.8516, as the cutting point x , such that $P(X < x) = 0.9972$.
- When $X \sim N(5, 4/9)$, using $\text{InvNorm}(0.3000, 5, 2/3)$ will give 4.6504, as the cutting point x , such that $P(X < x) = 0.3000$.
- When $X \sim N(200, 36)$, using $\text{InvNorm}(0.0614, 200, 6)$ will give 190.7412, as the cutting point x , such that $P(X < x) = 0.0614$, or $P(X > 190.7412) = 0.9386$.
- When $X \sim N(0, 4)$, using $\text{InvNorm}(0.7685, 0, 2)$ will give 1.4678, as the cutting point x , such that $P(X < x) = 0.7685$, or $P(X > 1.4678) = 0.2315 = 23.15\%$.



REMARK: Approximating the Binomial Distribution probabilities using the normal is not needed any more at this time of technology. Since a lot of software and calculators are accessible to students with more accuracy and less time consuming. Based on this notion, we will not discuss this topic here anymore.

e-learning for kids

- The number 1 MOOC for Primary Education
- Free Digital Learning for Children 5-12
- 15 Million Children Reached

About e-Learning for Kids Established in 2004, e-Learning for Kids is a global nonprofit foundation dedicated to fun and free learning on the Internet for children ages 5 - 12 with courses in math, science, language arts, computers, health and environmental skills. Since 2005, more than 15 million children in over 190 countries have benefitted from eLessons provided by EFKI. An all-volunteer staff consists of education and e-learning experts and business professionals from around the world committed to making difference. eLearning for Kids is actively seeking funding, volunteers, sponsors and courseware developers; get involved! For more information, please visit www.e-learningforkids.org.



THEOREM 5.2: If $X \sim N(\mu, \sigma^2)$, then the **MGF**, of X , is given by $M(t) = \exp(\mu t + \sigma^2 t^2/2)$.

(The proof is left as an exercise.)

5.2.3 STUDENT'S T-DISTRIBUTION

Student's t -distribution (or simply the **t -distribution**) is a family of continuous probability distributions. It arises when estimating the sampling distribution of a sample mean, taken from a normally distributed population when: a) the sample size is small and the population standard deviation is unknown. **The t -distribution** is characterized by its degrees of freedom. There is a completely different distribution for each value of the degrees of freedom ν . The **t -distribution** is different for each sample size, n . Hence, the following question is raised: What is the relationship between the sample size and the degrees of freedom for the t -distribution? The relationship is given by $\nu = n - 1$, and the larger the sample, the more the distribution resembles a normal distribution.

The **t -distribution** plays a big role in a number of widely used statistical analyses. It is used for assessing a statistical hypothesis about one population mean, or the difference between two means when the populations' variances are unknown, based on small samples of size $n < 30$ each. It is used under the name of **Student's t -Test**, as it will be seen in **Chapter 7**. In **Chapter 6**, the **t -distribution** will be used for constructing confidence intervals about one mean or about the difference between two population means. In **Chapter 10**, the **t -distribution** will appear again in testing on the significance of linear regression. Moreover, the **Student's t -distribution** also arises in the **Bayesian Analysis** of data distracted from a normal family.

If we take a sample of $n = \nu + 1$ observations from a normal distribution, see **Figure 11** (the black curve is representing a very large ν), compute the sample mean and plot it, and repeat this process infinitely many times (for the same n), we get the probability density function for that n , as shown in the image on the right.

The **t -distribution** is symmetric and bell-shaped, like the normal distribution. It has thicker tails and lower peak than the normal distribution due to the fact that it has a variance > 1 , when it is defined, check below. As the number of degrees of freedom grows, the t -distribution approaches the normal distribution with mean 0 and variance 1.

If we take a sample of $n = \nu + 1$ observations from a normal distribution (the black curve on the figure on the right of this page, representing a very large ν), compute the sample

mean and plot it, and repeat this process infinitely many times (for the same n), we get the probability density function for that n , as shown in the image on the right.

For more on the history of the t -distribution, as it was first derived as a posterior distribution in 1876 by **Helmert** (1875, 1876a, and 1876b), we refer the reader to **Pfanzag, J.** and **Sheynin, O. (1996)**, **Sheynin, O. (1995)**, and **Lüroth, J. (1876)**.

In the English-language literature it takes its name from [William Sealy Gosset's](#) 1908 paper in [Biometrika](#) under the pseudonym “Student”, see Student 1908. **Gosset** worked at the Guinness Brewery in **Dublin, Ireland**, and was interested in the problems of small samples, for example the chemical properties of barley where sample sizes might be as low as 3. One version of the origin of the pseudonym is that **Gosset's** employer preferred staff to use pen names when publishing scientific papers instead of their real name; therefore, he used the name “Student” to hide his identity. Another version is that Guinness did not want their competitors to know that they were using the t -test to test the quality of raw material, **Mortimer, Robert G. (2005)**. **Gosset's** paper refers to the distribution as the “frequency distribution of standard deviations of samples drawn from a normal population”. It became well-known through the work of **Ronald A. Fisher**, who called the distribution “Student’s distribution” and referred to the value as t , see **R.A. Fisher 1925**, and **Walpole, R. and al. (2002)**.

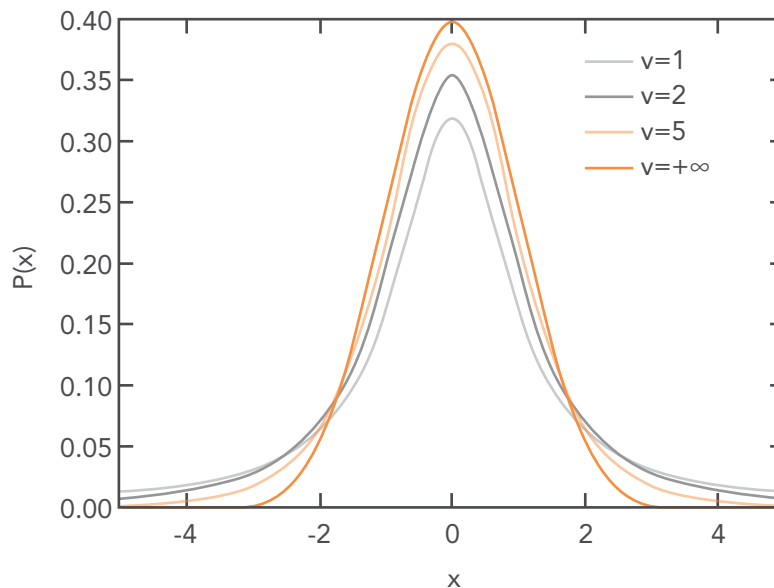


Figure 11 Student’s t Probability density function (Source: Internet)

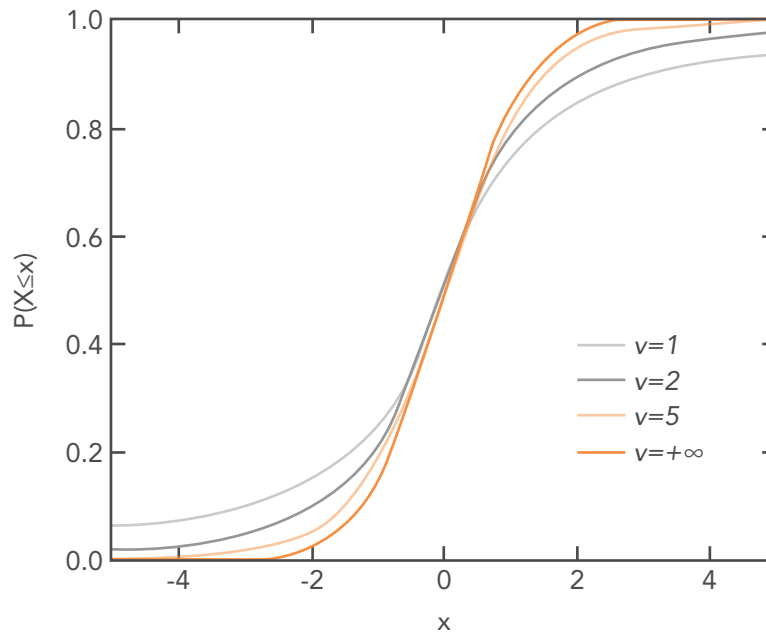


Figure 12 Student's t Cumulative distribution function (Source: Internet)

Student's **t-distribution** has the following pdf as given below:

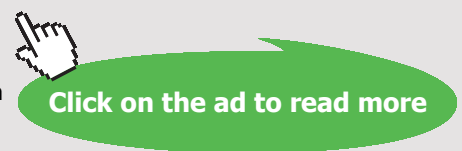
$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1+t^2/\nu\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty,$$

.....Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



where ν , a positive integer, is the number of degrees of freedom, and Γ is the gamma function. The above pdf can be derived when we let

$$T = \frac{Z}{\sqrt{U/\nu}},$$

with Z as the standard normal variable, that is, $Z \sim N(0, 1)$, and U the random variable that has a **Chi-square** distribution with ν degrees of freedom, when Z and U are independent. For a **t-distribution** with ν degrees of freedom, we see that the expected value, or the mean of the distribution, is 0 and the variance is given by $\nu/(\nu-2)$ if $\nu > 2$. The Coefficient of Skewness is 0, if $\nu > 3$, and the Coefficient of Kurtosis is $6/(\nu-4)$ if $\nu > 4$. (The interested reader may consult: Hogg and Tanis (2010), and Hogg and Craig (1978), for more.)

EXAMPLE 5.5: With $\nu = 4$, for 95% for *one-sided* (90% for *two-sided*), the value is “2.132”. Then the probability that T is less than 2.132 is 95% or $P(-\infty < T < 2.132) = 0.95$; this also means that $P(-2.132 < T < 2.132) = 0.9$.

Solution: This can be calculated by the symmetry of the distribution, as given below:

$$P(T < -2.132) = 1 - P(T > -2.132) = 1 - 0.95 = 0.05, \text{ or}$$

$$P(-2.132 < T < 2.132) = 1 - 2(0.05) = 0.9.$$



Most statistical textbooks list t distribution Tables. Nowadays, the better way to a full precise critical t -value or a cumulative probability, is the statistical function implemented in spreadsheets (Office Excel, TI Calc, Minitab, and R software programs), or an interactive calculating web page.

The following **Table** lists a few selected values for the t -distributions with ν degrees of freedom for a range of *one-sided* or *two-sided* critical regions.

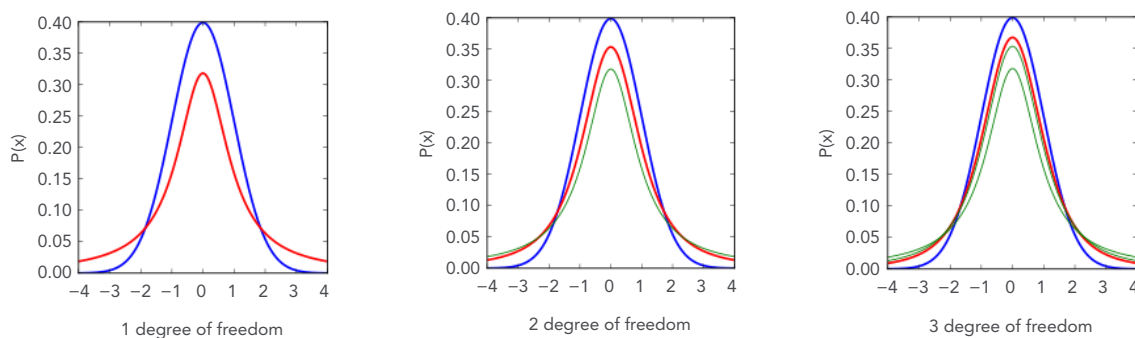
The first column is the number of degrees of freedom.

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587

11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

The following images show the density of the t -distribution for increasing values of ν . The normal distribution is shown as a blue line for comparison. Note that the t -distribution (red line) becomes closer to the normal distribution as ν increases.

Density of the t -distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue). Previous plots shown in green.



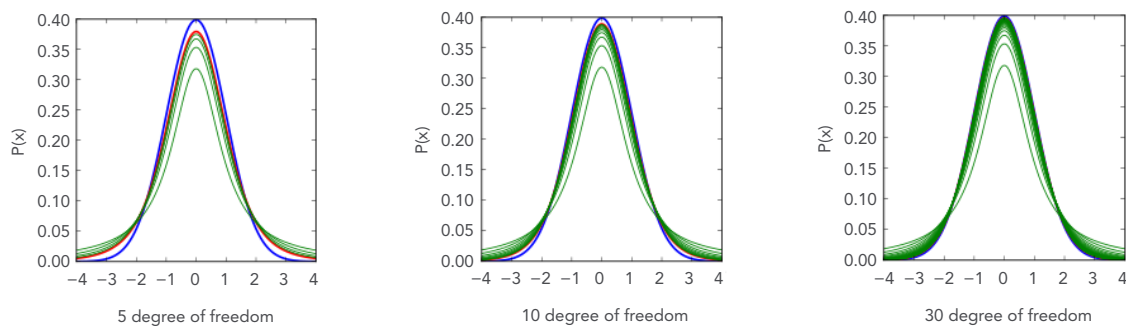


Figure 13 Student's t-Probability Density Function (Source: Internet)

5.3 NORMAL APPROXIMATION TO DISCRETE DISTRIBUTIONS

It has been shown, in **Chapter 4**, that a discrete random variable takes on countable values, finite or infinite as in the case of the **Binomial** and **Poisson** distributions respectively. In the continuous case for a random variable, that is defined over an interval on the real line, the probability to take a particular value, in an interval, is zero. This is based on the **Fundamental Theorem of Calculus**, see **Definition 5.2** for the conditions on $f(x)$ to be a **PDF** for a continuous random variable and how to calculate probabilities.



We have dealt with the Binomial Probability distribution in **Section 4.3.2**. In that discussion, we have found that formulas can be used to compute probabilities of events in a binomial experiment for a quite small range of values for the number of trials. As it has become very clear when we applied the formulas, the number of tables for the Binomial Probability Distribution is limited. Therefore, a large number of trials of a binomial experiment, makes this formula difficult to use. For example, given 300 trials of a binomial experiment, to compute the probability of 200 or more successes requires that we compute the following probabilities:

$$P(X \geq 200) = P(200) + P(201) + \dots + P(300).$$

This would be very tedious and time consuming to compute by hand. Fortunately, we have an alternative means for approximating binomial probabilities, provided that certain conditions are satisfied.

As it could be seen, in the **Figure 6 B of Chapter 4, below**; For a fixed p , as the number of trials n in a binomial experiment increases, the probability distribution of the random variable X becomes more nearly symmetric and bell shaped (**look back at Chapter 4, Figure 6**). As a rule of thumb, if $np(1 - p) \geq 10$, the probability distribution will be approximately symmetric and bell shaped, and thus, the normal approximation is in place, usually after standardization.

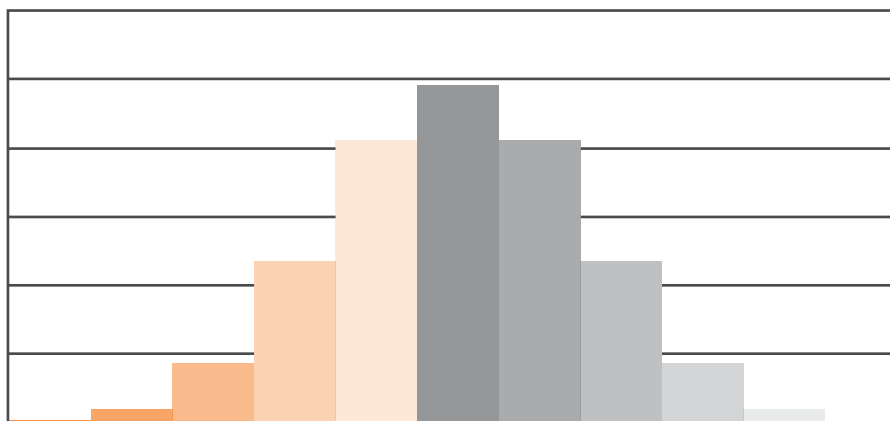


Figure 6 B – Chapter 4

The reader has to bear in mind that, in the continuous case, the probability to hit a point in the domain of the random variable is zero, and since this is true, we have to consider an interval around that value of the discrete random variable. Thus, If X has a binomial probability distribution, i.e., $X \sim \text{Bin}(n, p)$, Then to find $P(X = a)$, we need to calculate the continuous approximation given by: $P(a - 0.5 \leq X \leq b + 0)$. Therefore, we have the following Rule:

If $np(1 - p) \geq 10$, the binomial random variable X is approximately normally distributed, with mean $\mu_x = np$ and standard deviation $\sigma_x = \sqrt{np(1-p)}$.

In addition to the above approximation: $P(X = a) \approx P(a - 0.5 \leq X \leq a + 0.5)$, we have the following table for the case that might arise in practice.

Exact Probability Using Binomial	Approximate Probability Using Normal
$P(X = a)$	$P(a - 0.5 \leq X \leq a + 0.5),$
$P(X \leq a)$	$P(X \leq a + 0.5),$
$P(X \geq a)$	$P(X \geq a - 0.5),$ and
$P(a \leq X \leq b)$	$P(a - 0.5 \leq X \leq b + 0.5).$

Provided, in the above inequalities, that a and b are integers, in the domain of X .

To illustrate the above-mentioned cases, we look at the following examples.

EXAMPLE 5.6 According to the *American Red Cross*, 7% of people in the United States have blood type o-negative. What is the probability that, in a random sample of 500 people in the U.S., fewer than 30 have blood-type o-negative?

Solution: Clearly, each of the 500 independent trials has a probability of success equal to 0.07. This is a binomial distribution, with $n = 500$, and $p = 0.07$. To calculate $P(X < 30)$, it will take quite some time to do it by hand. Thus, let us check the condition for approximation. We see that $np(1 - p) = 500(0.07((1 - 0.07) = 32.55 > 10$. Hence, we can use the normal approximation. It is seen that $P(X < 30) = P(X \leq 29)$. This is approximately equal to the area under the normal curve to the left of $x = 29.5$, with $\mu_x = np = 500(0.07) = 35$, and $\sigma_x = \sqrt{np(1 - p)} = 5.71$. To complete the solution, convert $x = 29.5$ to a z-score to find that $z = -0.96$. By using the **Standard Normal Table, Table II, in Appendix A**, we see that $P(X < 30) = P(X \leq 29) = P(z \leq -0.96) = 0.1685$. Therefore the approximate probability that fewer than 30 people will have blood type o-negative is 16.85%

On the other side, since we are in the 21st Century, The Technology Century, using Technology Step-by-Step as was outlined in **Chapter 4** for the **Discreet Random Variables**, and with TI-84 Calculator, we have: $\text{Binomcdf}(500, 0.07, 29) = 0.1677676733$.

REMARK: When adding or subtracting 0.5 from the value for the random variable X, we are making correction for continuity.


It is needless to say, since we are in the Technology Century, and with the availability of calculators and software systems to solve any statistical problems, the normal approximation to the Binomial, or to any Discrete Random variable, is overlooked in academia nowadays.

NOTE: The aforementioned discussion concerning the Binomial Probability Distribution can be carried on any discrete Probability distribution once the mean and the variance of that distribution are known, then the standardization comes in.


EXAMPLE 5.7:

- a) Suppose we want to compute the probability that between 50 and 75 white blood cells will be neutrophils, where the probability that any one cell is neutrophil is 0.6. These limits are chosen as proposed limits to the range of neutrophils in normal people, and we like to predict what proportion of people will be in the normal range according to this definition.
- b) Suppose a neutrophil count is defined abnormally high if the number of neutrophils is ≥ 76 , and abnormally low if the number is ≤ 49 . Calculate the proportion of people whose neutrophil counts are abnormally high or low.

SIMPLY CLEVER



WE WILL TURN YOUR CV INTO AN OPPORTUNITY OF A LIFETIME



Do you like cars? Would you like to be a part of a successful brand? As a constructor at ŠKODA AUTO you will put great things in motion. Things that will ease everyday lives of people all around Send us your CV. We will give it an entirely new new dimension.

Send us your CV on www.employerforlife.com

Solution:

- a) The exact probability can be found by applying the binomial probability distribution formula. In other words, we have: $\text{Binomcdf}(100, 0.6, 75) - \text{Binomcdf}(100, 0.6, 50) = 0.9994384648 - 0.0270991975 = 0.9723392673 \approx 0.9723$. Since $np = 100 \cdot 0.6 = 60$, and $npq = 100 \cdot 0.6 \cdot 0.4 = 24 > 10$, thus, for the normal approximation we have, $P(50 \leq X \leq 75) = P(49.5 \leq X \leq 75.5) = \text{Normal cdf}(49.5, 75.5, 60, \sqrt{24}) = 0.9832$; hence 98.3% of the people will be normal.
- b) The probability of being abnormally high is given by $P(X \geq 76) \approx P(Y \geq 75.5)$, where $X \sim \text{Bin}(100, 0.6)$, and $Y \sim N(60, 24)$. Thus, we have $P(Y \geq 75.5) = \text{Normal cdf}(75.5, 1e99, 60, \sqrt{24}) = 0.000778$.

Similarly, the probability of being abnormally low is given by $P(X \leq 49) \approx P(Y \leq 49.5)$, where $X \sim \text{Bin}(100, 0.6)$, and $Y \sim N(60, 24)$. Thus, we have $P(Y \leq 49.5) = \text{Normal cdf}(-1E99, 49.5, 60, \sqrt{24}) = 0.01604$.

Therefore, 0.1% of people will have abnormally high neutrophil counts and 1.6% will have abnormally low neutrophil counts.

5.4 MORE CONTINUOUS DISTRIBUTIONS

In addition to the continuous random variables discussed so far in **Sections 5.2** there are more non-negative continuous random variables that can be used in real life applications. We will limit ourselves, in this text, to the distributions that have been presented and discussed so far. The interested reader, with a good background in statistics, can pursue those distributions and their applications by checking: **Shayib and Awad** (1990), **Shayib** (1991), **Shayib and Young** (1991), **Shayib and Aly** (1992), Shayib (2005), **Shayib and Haghghi** (2007), **Shayib and Haghghi** (2013), and the references cited therein also.

5.4.1 THE GAMMA DISTRIBUTION AND ITS TWINS

The **Gamma distribution** is a two-parameter, or one-parameter, family of non-negative continuous probability distributions. It took its name from the **Gamma Function** as defined below:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy, \quad \alpha > 0, \text{ and } y \geq 0.$$

The integrand, in the above definition of the gamma function, is positive and thus the integral itself is positive. For $\alpha > 1$, integration by parts gives

$$\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1).$$

The parameter α is called the shape parameter. Based on the above expression, we see that

$$\Gamma(6) = 5 \cdot \Gamma(5) = 5 \cdot 4 \cdot \Gamma(3).$$

Thus, for $\alpha = n$, a positive integer, we have, by repeating the formula $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$, that

$$\Gamma(n) = (n - 1) \cdot (n - 2) \cdot (n - 3) \cdots (2) \cdot (1) \cdot \Gamma(1).$$

Since $\Gamma(1) = 1$, (the interested reader with good calculus knowledge can show that), we can see that $\Gamma(n) = (n - 1)!$. Based on this, the **Gamma** function is called the **Generalized Factorial**. Moreover, since $\Gamma(1) = 1 = 0!$ This is consistent with the factorial convention. For more details on the **Gamma** function, and other special function, the interested reader is referred to **Whittaker and Watson (1965)**.

The above definition of the **Gamma** function, in terms of the one parameter α , will lead to another non-negative continuous random variable $Y \sim G(\alpha)$; or $Y \sim \text{Gamma}(\alpha)$, that will have the following pdf:

$$g(y) = \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}, \quad \alpha > 0, y \geq 0;$$

$$= 0; \text{ otherwise.}$$

In case we replace Y by X/β , where β will act as a scale parameter, we will get the two-parameter Gamma random variable X , i. e. $X \sim G(\alpha, \beta)$. Thus, X will have as its pdf the following function:

$$f(x) = \frac{x^{\alpha-1} \cdot e^{-x/\beta}}{\beta^\alpha \cdot \Gamma(\alpha)}, \quad \alpha, \beta > 0, \text{ and } x \geq 0;$$

$$= 0; \text{ otherwise.}$$

It was pointed out that the waiting time until the α^{th} change, in a Poisson process, has a gamma distribution with parameters α and $\beta = 1/\lambda$.

In addition to the above presentation for the gamma distribution, with α and β as cited above, there is another presentation with three parameters, adding a new location parameter μ , as given below:

The general formula for the Gamma pdf, with 3 parameters is

$$f(x) = \frac{(x-\mu)^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{x-\mu}{\beta}\right), \quad \alpha, \beta > 0, x \geq \mu.$$

Let us look at an Example.

EXAMPLE 5.8: Suppose that an average of 30 customers per hour arrive at a shop in accordance to a Poisson process. In other words, if a minute is our unit, then $\lambda = 1/2$. What is the probability that the shopkeeper will wait more than 5 minutes before both of the first two customers arrive?

Solution: If X denotes the waiting time, in minutes, until the second customer arrives, then X has a gamma distribution with $\alpha = 2$, and $\beta = 1/\lambda = 2$. Hence, we have

$$f(x) = 1/[\Gamma(2) \cdot 2^2] \cdot x^{2-1} \cdot e^{-x/2}, \quad 0 \leq x < \infty.$$

$$P(X > 5) = \int_5^\infty f(x) dx = 0.287.$$

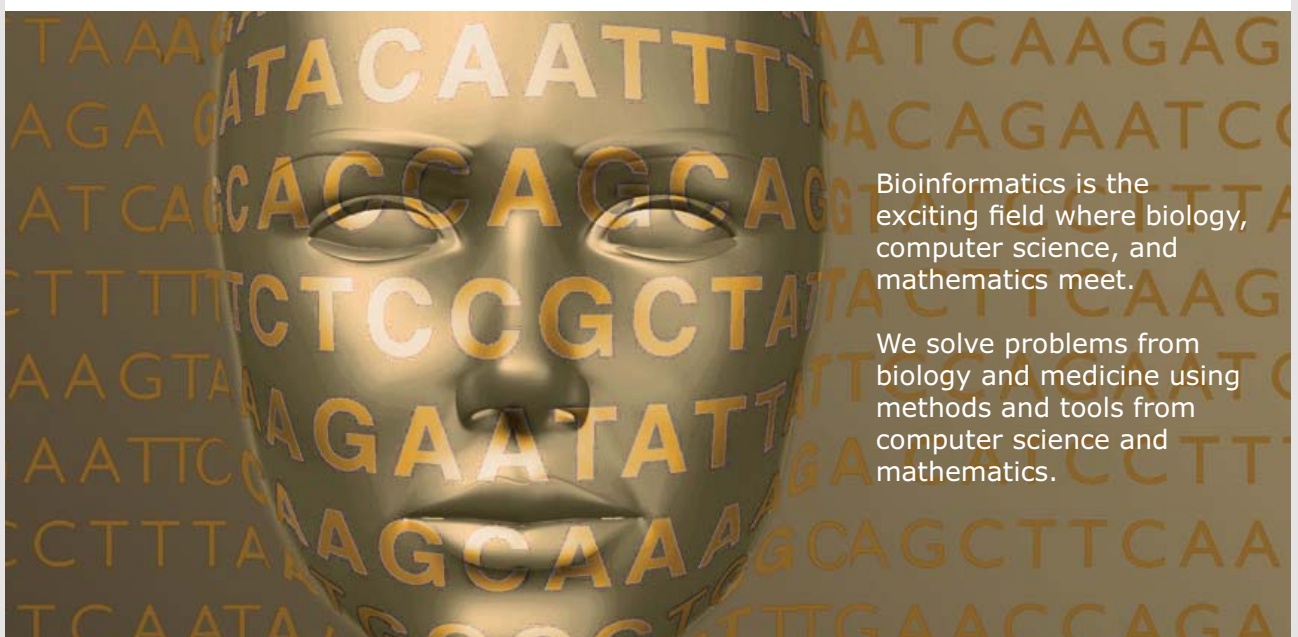


The gamma distribution is frequently used to model waiting times. The waiting time until death is a random variable that is frequently modeled with a gamma distribution, as in life



UPPSALA
UNIVERSITET

Develop the tools we need for Life Science Masters Degree in Bioinformatics



Bioinformatics is the exciting field where biology, computer science, and mathematics meet.

We solve problems from biology and medicine using methods and tools from computer science and mathematics.

Read more about this and our other international masters degree programmes at www.uu.se/master



testing, see Hogg and Craig (1978). In this sequel we will adopt the notation $X \sim G(\alpha, \beta)$, as the random variable X with a gamma pdf as defined in terms of α, β , and $f(x)$.

THEOREM 5.3: If $X \sim G(\alpha, \beta)$, and its pdf is

$$f(x) = \frac{x^{\alpha-1} \cdot e^{-x/\beta}}{\beta^\alpha \cdot \Gamma(\alpha)}, \quad \alpha, \beta > 0, \text{ and } x \geq 0;$$

$$= 0; \text{ otherwise.}$$

Then $E(X) = \mu_x = \alpha\beta$ and $\sigma_x^2 = \alpha\beta^2$. Also, the mfg of X is $M(t) = (1 - \beta t)^{-\alpha}$, $t < \beta^{-1}$.

(The proofs for the above expressions are left as an exercise to the interested reader.)

The above cited moments, with the mgf for the random variable X that is distributed as $G(\alpha, \beta)$, are given to be referred to later in the sequel. As we will see in the next two subsections, the gamma probability density function is a generalization of two more non-negative continuous random variables that frequently appear in real life applications, as special cases of the gamma function. Those two non-negative continuous random variables are A) the **Exponential** and B) the **Chi-square** random variables.

We will start with the **Exponential Distribution**.

5.4.2 EXPONENTIAL DISTRIBUTION

The Gamma distribution, as given above, is sometimes called the **Generalized Exponential distribution**. This is based on the special case, from the above pdf when $\alpha = 1$. In that case, the pdf will take the form

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad \beta > 0, \text{ and } x \geq 0;$$

$$= 0; \text{ otherwise.}$$

On the other hand, when we look at the **Poisson distribution** for the numbers of changes, or occurrences, of the discrete random variable, X ; the waiting time, was not discrete. The waiting times, between changes, are also random variables. Those waiting times are non-negative continuous-type random variables. Each can be any nonnegative real number. Another representation for the **Exponential** distribution is by letting $\lambda = 1/\beta$, see **Figure 19**. Based on this transformation, the pdf for the **Exponential** random variable, X , as shown in the figure below, will be presented as:

$$p(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, \text{ and } x \geq 0;$$

$$= 0; \text{ otherwise.}$$

Based on the presentation of the pdf, we see that the mean and variance of the **Exponential** random variable, X , are given by: $\mu_x = \beta$, and $\sigma_x^2 = \beta^2$. On the other hand, in terms of $\lambda = 1/\beta$ we have $\mu_x = 1/\lambda$, and $\sigma_x^2 = 1/\lambda^2$. (We leave it to the interested reader to compare these values with those of the gamma distribution, $\alpha = 1$.)

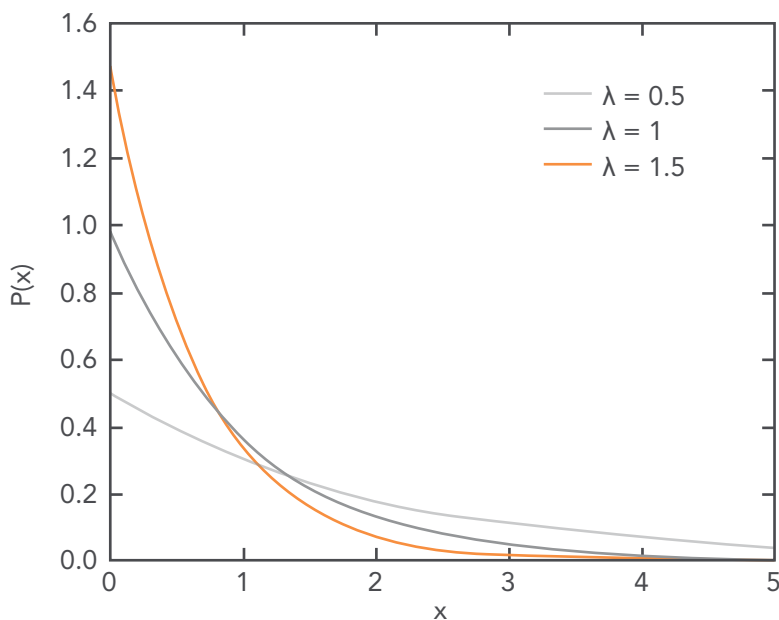


Figure 14 The Exponential Density Function (Source: Internet)

Here are two Examples on the Exponential Distribution.

EXAMPLE 5.9: Telephone calls enter a college switchboard according to a Poisson process on the average of 2 every 3 minutes. Let X denote the waiting time until the first call that arrives after 10 am.

- a) What is the pdf of X ? b) Find $P(X > 2)$.

Solution: a) X , has an Exponential distribution with parameter $\lambda = 2/3$ calls per minute. Thus, the pdf is given by

$$f(x) = (2/3) e^{-2x/3}, \quad 0 \leq x < \infty.$$

$$b) P(X > 2) = \int (2/3) e^{-2x/3} dx = e^{-4/3}.$$



EXAMPLE 5.10: Let X have an exponential distribution with a mean of $\beta = 20$. Then the pdf of X is

$$f(x) = \frac{1}{20} e^{-x/20}, \quad x \geq 0;$$

$$= 0; \text{ otherwise.}$$

Find $P(X < 18)$.

Solution: $P(X < 18) = \int_0^{18} f(x) dx = \int_0^{18} \frac{1}{20} e^{-x/20} dx = 1 - e^{-18/20} = 0.5934$.



5.4.3 CHI-SQUARE DISTRIBUTION

The **Chi-squared distribution** (also **Chi-square** or χ^2 -**distribution**) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. It can be derived as a special case of the Gamma distribution by letting $\alpha = k/2$, and $\beta = 2$, and thus the pdf will be given by

$$f(x) = \frac{x^{k/2-1} \cdot e^{-x/2}}{2^{k/2} \cdot \Gamma(k/2)}, \text{ k is a positive integer, and } x \geq 0; \text{ and } 0, \text{ otherwise.}$$

Figure 15-A and **Figure 15-B** below, show the graphs for the Chi-square pdf and cdf respectively.

UNIVERSITY OF COPENHAGEN



Copenhagen Master of Excellence

Copenhagen Master of Excellence are two-year master degrees taught in English at one of Europe's leading universities

Come to Copenhagen - *and aspire!*

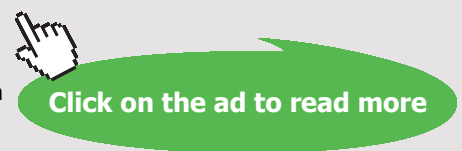
Apply now at
www.come.ku.dk



cultural studies

religious studies

science



The **Chi-square distribution** is one of the most widely used probability distribution in Hypothesis Testing as we will see in **Chapters 1 and 2 of Inferential Statistics – The Basics for Biostatistics**. It is used for constructing confidence intervals, as shown in **Chapter 1, Part II**. When there is a need to contrast it with the **Non-central Chi-square distribution**, this distribution is sometimes called the **Central Chi-square distribution**.

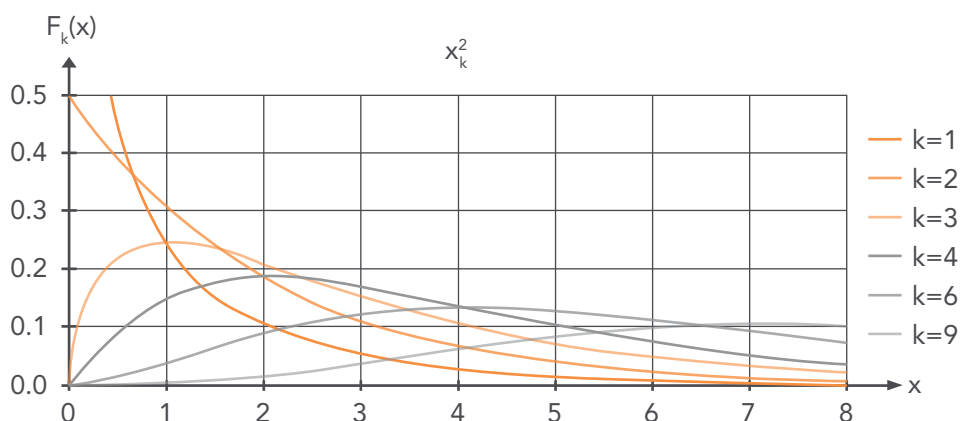


Figure 15-A Probability density function (Source: Internet)

As it can be seen from **Figure 15-A**, the graph of the **PDF** for the χ^2 random variable is not symmetric. It is positively skewed, as it was the case for the **Exponential** and **Gamma** distributions. As the number of degrees of freedom increases the graph will become flatter and almost symmetric about its mean. Because of the non-symmetry, values for the lower and upper tail areas need to be tabulated in order to use the **Chi-square** in calculating the confidence interval on the variances and standard deviations, as well as, for testing on them. Thus, we will discuss the following two examples.

EXAMPLE 5.11: Let X have a Chi-square distribution with $k = 4$ degrees of freedom. Find C if $P(X < C, k = 4)$ is: a) 0.95, b) 0.90 c) 0.975 d) 0.99.

Solution: C will represent the Chi-square value, $\chi_{1-\alpha, k}^2$, on the x -axis, such that the area to the left of it, between 0 and C , and under the curve is $1-\alpha$, with k degrees of freedom. Using the Chi-square Distribution **Table VI**, we find that

- a) $C = 9.488$, b) $C = 7.779$, c) $C = 11.14$, and d) $C = 13.28$.



Due to the non-symmetry property of the **Chi-square Distribution**, let us check the values of χ^2 such that the area to the right is α with k degrees of freedom. The values of the **Chi-square** random variable X , are given in **Table VI**, for some special values on the area to the right, and under the curve with degrees of freedom in the first column.

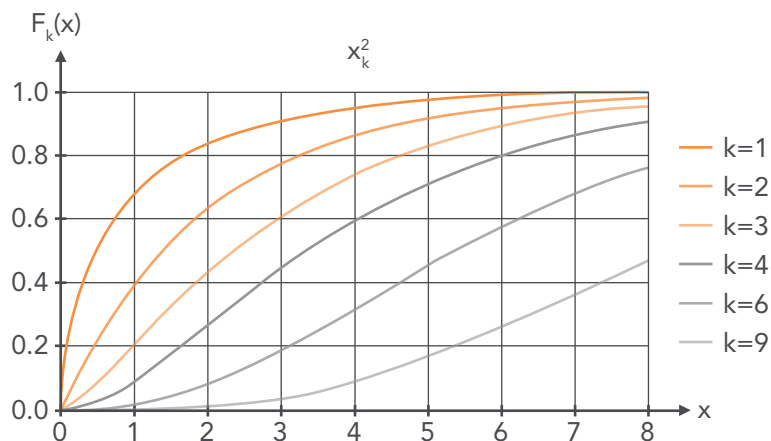


Figure 15-B Cumulative distribution function (Source: Internet)

EXAMPLE 5.12: Let X have a Chi-square distribution with $k = 4$ degrees of freedom. Find C if $P(X > C, k = 4)$ is: a) 0.9 b) 0.90 c) 0.975 d) 0.99.

Solution: C will represent the Chi-square value, $\chi_{\alpha,k}^2$, on the x -axis, such that the area to the right of it, between 0 and C , and under the curve is $1-\alpha$, with k degrees of freedom. Using the **Chi-square** Distribution Table, **Table VI**, we find that

- a) $C = 0.711$, b) $C = 1.064$, c) $C = 0.484$, and d) $C = 0.297$.



The **Chi-square** distribution is used in the common Chi-square tests for Goodness-of-Fit to a theoretical model, and for the independence of two, or more, criteria of classification of qualitative data. In addition to those tests, the **Chi-square distribution** is used in the calculations of the confidence interval on the variance and the standard deviation, of a normal distribution. Many other statistical tests also use this distribution, like Friedman's analysis of variance by ranks, which we are not going to present here.

5.4.4 F-DISTRIBUTION

The **F-distribution** is another non-negative continuous probability distribution. It is also known as **Snedecor's F-distribution** or the **Fisher-Snedecor distribution** (after R.A. Fisher and George W. Snedecor). The **F-distribution** arises frequently in the analysis of variance, when testing on the equality between more than two means. As it was the case with the **Exponential** and **Chi-square** distributions, being special cases of the Gamma distribution, we find that the **F-distribution** is related to the **Chi-square** distribution. It is being defined

as the ratio of two independent Chi-square random variables: $U \sim \chi^2(d_1)$ and $V \sim \chi^2(d_2)$, with d_1 and d_2 degrees of freedom respectively. Thus, the random variable defined by

$$F = \frac{U/d_1}{V/d_2},$$

has the **F-distribution**, with d_1 and d_2 as the degrees of freedom for the numerator and denominator respectively. As with the most non-negative continuous random variable, the **F-distribution** is positively skewed, when d_1 and d_2 are small. The interested reader is referred to **Hogg and Tanis (2010), Chapter 5**, for more details about the **F-distribution**, its derivation, and its properties. Similar to the Chi-square distribution, the tabulated values, for the **F-distribution** are restricted and few values are given in **Table VII**.

Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF

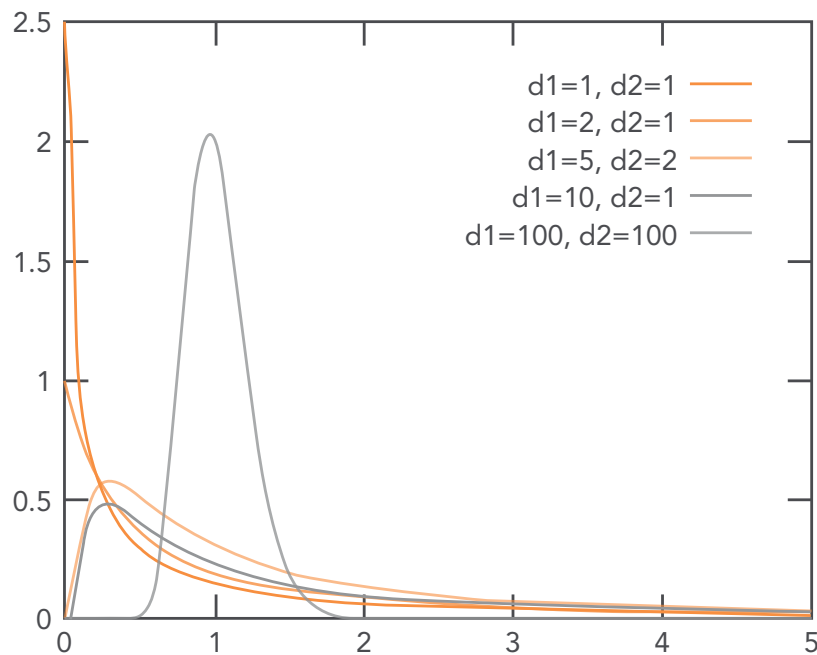


Figure 16 The pdf for the F-Distribution with degrees of freedom as noted
(Source: Internet)

The pdf, of a random variable with the F-distribution, is given by:

$$f(x; d_1, d_2) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}, \quad x \geq 0,$$

where d_1 , and d_2 are positive integers, with $B(m, n)$ being the **Beta** function, with parameters m and n , see **Whittaker and Watson (1965)**.

The **F-distribution** is introduced herein to familiarize the reader with it. Thus when we refer to it for testing on the ratio of two variances, or in the **ANOVA** analysis, (check **Chapter 2 and 12**), the reader has been made aware of the setting.

As for the cdf values of **Chi-square distributio** in **Table VI, Table VII**, for the corresponding F-distribution values, is also limited. Aside from the restrictions on the values in **Tables VI, and VII**, we still can use them adequately for most of the applications herein. For the notation to be adopted and used here we have; if X has an **F-distribution** with d_1 , d_2 degrees of freedom for the numerator and denominator, respectively, we say that $X \sim F(d_1, d_2)$. For a right-tail probability of α , we write

$$P[X \geq F_\alpha(d_1, d_2)] = \alpha.$$

For a left-tail probability of α , where α is generally small; .01, 0.05, 0.10, we see that if $X \sim F(d_1, d_2)$, then the distribution of $1/X$ is $F(d_2, d_1)$. Since

$$\alpha = P[X \leq F_{1-\alpha}(d_1, d_2)] = P[1/X \geq 1/F_{1-\alpha}(d_1, d_2)],$$

and

$$P[1/X \geq F_\alpha(d_2, d_1)] = \alpha,$$

it follows that

$$1/F_{1-\alpha}(d_1, d_2) = F_\alpha(d_2, d_1) \quad \text{or} \quad F_{1-\alpha}(d_1, d_2) = 1/F_\alpha(d_2, d_1).$$

EXAMPLE 5.13: Let the distribution of X be $F(4, 6)$. Find

- a) $F_{0.05}(4, 6)$ b) $F_{0.99}(4, 6)$ c) $F_{0.90}(4, 6)$ d) $F_{0.10}(6, 4)$ e) $F_{0.05}(6, 4)$

Solution: From **Table VII**, we see that

- a) $F_{0.05}(4, 6) = 4.5$ b) From **Table VII**, we see that $P(0 \leq X \leq 9.15) = 0.99$. That is $F_{0.01}(4, 6) = 9.15$.
 c) From **Table VII**, we have $P(0 \leq X \leq 3.18) = 0.90$, thus $F_{0.10}(4, 6) = 3.18$.
 d) $F_{0.10}(6, 4) = 4.01$, i.e., $P(X \geq 4.01; X \sim F(6, 4)) = 0.10$, and
 e) $F_{0.05}(6, 4) = 6.16$, i.e., $P(X \geq 6.06; X \sim F(6, 4)) = 0.05$

5.5 BIVARIATE RANDOM VARIABLES (OPTIONAL)

We have introduced, in **Chapter 4**, that observations, or distributions, on two or more quantitative variables are often recorded and found in real life, and in many applications of sciences. In the following subsection, as it was the case in **Chapter 4** when two discrete random variables can be related and get to study the behavior based on their probability mass functions, we will devote the discussion to the continuous case, based on two random variables. We refer the interested reader to check on the volumes of univariate and multivariate continuous distributions written by **Johnson, N.L. and Kotz, S. (1969–1972)**.

5.5.1 BIVARIATE CONTINUOUS RANDOM VARIABLES

The notion that was explored in **Section 4.5.1**, for two discrete random variables, can be applied, and extended, to two continuous random variables. No doubt now that the joint space for X and Y will be a part, or a subset, of the **Cartesian** xy -plane. Thus, the subspace for either variable will be an interval on the real line. When X and Y are continuous random variables, the joint probability function, $f(x, y)$ will be labeled as the joint probability density function, pdf, for the two continuous random variables X and Y , with the following properties:

1. $f(x, y) \geq 0$ over the common space of X , and Y .

$$2. \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

$$3. P[(X, Y) \in A] = \iint_A f(x, y) dx dy$$

where $\{(X, Y) \in A\}$ is an event defined within the space of X and Y , or where A is a region in the xy -plane.

Trust and responsibility

NNE and Pharmaplan have joined forces to create NNE Pharmaplan, the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries.

Inés Aréizaga Esteva (Spain), 25 years old
Education: Chemical Engineer

– You have to be proactive and open-minded as a newcomer and make it clear to your colleagues what you are able to cope. The pharmaceutical field is new to me. But busy as they are, most of my colleagues find the time to teach me, and they also trust me. Even though it was a bit hard at first, I can feel over time that I am beginning to be taken seriously and that my contribution is appreciated.



NNE Pharmaplan is the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries. We employ more than 1500 people worldwide and offer global reach and local knowledge along with our all-encompassing list of services.
nnepharmaplan.com

nne pharmaplan®



In addition to the above conditions on the joint pdf for the two continuous random variables X and Y we have, in analogous with the discrete case above:

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)dy, \quad \text{and} \quad f_y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

The two new defined functions, namely, $f_x(x)$ and $f_y(y)$, are the marginal pdfs for X and Y , respectively.

It is to be noticed in this section, that the summation sign that appeared in **Section 4.5.1** will be replaced by the integral sign in order to define the joint pdf for the two continuous random variables X and Y . Recalling how we defined the expected value and the variance for one random variable in **Section 4.2**, the notion will be carried out here for the two continuous random variables. Thus, we see that, when A is the set where X and Y are both defined:

$$\begin{aligned} E(X+Y) &= \iint_A (x+y)f(x, y)dxdy \\ &= \int_{-\infty}^{\infty} x \left\{ \int_{-\infty}^{\infty} f(x, y)dy \right\} dx + \int_{-\infty}^{\infty} y \left\{ \int_{-\infty}^{\infty} f(x, y)dx \right\} dy \\ &= \int_{-\infty}^{\infty} x \cdot f_x(x)dx + \int_{-\infty}^{\infty} y \cdot f_y(y)dy \\ &= E(X) + E(Y). \end{aligned}$$

As it can be seen, from the above expression for $E(X+Y)$, it is the same for two discrete or continuous random variables. We just replace the summation signs by the integral signs. Utilizing the above **REMARK** and **THEOREM**, we find out that they apply to the case when we have two continuous random variables, as well. To clarify all the concepts, let us give an example.

EXAMPLE 5.14: Let X and Y have the joint pdf that is given by $f(x, y) = x + y$, $0 \leq x \leq 1$, $0 \leq y \leq 1$; and 0 otherwise.

- Find the marginal pdfs for X and Y , and show that X and Y are not independent.
- Compute the mean and the variance for each of X and Y .
- Calculate the correlation coefficient $\rho(X, Y)$.

Solution: a) The pdf of X is given by:

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)dy = \int_0^1 (x+y)dy = [xy + y^2/2]_0^1 = x + 1/2, \quad 0 \leq x \leq 1.$$

Similarly, the pdf of Y is

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 (x + y) dx = [x^2/2 + xy]_0^1 = y + 1/2, 0 \leq y \leq 1.$$

Clearly on multiplying $f_x(x) \cdot f_y(y) = (x + 1/2) \cdot (y + 1/2) = xy + (x + y)/2 + 1/4 \neq x + y = f(x, y)$. Thus, proving that X and Y are not independent.

d) For the mean of X we have: $\mu_x = \int_0^1 x \cdot (x + 1/2) dx = 7/12$, similarly we find $\mu_y = 7/12$.

For the variances, we have: $\sigma_x^2 = E(X^2) - \{E(X)\}^2$.

Thus, $E(X^2) = \int_0^1 x^2 \cdot (x + 1/2) dx = 5/12$, and therefore $\sigma_x^2 = 5/12 - (7/12)^2 = 11/144$. In the same way, we find that $\mu_y = 7/12$, and $\sigma_y^2 = 11/144$.

e) To calculate ρ we need to find the Cov (X, Y). Since $\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$, we have $E(XY) = \int_0^1 \int_0^1 x \cdot y \cdot (x + y) dx dy = \int_0^1 [\int_0^1 (x^2 y + xy^2) dx] dy = \int_0^1 [y/3 + y^2/2] dy = 1/3$

Thus $\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y) = 1/3 - (7/12)^2 = -1/144$. Therefore $\rho(X, Y) = - (1/144) / (144/11) = -1/11$.

❖ -----

5.5.2 CONDITIONAL PROBABILITY OF TWO RANDOM VARIABLES

We have discussed conditional probability of two events earlier in **Chapter 3**. We did not need to specify the type of the event as discrete or continuous, since events were subsets, or parts, of the common sample space of the experiment. In this section, we like to explore the conditional probabilities for two continuous random variables that are jointly related by a **probability density function** in the continuous case. As was the setup in the above two subsections of this chapter for the continuous random variables, we need to introduce the notion for the conditional probability for two random variables. Recalling the earlier notation,

$$P(A|B) = P(A \text{ and } B)/P(B), \text{ provided } P(B) \neq 0.$$

For our discussion here we will adopt the following notation: Let A be the region, in the Cartesian plane, where X and Y are defined. Building on this convention, we see

$$A \cap B = \{X = x, Y = y\}, \text{ and } P(A \cap B) = P(X = x, Y = y) = P(x, y), \text{ with } P(B) = P_y(y) > 0.$$

Therefore, the above conditional probability can be expressed as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(x, y)}{P_y(y)}, \text{ with } P_y(y) > 0.$$

Thus, we have the following definition:

The conditional probability mass function of the discrete random variables X , given $Y = y$, is defined as

$$P(x|y) = P(x, y) / P_y(y), \text{ when } P_y(y) > 0.$$

By the same convention, we see that the conditional probability mass function of Y , given $X = x$, is

$$P(y|x) = P(x, y) / P_x(x), \text{ when } P_x(x) > 0.$$

As it was the case when we dealt with two discrete random variables, **Chapter 4**, the process can be extended to the case of two continuous random variables just by replacing the summation sign by the integral sign. For a complete illustration, and to show the way for deriving the conditional probabilities, we give the example below.

This e-book
is made with
SetaPDF





SETASIGN

PDF components for PHP developers

www.setasign.com



EXAMPLE 5.15: Let $f(x, y) = 1/40$, $0 \leq x \leq 10$, $10 - x \leq y \leq 14 - x$, be the joint pdf of X and Y .

- Find the marginal pdf's $f_1(x)$.
- Determine the conditional pdf $h(y|x)$, of Y given $X = x$.
- Calculate $E(Y|x)$, the conditional mean of Y , given that $X = x$.

Solution:

- By using the notation above, we can write $f_1(x) = \int_{10-x}^{14-x} f(x, y) dy = \int_{10-x}^{14-x} (1/40) dy = 1/10$, $0 \leq x \leq 10$.
- The conditional pdf $h(y|x)$, of Y given $X = x$ is given by

$$h(y|x) = f(x, y) / f_1(x) = 1/4, \quad 10 - x \leq y \leq 14 - x.$$

- Thus, $E(Y|x) = \int_{10-x}^{14-x} \frac{y}{4} dy$. On integration, we have $E(Y|x) = 12 - x$.

In case you guessed the answer in **Example 4.25**, can you guess, for this example, how many $h(y|x)$ are there?



REMARK: In **Chapter 7**, when $E(Y|x)$ is linear in x and dependent on x , we will have what is then called the **Simple Linear Regression**.

CHAPTER 5 EXERCISES

- Weights of fish caught by a certain method are approximately normally distributed with mean of 4.5 lbs. and a standard deviation of 0.50 lbs. Find the
 - Percentage of fish will weigh less than 4 lbs?
 - Percentage of the fish will weigh within one lb. of the average weight?
 - What is the chance that one fish will weigh more than 5 lbs?
- The inside diameter of a piston ring is normally distributed with mean of 4 inches and a standard deviation of 0.01 inches.
 - What percentage of the rings will have an inside diameter exceeding 4.025 inches?
 - What is the probability that a piston ring will have an inside diameter between 3.99 and 4.01 inches?

- c) Below what value of the inside diameter will 15% of the rings fall?
- 5.3 Gauges are used to reject all components in which a certain dimension is not within the specifications of $1.5 - d$ and $1.5 + d$. It is known that this dimension is normal distributed with mean 1.50 and standard deviation 0.2. Determine the value of d such that the specifications cover
- 95% of the components,
 - 90% of the components,
 - 99.7% of the components
- 5.4 A sample space consists of five simple events, E_1 , E_2 , E_3 , E_4 , and E_5 .
- If $P(E_1) = P(E_2) = 0.15$, $P(E_3) = 0.4$, and $P(E_4) = 2P(E_5)$, find the probabilities of $P(E_4)$ and $P(E_5)$.
 - If $P(E_1) = 3P(E_2) = 0.3$, find the probabilities of the remaining simple events if you know that the remaining simple events are equally probable.
- 5.5 Using Table IV, in **Appendix A**, find:
- The upper 0.05 point when d.f. = 5,
 - The lower 0.025 point when d.f. = 18,
 - The lower 0.01 point when d.f. = 11
 - The upper 0.10 point when d.f. = 16.
- 5.6 Name the t -percentile and find the values for t when:
- The upper tail area under the curve is 0.10, and d.f. = 18,
 - The upper tail area under the curve is 0.05, and d.f. = 14,
 - The lower tail area under the curve is 0.10, and d.f. = 21,
 - The lower tail area under the curve is 0.025, and d.f. = 17,
 - The upper and the lower tail areas each is 0.025, and d.f. = 18,
 - The upper and the lower tail areas under the curve each is 0.05, and d.f. = 24.
- 5.7 By using **Table IV**, in **Appendix A**, find:
- The 90th percentile of the t -distribution when d.f. = 15,
 - The 99th percentile of the t -distribution when d.f. = 8,
 - The 5th percentile of the t -distribution when d.f. = 11,
 - The 10th percentile of the t -distribution when d.f. = 16,
 - The 95th percentile of the t -distribution when d.f. = 17,

f) The lower and upper quartiles of the t-distribution when d.f. = 19.

5.8 Find the following probabilities

- a) $P(T < -1.761)$ when d.f. = 14,
- b) $P(|T| > 2.306)$, when d.f. = 8,
- c) $P(-1.734 < T < 1.734)$, when d.f. = 18,
- d) $P(-1.812 < T < 2.764)$, and d.f. = 10.

5.9 In each case below, find the constant c that satisfies the given condition:

- a) $P[T < c] = 0.95$, when d.f. = 9,
- b) $P[T > c] = 0.95$, when d.f. = 16,
- c) $P[-c < T < c] = 0.95$, when d.f. = 18,
- d) $P[T > c] = 0.05$, when d.f. = 26,
- e) $P[T > c] = 0.99$, when d.f. = 15,
- f) $P[T > c] = 0.01$, when d.f. = 11.



Sharp Minds - Bright Ideas!

Employees at FOSS Analytical A/S are living proof of the company value - First - using new inventions to make dedicated solutions for our customers. With sharp minds and cross functional teamwork, we constantly strive to develop new unique products - Would you like to join our team?

FOSS works diligently with innovation and development as basis for its growth. It is reflected in the fact that more than 200 of the 1200 employees in FOSS work with Research & Development in Scandinavia and USA. Engineers at FOSS work in production, development and marketing, within a wide range of different fields, i.e. Chemistry, Electronics, Mechanics, Software, Optics, Microbiology, Chemometrics.

We offer
A challenging job in an international and innovative company that is leading in its field. You will get the opportunity to work with the most advanced technology together with highly skilled colleagues.

Read more about FOSS at www.foss.dk - or go directly to our student site www.foss.dk/sharpminds where you can learn more about your possibilities of working together with us on projects, your thesis etc.

Dedicated Analytical Solutions

FOSS
Slangerupgade 69
3400 Hillerød
Tel. +45 70103370
www.foss.dk

The Family owned FOSS group is the world leader as supplier of dedicated, high-tech analytical solutions which measure and control the quality and production of agricultural, food, pharmaceutical and chemical products. Main activities are initiated from Denmark, Sweden and USA with headquarters domiciled in Hillerød, DK. The products are marketed globally by 23 sales companies and an extensive net of distributors. In line with the corevalue to be 'First', the company intends to expand its market position.





- 5.10 Customers arrive randomly at a bank teller's window. Given that one customer arrived during a particular 10-minute period, let X equal the time within the 10 minutes that the customer arrived. Find: a) The pdf of X , b) $P(X \geq 8)$, c) $P(2 \leq X < 8)$ d) $E(X)$, e) $\text{Var}(X)$.
- 5.11 If the MGF of X is $M(t) = (1/t)^*[e^{5t} - e^{4t}]$, $t \neq 0$, and $M(0) = 1$, find:
- a) $E(X)$, b) $\text{Var}(X)$, and c) $P(4.2 < X \leq 4.7)$.
- 5.12 What are the PDF, the mean and the variance of X if the MGF is given by the following?
- a) $M(t) = (1 - 3t)^{-1}$, b) $M(t) = 3(3 - t)^{-1}$.
- 5.13 If X has a gamma distribution with a scale and shape parameters of 4 and 2 respectively, find $P(X < 5)$.
- 5.14 Give the proof for MGF the mean and the variance for the gamma distribution as depicted in the theorem of that section.
- 5.15 If $X \sim \chi(17)$, find: a) $P(X < 7.564)$, b) $P(X > 27.59)$, c) $P(6.408 < X < 27.59)$ d) $\chi_{0.95}^2(17)$, e) $\chi_{0.025}^2(17)$.
- 5.16 If X is $\chi^2(12)$, find constants a and b such that $P(a < X < b) = 0.90$ and $P(X < a) = 0.05$.
- 5.17 If the moment generating function of X is: $M(t) = (1 - 2t)^{-12}$, $t < 1/2$, find:
- a) $E(X)$, b) $\text{Var}(X)$, c) $P(15.66 < X < 42.98)$.
- 5.18 Find the mean and variance of an F random variable with d_1 and d_2 degrees of freedom by first finding $E(U)$, $E(1/V)$, $E(U^2)$, and $E(1/V^2)$.
- 5.19 Let the distribution of Y be $F(9, 24)$. Find the following:
- a) $F_{0.05}(9, 24)$,
 b) $F_{0.95}(9, 24)$,
 c) $F_{0.05}(24, 9)$,
 d) $F_{0.95}(24, 9)$, and

- e) $P(0.277 \leq Y \leq 2.70)$.
- 5.20 Let the distribution of Y be $F(d_1, d_2)$ such that $P(0.198 < Y < 8.98) = 0.95$.
- 5.21 Let $f(x, y) = 3/2, x^2 \leq y \leq 1, 0 \leq x \leq 1$, be the joint pdf of X and Y . Find:
- $P(0 \leq X \leq 1/2)$,
 - $P(1/2 \leq Y \leq 1)$,
 - $P(1/2 \leq X \leq 1, 1/2 \leq Y \leq 1)$, and
 - $P(X \geq 1/2, Y \geq 1/2)$,
 - Calculate $\rho(X, Y)$.
- 5.22 Let $f(x, y) = (3/16)xy^2, 0 \leq x \leq 2, 0 \leq y \leq 2$, be the joint pdf of X and Y . Find:
- the marginal pdf for X and Y ,
 - $\text{Cov}(X, Y)$,
 - Are X and Y independent? Why or why not?
- 5.23 Let X and Y have the joint pdf: $f(x, y) = 2, 0 \leq y \leq x \leq 1$.
- Find the marginal pdf of X and Y ,
 - Compute the mean for each of X and Y ,
 - calculate: i) the variance for each of X and Y , ii) $\text{Cov}(X, Y)$, iii) the correlation Coefficient.
- 5.24 Let X and Y have the joint pdf: $f(x, y) = 1/8, 0 \leq y \leq 4, y \leq x \leq y + 2$.
- Find the marginal pdfs of X and Y ,
 - Determine $h(y | x)$, the conditional pdf of Y , given that $X = x$, and compute $E(Y | x)$.
 - Determine $g(x | y)$, the conditional pdf of X , given that $Y = y$, and compute $E(X | y)$.
- 5.25 Let X have a uniform distribution on the interval $(0, 1)$. Given that $X = x$, let Y have a uniform distribution on the interval $(0, x + 1)$.
- Find $f(x, y)$, the joint pdf of X and Y .
 - Find $E(Y | x)$.
 - Find $f_2(y)$ and specify its domain.
- 5.26 Let X and Y have the joint pdf: $f(x, y) = cx(1 - y), 0 < y < 1, \text{ and } 0 < x < 1 - y$.
- Determine c , in order for $f(x, y)$ to be a joint PDF for X and Y .

b) Compute $P(Y < X | X < \frac{1}{4})$.

5.27 Using what has been found in the **Example 5.15**, calculate $\rho(X, Y)$.

Understanding the Concepts Exercises CHAPTER 5

1. What is a random variable?
2. What is the difference between a discrete random variable and a continuous random variable?
3. Can you set the differences between the pmf and the pdf of random variables?
4. Two characteristics describe the graph of the probability density function, can you name them?
5. Why do we use the Z-scores to find areas under the normal curve?
6. Do you need to calculate the mean and the variance for a normal distribution?
7. What are the similarities between the Normal Distribution and the Student's t-distribution?
8. Is there a difference between positive and non-negative Random variables?
9. Why are the mean and the variance for a random variable important?



"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

10. Can you name a few of continuous random variables other than the normal and Student's t ?

TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP Finding the Mean and Standard Deviation of a Discrete Random Variable

TI-83/84 Plus

1. Enter the values of the random variable in L1 and their corresponding probabilities in L2.
2. Press **STAT**, highlight **CALC**, and select **1: 1-Var Stats**.
3. With 1-VarStats on the HOME screen, type L1 followed by a comma, followed by L2 as follows: 1-Var Stats L1, L2

Hit **ENTER**.

TECHNOLOGY STEP-BY-STEP Computing Binomial Probabilities via Technology

TI-83/84 Plus

Computing $P(x)$

1. Press **2nd VARS** to access the probability distribution menu.
2. Highlight **0: Binompdf** (and hit **ENTER**).
3. With **Binompdf** (on the HOME screen, type the number of trials n , the probability of success p , and the number of successes, x , for example, with $n = 15$, $p = 0.3$, and $x = 8$, type **Binompdf (15, 0.3, 8)** Then hit **ENTER**.

Computing $P(X \leq x)$

1. Press 2nd **VARS** to access the probability distribution menu.
2. Highlight A: **Binomcdf** (and hit **ENTER**).
3. With **Binomcdf** (on the HOME screen, type the number of trials n , the probability of success p , and the number of successes, x , for example, with $n = 15$, $p = 0.3$, and $x \leq 8$, type **Binomcdf (15, 0.3, 8)** Then hit **ENTER**.

Excel

Computing $P(x)$

1. Click on the **fx** icon. Highlight **Statistical** in the Function category window. Highlight **BINOMDIST** in the Function name window
2. Fill in the window with the appropriate values. For example, if $x = 5$, $n = 10$, and $p = 0.2$, fill in the window. Click **OK**

Computing $P(X \leq x)$

Follow the same steps as those presented for computing $P(x)$. In the **BINOMDIST** window, type **TRUE** in the cumulative cell.

TECHNOLOGY STEP-BY-STEP Computing Poisson Probabilities via Technology

TI-83/84 Plus

Computing $P(x)$

1. Press 2nd **VARS** to access the probability distribution menu.
2. Highlight: **Poissonpdf** (and hit **ENTER**).
3. With **Poissonpdf** (on the HOME screen, type the value of Lambda, λ (the mean of the distribution), followed by the number of successes x , for example, type **Poissonpdf (10, 8)**
Then hit **ENTER**.

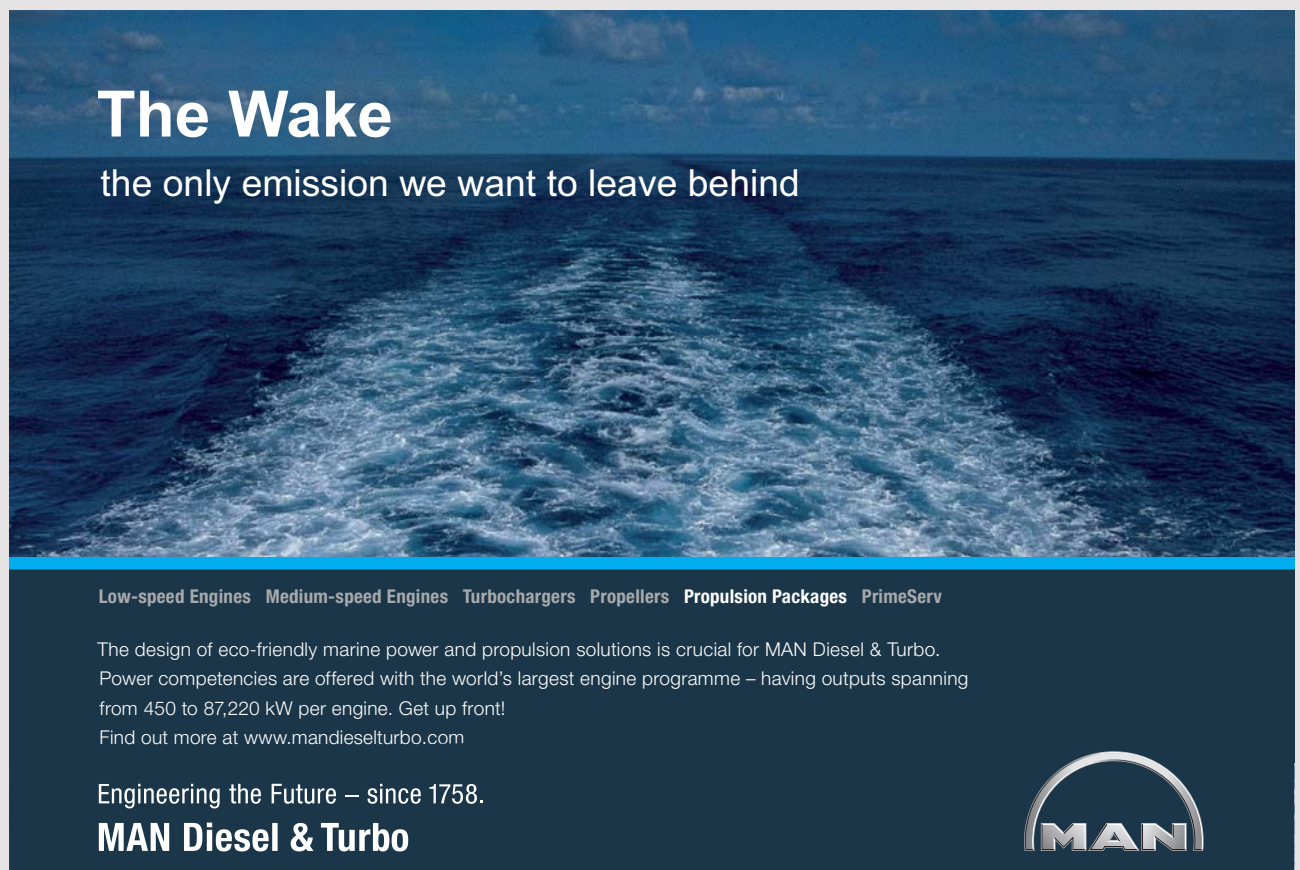
Computing $P(X \leq x)$

1. Press 2nd VARS to access the probability distribution menu.
2. Highlight: **Poissoncdf** (and hit ENTER).
3. With **Poissoncdf** (on the HOME screen, type the value of Lambda, λ (the mean of the distribution), followed by the number of successes x , for example, type **Poissoncdf (10, 8)** Then hit ENTER.

Excel

Computing $P(x)$

- a. Enter the desired values of the random variable X in column A.
- b. With cursor in cell B1, select the Formulas tab, Select more Formulas. Highlight Statistical, and then highlight POISSON in the function name menu.
- c. In the cell labeled X , enter A1 In the cell labeled mean, enter the mean (λ , Lambda). In the cell labeled cumulative, type FALSE. Click OK.



The Wake


the only emission we want to leave behind

Low-speed Engines Medium-speed Engines Turbochargers Propellers Propulsion Packages PrimeServ

The design of eco-friendly marine power and propulsion solutions is crucial for MAN Diesel & Turbo. Power competencies are offered with the world's largest engine programme – having outputs spanning from 450 to 87,220 kW per engine. Get up front! Find out more at www.mandieselturbo.com

Engineering the Future – since 1758.

MAN Diesel & Turbo



Computing $P(X \leq x)$

Follow the same steps as those presented for computing $P(x)$. In the **POISSON** window, type TRUE in the cumulative cell.

TECHNOLOGY STEP-BY-STEP the Standard Normal Distribution

TI-83/84 Plus

Finding Areas under the Standard Normal Curve

1. Press 2nd VARS to access the Distribution menu.
2. Select 2: Normalcdf (
3. With Normalcdf (on the HOME screen type *lower bound, upper bound, 0, 1*), for the standard Normal Distribution based on a given value for Z. For Example, to find the area left of $z = 1.26$ under the standard normal curve, type Normalcdf (-1E99, 1.26, 0, 1),
And hit ENTER.

Note: When there is no lower bound, enter -1E99. When there is no upper bound, enter 1E99. The E shown is for the scientific notation. It is selected by pressing 2nd then ‘.

Finding Z-Scores Corresponding to an Area

1. Press 2nd VARS to access the Distribution menu.
2. Select 3: Invnorm (
3. With Invnorm (on the HOME screen type “*area left*”, 0, 1). For example, to find the z-score such that the area under the normal curve to the left of the score is 0.79, type
Invnorm (0.79, 0, 1)
And hit ENTER.

Excel

Finding Areas under the Standard Normal Curve

1. Select the fx button from the tool bar. In **Function Category:** select “Statistical”. In **Function Name:** select “NORMDIST”. Click **OK**.

2. Enter the specified z-score. Click **OK**.

Finding Z-Scores Corresponding to an Area

1. Select the fx button from the tool bar. In **Function Category**: select “Statistical”.
In **Function Name**: select “**INVNORM**”. Click **OK**.
2. Enter the specified area. Click **OK**.

TECHNOLOGY STEP-BY-STEP The Normal Distribution

TI-83/84 Plus

Finding Areas under the Normal Curve

1. Press 2nd **VARS** to access the Distribution menu.
2. Select 2: **Normalcdf** (
3. With **Normalcdf** (on the HOME screen type *lower bound, upper bound, mu, sigma*).
For example, to find the area to the left of $x=35$ under the normal curve, with $\mu = 40$ and $\sigma = 10$, type
Normalcdf (-1E99, 35, 35, 10)
And hit **ENTER**.

Note: When there is no lower bound, enter -1E99. When there is no upper bound, enter 1E99. The E shown is scientific notation; it is selected by pressing 2nd then ‘.

Finding Normal Values Corresponding to an Area

1. Press 2nd **VARS** to access the Distribution menu.
2. Select 3: **Invnorm** (
3. With **Invnorm** (on the HOME screen type “*area to the left*”, 0, 1). For example, to find the z-score such that the area under the normal curve to the left of the value 0.68, with $\mu = 40$ and $\sigma = 10$, type
Invnorm (0.68, 40, 10)
And hit **ENTER**.

Excel

Finding Areas under the Normal Curve

1. Select the fx button from the tool bar. In **Function Category**: select “Statistical”. In **Function Name**: select “NORMDIST”. Click **OK**.
2. Enter the specified observation, mu, and sigma, and set **Cumulative** to TRUE. Click **OK**.

Finding Normal Values Corresponding to an Area

1. Select the fx button from the tool bar. In **Function Category**: select “Statistical”. In **Function Name**: select “NORMINV”. Click **OK**.
2. Enter the specified area left of the unknown normal value, mu, sigma, Click **OK**.

TECHNOLOGY STEP-BY-STEP Normal Probability Plots

gaiteye[®]
Challenge the way we run

**EXPERIENCE THE POWER OF
FULL ENGAGEMENT...**

.....

**RUN FASTER.
RUN LONGER..
RUN EASIER...**

READ MORE & PRE-ORDER TODAY
WWW.GAITEYE.COM

The advertisement features a runner in a red shirt and black leggings on a dirt path. Technical diagrams, including circles and lines, are overlaid on the runner's feet, suggesting motion analysis or shoe technology. A yellow button with a hand cursor icon is positioned at the bottom right of the ad.

TI-83/84 Plus

1. Enter the raw data into L1.
2. Press 2nd Y= to access **STATPLOTS**.
3. Select 1: Plot 1.
4. Turn plot 1 on by highlighting on and pressing ENTER. Press the down-arrow key. Highlight the normal probability plot icon. It is the icon in the lower-right corner under Type: Press ENTER to select this plot type. The Data List should be set at L1. The Data axis should be the x-axis.
5. Press ZOOM, and select 9: Zoom Stat.

Excel

1. Install Data Desk XL.
 2. Enter the raw data into column A
 3. Select the **DDXL** menu. Highlight **Charts and Plots**.
 4. In the pull-down menu, select Normal Probability Plots. Drag the column containing the data to the Quantitative Variable cell and click **OK**. If the first row contains the variable Name, check the “First Row is variable name” box.
-

6 SAMPLING DISTRIBUTIONS

6.1 INTRODUCTION

Due to the difficulty of dealing with large populations, and the scarcity of funds and time, it is almost certain that we appeal to a **Simple Random Sample, SRS**, from that population in order to estimate those parameters of interest. If the sample is taken, and calculations have been done, will that be enough to get all the information needed just by using that sample? What if another simple random sample has been taken earlier, or later, will the results be the same? No doubt, there will be as many different estimated values, for the parameters of that population, as there are samples. Those observed values of the statistics, calculated from different samples, even taken from the same population will be different, and there is some variability in them. This variability is called the **Sampling Variability**. Based on that, it is quite important, in statistics, to check on the sampling distribution of the different statistics that were calculated from different samples, when we are estimating the parameters in the population. This is the topic of **Chapter 6**.

In **Chapter 3** we dealt with **Probability**. **Chapters 4 and 5** were devoted to discuss **Random Variables, Discrete and Continuous**, and their distributions respectively. In those two chapters, **Chapters 4 and 5**, the discussion was done, in general, on probability distributions which describe the populations. In **Chapter 6** we are going back to use data obtained by a simple random sample in order to see how the statistics of those samples, and the populations' parameters are related.

In general, we will be looking at the **Sampling distributions** of some statistics.

Definition 6.1: The Sampling Distribution of a statistic is a probability distribution for all possible values of that statistic computed from a simple random sample of size n .

In **Section 2**, we will display, and show how to construct, the **Sampling Distribution** of the sample mean \bar{X} . **Section 3** will be devoted to a very famous theorem in Statistics; namely, the **Central Limit Theorem**, and its applications for finding a **Sampling Distribution** of a statistic. **Section 4** will have the description on the **Sampling Distribution** of the sample proportion \hat{p} . In **Section 5**, the discussion will be about the sampling distribution of the sample variance.

6.2 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Let us begin with the definition of **The Sampling distribution**.

Definition 6.2: The **Sampling Distribution** of the sample mean, \bar{X} , for a given sample size n , is the probability distribution of all possible values of the random variable \bar{X} computed from simple random samples taken from a population with mean μ and standard deviation σ .

The procedure for getting a sampling distribution of the sample mean \bar{X} , is given by the following steps:

1. Obtain a simple random sample of size n .
2. Calculate the sample mean.
3. Assume that we have a finite population, in order that we can do something practical, and in a short time.
4. Repeat Steps 1 and 2 until all simple random samples of size n have been obtained, with the understanding that once a particular sample has been obtained, it cannot be considered again.

For illustration, consider the following example.



**Technical training on
WHAT you need, *WHEN* you need it**

At IDC Technologies we can tailor our technical and engineering training workshops to suit your needs. We have extensive experience in training technical and engineering staff and have trained people in organisations such as General Motors, Shell, Siemens, BHP and Honeywell to name a few.

Our onsite training is cost effective, convenient and completely customisable to the technical and engineering areas you want covered. Our workshops are all comprehensive hands-on learning experiences with ample time given to practical sessions and demonstrations. We communicate well to ensure that workshop content and timing match the knowledge, skills, and abilities of the participants.

We run onsite training all year round and hold the workshops on your premises or a venue of your choice for your convenience.

For a no obligation proposal, contact us today at training@idc-online.com or visit our website for more information: www.idc-online.com/onsite/

OIL & GAS ENGINEERING
ELECTRONICS
AUTOMATION & PROCESS CONTROL
MECHANICAL ENGINEERING
INDUSTRIAL DATA COMMS
ELECTRICAL POWER

Phone: +61 8 9321 1702
Email: training@idc-online.com
Website: www.idc-online.com

IDC TECHNOLOGIES

EXAMPLE 6.1: In a graduate class of Stat 5335 with 8 students, the instructor is interested in finding the distance, in miles, that each student had driven to come to class. The data came up as follows: 2, 4, 6, 8, 5, 7, 9, and 10.

- Construct a Sampling distribution of the sample mean \bar{x} for samples of size $n = 2$.
- What is the probability of obtaining a sample mean between 5 and 7 miles inclusive, i.e, what is $P(5 \leq \bar{x} \leq 7)$?

Solution: We will follow the above procedure to construct the sampling distribution. There are 8 individuals in the population. We are selecting them two at a time without replacement. Therefore, the sample size to be taken, in this case, is $n = 2$. Thus there are ${}_8C_2 = 28$ samples to be considered. Those samples are listed below with their corresponding means, as shown in **Table 1**:

Sample:	2, 4	2, 5	2, 6	2, 7	2, 8	2, 9	2, 10	4, 5	4, 6	4, 7	4, 8
Mean:	3	3.5	4	4.5	5	5.5	6	4.5	5	5.5	6
Sample:	4, 9	4, 10	5, 6	5, 7	5, 8	5, 9	5, 10	6, 7	6, 8	6, 9	6, 10
Mean:	6.5	7	5.5	6	6.5	7	7.5	6.5	7	7.5	8
Sample:	7, 8	7, 9	7, 10	8, 9	8, 10	9, 10					
Mean:	7.5	8	8.5	8.5	9	9.5					

Table 1 The 28 Samples with their Means

Table 2 will show the sampling distribution of the sample mean, \bar{x} based on the values in **Table 1**.

From **Table 2** we can compute

$$P(5 \leq \bar{x} \leq 7) = 2/28 + 3/28 + 3/28 + 3/28 + 3/28 = 14/28 = 0.50.$$

What the above probability is saying is that in case we can take 10 simple random samples of size 2, from the above population, there is a 50% chance that the mean will be between 5 and 7, inclusive.

Now, let us consider the above values of the sample means as a sample by themselves. In this case, we find that the mean of the sample means is exactly equal to the population mean. In other words, we have $\mu_{\bar{x}} = \mu = 6.375$. But how did that happen. Recall from **Chapter 2**, we showed there how to find the sample and the population means. In that sense, we have, for the sample mean and the population mean, respectively:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \text{and} \quad \mu = \frac{\sum_{i=1}^N x_i}{N}.$$

With the variances for the sample and the population as

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}, \quad \text{and} \quad \sigma^2 = \frac{\sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N}}{N}.$$

In addition, recall that we are taking the simple random sample from the population whose mean is μ and its finite and known variance is σ^2 .

Considering the population data of 2, 4, 6, 8, 5, 7, 9, and 10 we see that $\mu = 6.375$, and $\sigma^2 = 399/64$. Now let us look at the values of the sample means, and find their mean and their variance. Using Technology, we have $\mu_{\bar{x}} = \mu = 6.375$, $\sigma_{\bar{x}}^2 = 2.77083333$. These calculations will become very close to those of the population as n increases. Utilizing the properties of the expected value, and the variance of independent variables, we find

$$\mu_{\bar{x}} = E(\bar{x}) = E\left(\sum_{i=1}^n X_i / n\right) = \mu,$$

and

$$\sigma_{\bar{x}}^2 = Var(\bar{x}) = Var\left(\sum_{i=1}^n X_i / n\right) = (1/n^2) \cdot Var\left(\sum_{i=1}^n X_i\right) = \frac{\sigma^2}{n}.$$

These results should not be surprising based on a very famous theorem called **The Central Limit Theorem**, as we will see it in the next section. Despite the fact that the sample size that was considered in the above case was so small, just $n = 2$, we see that the bar graphs for the values of the means of those 28 samples are symmetric, and can be looked at as a roughly bell-shaped curve, see **Figure 1**.

6.3 CENTRAL LIMIT THEOREM

In **Section 2**, the sample mean, \bar{X} , of a sample of size n from a population whose mean is μ and its variance σ^2 , is itself a random variable until the sample is obtained. It was found that this random variable has a distribution in terms of the parameters of the population from which it has been drawn. That random variable has the following properties:

Sample Mean	Frequency	Prob.	Sample Mean	Frequency	Prob.
3	1	1/28	6.5	3	3/28
3.5	1	1/28	7	3	3/28
4	1	1/28	7.5	3	3/28
4.5	2	2/28	8	2	2/28
5	2	2/28	8.5	2	2/28
5.5	3	3/28	9	1	1/28
6	3	3/28	9.5	1	1/28

Table 2 The Probability Distribution of the Sample Means

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements



MAERSK

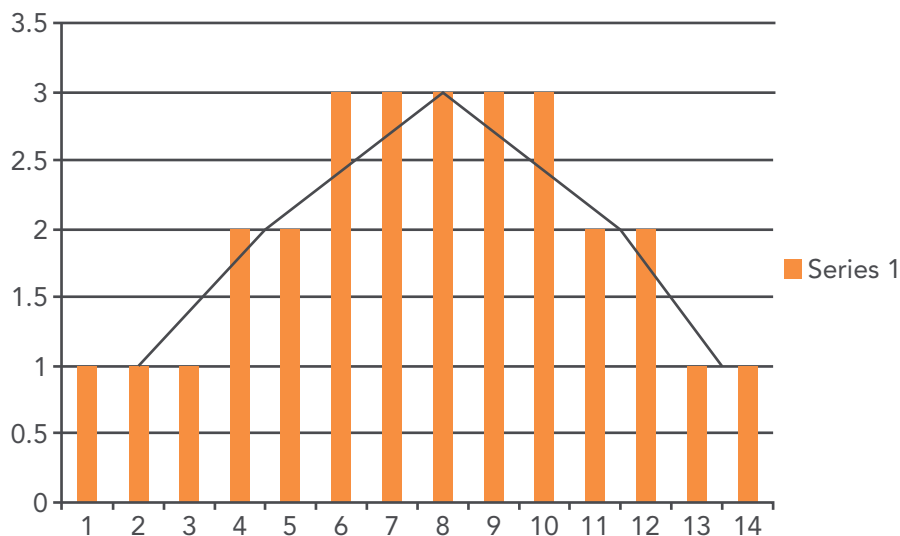


Figure 1 The bar graph of the Sample means

$$E(\bar{X}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

From the above expressions, it is clearly obvious that the distribution of \bar{X} depends on n , the sample size. Moreover, the variance, of the sampling distribution of \bar{X} , gets smaller and smaller as n increases. Thus, for every sample size we have a distribution for \bar{X} . If sampling has been taken at random from a normal population, then we have $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, see **Hogg and Tanis**, 2010. But what if the population, we are sampling from, were not normal, then what will the distribution of \bar{X} be? It turns out that whether we are sampling from a normal population or not, finite population or infinite, the sampling distribution of \bar{X} is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large, $n \geq 30$. This wonderful, and very useful result, is an immediate consequence of a very famous theorem, called the **Central Limit Theorem, CLT**, which is stated as follows.

THEOREM 6.1: (Central Limit Theorem) If \bar{X} is the mean of a simple random sample, of size n , that was taken from a population with mean μ and variance σ^2 , then the limiting distribution of the following random variable defined by, as n increases, we see that the variable given by

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

has the **Standard Normal Distribution**, i.e. $Z \sim N(0, 1)$.

Let us see an application of the **Central Limit Theorem**.

EXAMPLE 6.2: An electric company manufactures special light bulbs that are normally distributed with life time of 1000 hours and a standard deviation of 100 hours. If a random sample of size 25 bulbs is taken from that firm production, find the probability that the average of those bulbs is less than 950 hours.

Solution: From the above discussion and presentation, we that $\bar{X} \sim N(1000, 400)$. Thus $P(\bar{X} < 950) = \text{Normalcdf}(-1E99, 950, 1000, 20) = 0.0062$, which is quite small.

EXAMPLE 6.3: Samples are selected from a uniform distribution on the interval [0, 12].

- a) Find the probability that the mean of a sample of 25 observations is at least 6.75.
- b) Find the probability that the mean of a sample of 49 observations is at least 6.75.

Solution: We know that the mean of X is 6 and its variance is 12. Thus, we have for a) when $n = 25$, $\mu_{\bar{x}} = 6$, and $\sigma_{\bar{x}}^2 = 12/5$. Therefore $P(\bar{X} \geq 6.75) = \text{Normalcdf}(6.75, 1E99, 6, \sqrt{12/25}) = 0.1395$, by using Technology.

Another way is by standardizing, i.e. using the transformation $Z = (\bar{X} - \mu_{\bar{x}})/\sigma_{\bar{x}}$, and hence $P(\bar{X} \geq 6.75) = P[(\bar{X} - \mu_{\bar{x}})/\sigma_{\bar{x}} \geq (6.75 - 6)/\sqrt{12/25}] = P(Z \geq 1.08) = 1 - P(Z \leq 1.08) = 1 - 0.8599 = 0.1401$. It is clearly noticeable that there is a difference in the answers. This is due to the rounding in the Z -value to two decimal places, in order to use the Standard Normal Table.

For Part b) we have, by using the Standard Normal Table: $P(\bar{X} \geq 6.75) = P[(\bar{X} - \mu_{\bar{x}})/\sigma_{\bar{x}} \geq (6.75 - 6)/\sqrt{12/25}] = P(Z \geq 1.5155) = 1 - P(Z \leq 1.5155) = 1 - P(Z \leq 1.52) = 1 - 0.9357 = 0.0643$. (The reader can check, using Technology, that the answer is 0.0648.)



6.3.1 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN WHEN σ IS UNKNOWN

We have seen when we are given the population variance that the random variable

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution with mean 0 and variance 1. As it was clear, we did not pay attention to the sample size n , as long as n is large enough. The question that arises is what is the distribution of the sample mean, \bar{X} , when the sample size is small and the population variance is unknown. This case is worth looking at from a practical point of view.

As we found out that when n is large the **Central Limit Theorem** applies, and we can have a standard normal distribution to rely on regardless whether the population variance is known or not. This is so because we can use s for σ . However, when it comes to the case when n is small and σ is not known, the random variable, T , given by $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$, is not quite known, and its sampling distribution is yet to be determined. We find that we have to assume that the population from which the sample has been taken to a normal population. Along that line, we have the following Theorem.

Theorem 6.2 If \bar{X} is the mean of a random sample of size n taken from a normal population with mean μ , and unknown variance σ^2 , with

$S^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}$, then the random variable $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ will have a Student's t-distribution with parameter $\nu = n-1$, as its degrees of freedom. The t-distribution has been presented and discussed in **Chapter 5**.



www.job.oticon.dk

oticon
PEOPLE FIRST



6.4 DISTRIBUTION OF THE SAMPLE PROPORTION

It is quite often that the interest of the investigator does not involve a quantitative characteristic of a population, rather than it is related to a qualitative property, or aspect, in that population. For example, on the first day of a class in **Stat 3223**, the instructor likes to know: how many students in the class who have a calculator? In another aspect, how many students have a foreign-made car? Being in the age of technology, the question might be of interest is that: How many students are there who have a cell phone? As it can be seen, and it will not be surprising to see that there are more students with a cellphone than there are students with a calculator. Thus, as the questions have indicated that we are interested in the proportion of the individuals in that population that has a certain property.

Suppose that a random sample of size n is obtained from a population in which the items, in that population, have a certain property of interest to the researcher. We will assume that each individual, of that population, either does or does not have that characteristic. Thus, we see that the **Sample Proportion**, denoted by \hat{p} , is given by $\hat{p} = x/n$, where x is the number of individuals, of that population, that have the characteristic of interest. Without any doubt, \hat{p} is a random variable until the sample is taken. Being dependent on the observed sample values, and its size, we find that \hat{p} is a statistic which will estimate the population proportion, p . The sample size, n , as it was stated in **Chapter 4**, plays an important role in finding the sampling distribution of the statistic \hat{p} , when we dealt with the **Binomial distribution**, by looking at the probability of success as the proportion of the items with the specific property of interest. From **Chapter 4**, it was found that

$$E(\hat{p}) = p, \text{ and } \text{Var}(\hat{p}) = p(1-p)/n.$$

Utilizing the above famous theorem, i.e., **CLT**, we find that the sampling distribution of \hat{p} has the following properties, when a simple random sample of size n is taken from a population with proportion p :

- 1) The shape of the sampling distribution of \hat{p} is approximately normal provided $np(1-p) \geq 10$.
- 2) The mean of the sampling distribution is $\mu_{\hat{p}} = E(\hat{p}) = p$.
- 3) The variance of the sampling distribution of \hat{p} is $\sigma_{\hat{p}}^2 = \text{Var}(\hat{p}) = p(1-p)/n$.
- 4) On top of the above conditions, the sampled values must be independent as it was the case for the **Binomial distribution**. In other words, one outcome does not affect the success or failure of any other outcome. This condition will be satisfied when sampling from a finite population of size N by verifying that the sample size n is no more than $0.05N$, i.e., $n \leq 0.05N$.

After describing the sampling distribution of the sample proportion, we can do some examples on calculating the probabilities of a certain characteristic in a population.

EXAMPLE 6.4: A simple random sample of size $n = 75$ is taken from a population with $N = 10,000$ and with a specified characteristic given by $p = 0.8$.

1. Describe the sampling distribution of \hat{p} .
2. What is the probability of getting at least 63 individuals with the characteristic?
3. What is the probability of getting at most 51 individuals with the characteristic?

Solution: Based on the above discussion, we find that the sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.8$ and variance $\sigma_{\hat{p}}^2 = 0.8(1 - 0.8)/75 = 16/7500 = 4/1875$, and thus $\sigma_{\hat{p}} = 0.046188$.

1. For $x \geq 63$, we need to find $P(X \geq 63) = P(\hat{p} \geq 0.84) = \text{Normalcdf}(0.84, 1E99, 0.8, 0.046188) = 0.1932379586$, based on Technology and using **TI-83 Plus Calculator**. Interested reader, using the **Standard Normal Table**, can check the answer of $P(X \geq 63) = P(\hat{p} \geq 0.84) = 0.1922$.
2. For $x \leq 51$, we need to find $P(X \leq 51) = P(\hat{p} \leq 0.68) = \text{Normalcdf}(-1E99, 0.68, 0.8, 0.046188) = 0.0046874011$ based on Technology and using TI-83 Plus Calculator. Interested reader, using the **Standard Normal Table**, can check the answer of $P(X \leq 51) = P(\hat{p} \leq 0.68) = 0.0047$.

❖ -----

Here is another example on the sampling distribution of the sample proportion.

EXAMPLE 6.5: According to the National Center for Health Statistics, 15% of all Americans have hearing trouble.

- a) In a random sample of 120 Americans, what is the probability of at most 12% have hearing trouble?
- b) Suppose that a random sample of 120 Americans who regularly listen to music using headphones results in 26 having hearing trouble. What might you conclude?

Solution: Based on some census statistics, there are more than 300 million people in USA. Since the sample size is $n = 120 < 0.05(300 \text{ Million}) = 15 \text{ million}$, and $np(1 - p) = 120 \cdot (0.15) \cdot (0.85) = 15.3 > 10$, then the shape of the distribution of the sample proportion is approximately normal. Thus, we have:

- a) Since we are given $\hat{p} = 0.15$, and $\sigma_{\hat{p}} = \sqrt{\frac{0.15 \cdot (1-0.15)}{120}} = 0.032596012$, we can look up the table values. Let us use the Standard Normal Table for finding the probability that $\hat{p} \leq 0.12$, i.e. What is $P(\hat{p} \leq 0.12)$? It is the same as $P(Z \leq -0.92) = 0.1788$.
- b) A random sample of 120 Americans resulted in 26 having hearing trouble implied that $\hat{p} = 26/120 = 0.217$. We like to check on this value of \hat{p} being unusual by finding the following probability $P(\hat{p} \geq 0.217)$. By using the **Standard Normal Table**, as before, we find that $P(\hat{p} \geq 0.217) = P(Z \geq 2.06) = 0.0197$. This is a low probability, and implies that getting 26 in 120 Americans with hearing trouble is unusual when the proportion is 0.15.



6.5 SAMPLING DISTRIBUTION OF THE SAMPLE VARIANCE

So far, in the preceding sections of this Chapter, we have discussed the sampling distributions of two important statistics that relate to two important parameters of any population, namely, the sample mean and the sample proportion. It is clearly now that there is another statistic that relates to an important parameter of the population. Precisely, we are talking



Schlumberger

WHY WAIT FOR PROGRESS?

DARE TO DISCOVER

Discovery means many different things at Schlumberger. But it's the spirit that unites every single one of us. It doesn't matter whether they join our business, engineering or technology teams, our trainees push boundaries, break new ground and deliver the exceptional. If that excites you, then we want to hear from you.

careers.slb.com/recentgraduates



about the sample variance, S^2 , and its counterpart in the population, namely, the population variance, σ^2 . Thus, in this section we will address the sampling distribution of the sample variance for random samples, of size n , taken from normal populations. Without any doubt, the sampling distribution of Sample variance can have neither a normal distribution, nor a t - distribution. Why? S^2 is never negative. Checking the continuous distributions in **Chapter 5**, we find ourselves that this distribution is related to another continuous random variable, namely the **Gamma** distribution with special values for its parameters. Recalling from **Chapter 5**, when $X \sim \text{Gamma}(\alpha, \beta)$, the random variable with $\alpha = \nu/2$ and $\beta = 2$ has a **Chi-square** distribution, χ^2 , such using the square of the Greek letter (Chi), χ . Hence, we have the following theorem. The importance of studying the sampling distribution of the sample variance is quite relevant and important especially when we study the variation in the measurements we take.

THEOREM 6.3: If S^2 is the sample variance, that is calculated based on a sample of size n taken from a normal population with finite and unknown variance σ^2 then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

Will have a **Chi-square** distribution with $\nu = n - 1$, degrees of freedom.

Proof: By utilizing the definition of the sample variance, we have

$$\begin{aligned} \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} &= \sum_{i=1}^n \left[\frac{(X_i - \bar{X} + \bar{X} - \mu)^2}{\sigma^2} \right] \\ &= \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n} \end{aligned}$$

Again, based on the left-hand side, above, being the sum of the squares of independent standard normal random variables that has a **Chi-square** distribution with n degrees of freedom, we find that the right-hand side is a sum of two independent **Chi-square** distributions with degrees of freedom $n - 1$ and 1 respectively. Hence the result as ascertained in the **THEOREM 6.3**. (Interested readers might like to check references on mathematical statistics.)

Corollary 6.3.1:

- The expected value of the sample variance, S^2 , is the population variance σ^2 .
- The variance of the sample variance is given by $2\sigma^4/(n - 1)$.

EXAMPLE 6.6: A sample of size 9 is selected from a normal distribution $N(14, 4)$. Find $P(S^2 > 6)$.

Solution: $P(S^2 > 6) = P\left(\frac{(n-1)S^2}{\sigma^2} > 8 \cdot 6/4\right) = P(\chi_8^2 > 12) = 1 - P(\chi_8^2 < 12) = 1 - \chi^2 \text{cdf}(0, 12, 8) = 1 - 0.8487961172 = 0.1512038828$, by using a **TI-84 Plus Silver Edition** calculator.



Understanding the Concepts Exercises CHAPTER 6

1. What is a parameter? What is a statistic?
2. When will the boxplot indicate that the data is positively skewed?
3. What does it mean to have a sampling distribution?
4. Regardless of the distribution of the population from which we had taken the sample, what are the mean and the standard deviation of the sampling distribution of the sample mean?
5. There are two parameters of utmost interest in statistics that we care about their sampling distributions, what are they?
6. There is another parameter that we are interested in its sampling distribution, can you name it?
7. Under what conditions is the sampling distribution of the sample mean is normal?
8. Why do we use samples not populations?
9. What do we call the characteristics of a population?
10. What do we call the characteristics of a sample?

CHAPTER 6 EXERCISES

- 6.1 The IQ, X , of human beings is approximately normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. Compute the probability that a simple random sample of size $n = 10$ results in a sample mean greater than 110.
- 6.2 Determine $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ from the given parameters of the population, and the given sample size.
 - a) $\mu = 80, \quad \sigma = 14, \quad n = 49,$
 - b) $\mu = 64, \quad \sigma = 18, \quad n = 36.$
 - c) $\mu = 50, \quad \sigma = 10, \quad n = 25.$
 - d) $\mu = 27, \quad \sigma = 6, \quad n = 36.$

- 6.3 A simple random sample of size 10 is obtained from a normally distributed population with $\mu = 30$ and $\sigma = 8$. What is the sampling distribution of the sample mean \bar{x} ?
- 6.4 A simple random sample of size $n = 49$ is obtained from a population with $\mu = 80$ and $\sigma = 14$.
- Describe the sampling distribution of \bar{x} .
 - What is $P(\bar{x} > 83)$?
 - Calculate $P(78.3 < \bar{x} < 85.1)$?
 - Find $P(\bar{x} \leq 75.8)$.
 - True or False: $P(\bar{x} > 83) = 0.9332$
- 6.5 A simple random sample of size $n = 36$ is obtained from a population with $\mu = 64$ and $\sigma = 18$.
- Describe the sampling distribution of \bar{x} .
 - What is $P(\bar{x} < 62.6)$?
 - Calculate $P(59.8 < \bar{x} < 65.9)$?
 - Find $P(\bar{x} \geq 68.7)$.



PREPARE FOR A LEADING ROLE.

English-taught MSc programmes in engineering: Aeronautical, Biomedical, Electronics, Mechanical, Communication systems and Transport systems. No tuition fees.

→ liu.se/master

li.u LINKÖPING UNIVERSITY

- 6.6 The length of human pregnancies is approximately normally distributed with $\mu = 266$ days and standard deviation $\sigma = 16$ days.
- What is the probability that a randomly selected pregnancy lasts less than 260 days?
 - Suppose a random sample of size 20 pregnancies is obtained. Describe the sampling distribution of the sample mean length of human pregnancies.
 - What is the probability that a random sample of 20 pregnancies has a mean gestation period of 260 days or less?
 - What is the probability that a random sample of 50 pregnancies has a mean gestation period of 260 days or less?
 - What would you conclude if a random sample of 50 pregnancies resulted in a mean gestation period of 260 days or less?
 - What is the probability that a random sample of 15 pregnancies will have a mean gestation period within 10 days of the mean?
- 6.7 According to **ATMDeposit.com**, the mean ATM withdrawals is \$60, with a standard deviation of \$35.
- Do you think that the variable “ATM withdrawal” is normally distributed? If not, describe the shape of the variable.
 - If a random sample of 50 ATM withdrawals is obtained, describe the sampling distribution of \bar{x} , the mean of the withdrawals amount.
 - Determine the probability of obtaining a sample mean withdrawal amount between \$70 and \$75.
- 6.8 From the set of numbers {3, 5, 7}, a random sample of size 2 will be selected with replacement.
- List all possible samples and evaluate the mean of each.
 - Determine the distribution of \bar{x} .
- 6.9 A certain type of robe is made with mean tensile strength of 80 pounds and a standard deviation of 6 pounds. How is the variance of the sample mean changed when the sample size is
- Increased from 64 to 200?
 - Decreased from 750 to 49?
- 6.10 A random variable X, representing the number of chocolate chips in a piece of cake, has the following probability distribution.
- | | | | | |
|------|-----|-----|-----|-----|
| X | 4 | 5 | 6 | 7 |
| P(x) | 0.2 | 0.4 | 0.3 | 0.1 |
- Find the mean and the variance of X.

- b) Find the mean and the variance of the sample mean for random samples of size 36 chocolate cakes.
- c) Find the probability that the average number of chocolate chips in 36 chocolate cakes will be less than 5.5.
- 6.11 A random sample of size 25 is taken from a normal population with mean of 80 and variance 25. A second random sample of size 36 is taken from a different normal population have a mean of 75 and variance 9. Find the probability that the sample mean of the first sample will exceed that sample mean of the second sample by at least 3.4 but less than 5.9.
- 6.12 A machine used to fill plastic bottles with a soft drink has a known standard deviation of $\sigma = 0.05$ liters. The target mean fill volume is $\mu = 2.0$ liters.
- a) Describe the sampling distribution of the sample mean fill volume for a random sample of 45 such bottles.
- b) A quality-control manager obtains a random sample of 45 bottles. He will shut down the machine if the sample mean volume of these bottles is less than 1.98 liters or greater than 2.02 liters. What is the probability that the quality-control manager will shut down the machine even though the machine is correctly calibrated?
- 6.13 Explain the difference between p and \hat{p} , and give examples.
- 6.14 A random sample of 100 employees of a large company included 40 who had worked for the company for more than two years. For this sample $\hat{p} = 0.4$. If a different sample of 100 employees were selected, would you expect that \hat{p} be 0.40? Explain why or why not.
- 6.15 For which of the following combinations of sample size and population proportion would the standard deviation of \hat{p} be the smallest:
- a) $n = 40$ $p = 0.3$,
- b) $n = 60$ $p = 0.4$,
- c) $n = 100$ $p = 0.5$.
- 6.16 Give the conditions under which the shape of the sampling distribution of \hat{p} is approximately normal.
- 6.17 Describe the sampling distribution of \hat{p} . Let $N = 25,000$, in each case below.
- a) $n = 500$, and $p = 0.4$.

- b) $n = 300$, and $p = 0.7$.
c) $n = 1000$, and $p = 0.102$.
d) $n = 1100$, and $p = 0.85$
- 6.18 A simple random sample of size $n = 75$ is obtained from a population with $N = 10,000$, and $p = 0.8$.
- Describe the sampling distribution of \hat{p} .
 - What is the probability of obtaining $x = 63$ or more individuals with the characteristic? That is, find $P(\hat{p} \geq 0.84)$?
 - What is the probability of obtaining $x = 51$ or less individuals with the characteristic? That is, find $P(\hat{p} \leq 0.68)$?
- 6.19 A simple random sample of size $n = 300$ is obtained from a population with $N = 25,000$, and $p = 0.65$.
- Describe the sampling distribution of \hat{p} .
 - What is the probability of obtaining $x \geq 136$ individuals with the characteristic? That is, find $P(\hat{p} \geq 0.68)$?
 - What is the probability of obtaining $x \leq 118$ individuals with the characteristic? That is, find $P(\hat{p} \leq 0.59)$?

Click here
to learn more


TAKE THE
RIGHT TRACK

Give your career a head start
by studying with us. Experience the advantages
of our collaboration with major companies like
ABB, Volvo and Ericsson!

Apply by
15 January

World class
research

www.mdh.se


MÄLARDALEN UNIVERSITY
SWEDEN

- 6.20 A simple random sample of size $n = 1,000$ is obtained from a population with $N = 1,000,000$, and $p = 0.35$.
- Describe the sampling distribution of \hat{p} .
 - What is the probability of obtaining $x \geq 390$ individuals with the characteristic?
 - What is the probability of obtaining $x \leq 320$ individuals with the characteristic?
- 6.21 According to the National Center for Health Statistics (2004), 22.4% of adults are smokers. A random sample of 300 adults is obtained.
- Describe the sampling distribution of \hat{p} .
 - In a sample of $n = 300$, what is the probability of $x \geq 50$ individuals are smokers?
 - Would it be unusual if in a random sample of 300 results in 18% or less being smokers?
- 6.22 According to a USA *Today* "Snapshots" 26% of adults do not have any credit card. A simple random sample of 500 adults is obtained.
- Describe the sampling distribution of \hat{p} , the sample proportion of adults who do not have a credit card.
 - In a random sample of 500 adults, what is the probability that less than 24% have no credit card?
 - Would it be unusual if a random sample of 500 adults' results in 150 having no credit card?
- 6.23 According to a study conducted by a social organization, the proportion of Americans who are satisfied with the way things are going in their lives is 0.82. Suppose that a random sample of 100 Americans is obtained.
- Describe the sampling distribution of \hat{p} , the sample proportion of Americans who are satisfied with the way things are going.
 - What is the probability that at least 85 Americans in the sample are satisfied with their lives?
 - What is the probability that 75 or fewer Americans in the sample are satisfied with their lives? Is this result unusual?
- 6.24 If samples of size 9 are selected randomly for a normal population $N(14, 4)$. Find the values of a and b such that:
- $P(a < S^2 < b) = 0.90$,
 - $P(a < S^2 < b) = 0.95$,
 - $P(a < S^2 < b) = 0.975$,
 - $P(a < S^2 < b) = 0.995$.

(Hint: Consult, **Kinney, J.J.** (2002) *Statistics for Science and Engineering*, Addison, Wesley.)

- 6.25 If X has a normal distribution with mean 5 and variance 10, find $P(0.04 < (X - 5)^2 < 38.4)$.
- 6.26 If X is $N(0, 4)$, compute:
- $P(1 < X^2 < 9)$,
 - $P(10 < X^2 < 90)$.
- 6.27 If X is $N(1, 4)$, compute:
- $P(1 < X^2 < 9)$,
 - $P(2 < X^2 < 4)$.

TECHNOLOGY – STEP-BY-STEP

The reader should consult the section of **TECHNOLOGY – STEP-BY-STEP** that relates to the sampling distribution under consideration, as shown in **Chapter 5**.

7 SIMPLE LINEAR REGRESSION AND CORRELATION

7.1 INTRODUCTION

It is quite often that the investigator likes to check on how variables are related, and to “find if he can” a relationship between (or among) the variables on hand. An interesting question might be raised by a researcher such as: am I able to predict the value of a random variable if I have the value of one or more variables available? This kind of study is what we call **Regression Analysis**. As we know variables are two types, either independent or dependent, based on how the value of that variable is attained or obtained? For example, the relationship between heights and weights of individuals, temperature and pressure of any gas, the weight of a fish and its breathing capacity, the annual food expenditure and the annual income of the family, and so on, might be of interest to the scientist. In the above examples, there were two quantities, and we could label them as: input and output, independent and dependent, or explanatory and response, predictor and predicted. For



How will people travel in the future, and how will goods be transported? What resources will we use, and how many will we need? The passenger and freight traffic sector is developing rapidly, and we provide the impetus for innovation and movement. We develop components and systems for internal combustion engines that operate more cleanly and more efficiently than ever before. We are also pushing forward technologies that are bringing hybrid vehicles and alternative drives into a new dimension – for private, corporate, and public use. The challenges are great. We deliver the solutions and offer challenging jobs.

www.schaeffler.com/careers

SCHAEFFLER

example, the family income is the independent or explanatory variable while the food expenditure will be called the dependent or response variable.

The **Response Variable** is that variable whose value can be explained by the value of one or more **Explanatory** or **Predictor Variables**. There are many types of regression, based on the number of variables involved. When we have one response and one predictor, this is the case for **Simple Linear Regression**. In case we have one response and many predictor variables, this is the case of multiple regression models. In this chapter, we are interested in the case of two variables, and when the relationship between them is assumed to be linear. In addition to a linear simple regression between two variables there might be a logarithmic, an exponential, a quadratic, cubic and so on regressions, on just the two variables. In all the types of regression that were mentioned above, the relationship relating the response variable to the explanatory variables(s) will be a linear as it will be shown later.

The purpose of regression analysis is to determine the existence, the kind, or the extent of a relationship (in the form of a mathematical equation) between two or more variables. For example, if X and Y denote two variables under study, then, in general, we want to determine the best equation (relation), $y = f(x)$, that describes the relationship between x and y . Let Y be a random variable whose distribution depends on another nonrandom variable x . The relationship between these two variables is set as the mean of Y given x , $\mu_{y|x}$, varies linearly with x . Thus, the relationship would be like the following equation:

$$\mu_{y|x} = \beta_0 + \beta_1 x .$$

Thus, for a fixed value of x , the explanatory variable, Y as the response variable, will vary randomly around its mean: $\mu_{y|x}$. The term “**Regression**” goes back to **Sir Francis Galton**. In his work on hereditary, Galton noted that the sons of very tall fathers tended to be tall, but not quite as tall as their fathers. Also, sons of very short fathers tended to be short, but not quite as short as their fathers. He called this tendency of individuals to be not quite tall or as short as their fathers, the tendency of “regression toward mediocrity”, i.e., going back to the average.

Regression: Is a functional relationship between two or more correlated variables that is often empirically determined from data and is used specially to predict values of one variable when given values of the others (the \sim of y on x is linear). More specifically: **Regression** is a function that yields the mean value of a random variable under the conditions that one or more independent variables have specified values. (Webster’s New Collegiate Dictionary, Merriam, 1981, p. 966.)

Linear: an equation of the first degree in any number of variables is linear, as far as the coefficients in the equation are linear. Those coefficients are called the parameters of the

linear model. The model is linear despite the fact that it will have any powers of the explanatory variables. This will become completely clear when we address the **Multiple Linear Regression in Chapter 5 of Inferential Statistics – The Basics for Biostatistics**.

Simple: in this chapter, we will be dealing ONLY with one explanatory variable and one response variable, and the coefficients in the linear relationship will be only taken to the first power, or first degree.

In many of the experiments and investigative work, it is desired to check, and possibly see, how the changes in one variable affect another variable, or introduce some changes in that variable. One can distinguish between exact and non-exact relationships.

7.1.1 EXACT RELATIONSHIP

Sometimes we find that two variables are linked by an exact relationship. For example, if the resistance “R” of a simple circuit is kept constant, the current intensity, “I”, is related to the voltage “V” by Ohm’s Law: $I = V/R$, which is an equation of first degree and the graph of it is a straight line. This is an exact law. However, if we want to verify it empirically by making changes in V and observing I, while R is kept constant, we notice that the plot of the empirical points (V, I) will not fall exactly on a straight line. The reason behind this is that the measurements may be subject to slight errors and thus the plotted points would not fall exactly on the straight line but would vary randomly around it. For purposes of predicting the value of I for a particular value of V, (with R as a fixed constant) we should use the plotted straight line.

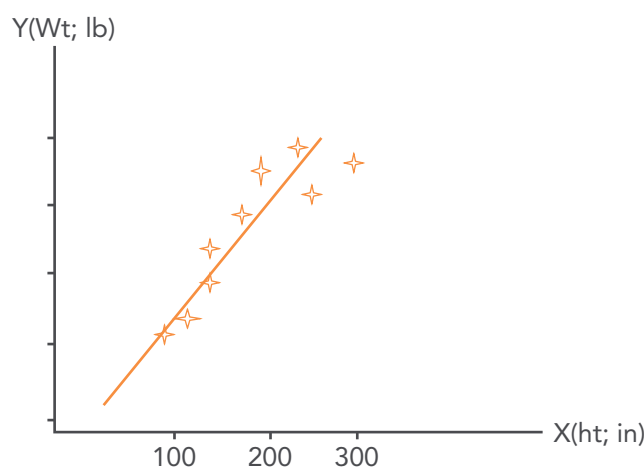


Figure 1 Wt versus Ht

7.1.2 NON-EXACT RELATIONSHIP

Sometimes the relationship is not exact even apart from the errors in measurements. For example, suppose we consider the height and weight of adult males of some population. If we plot the ordered pair (weight, height) = (x, y), a diagram, like **Figure 1**, will result. Such a representation is conventionally called a scatter diagram.

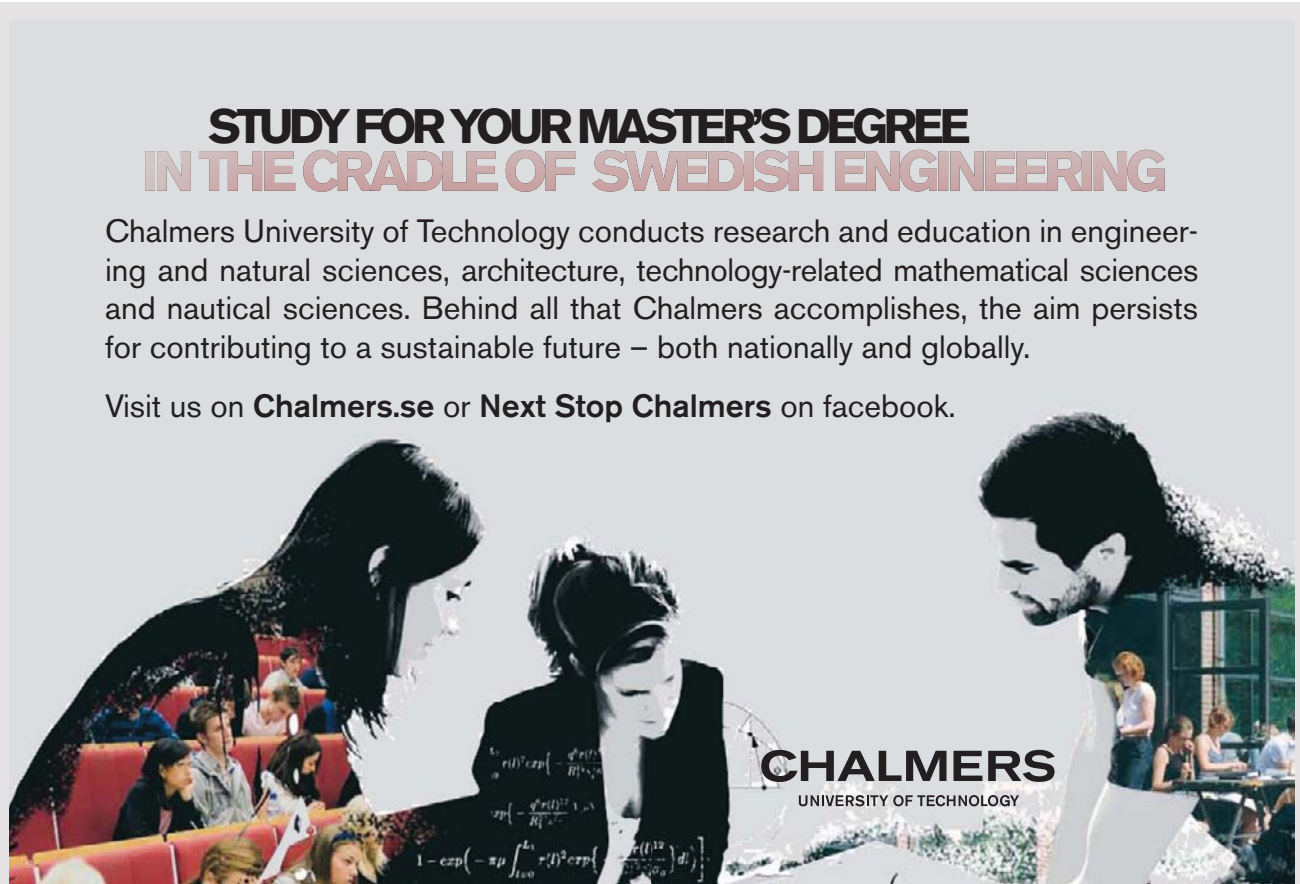
Note that for any given height there is a range of observed weights and vice versa. This variation will be partially due to measurement errors but primarily due to variation among individuals. If you consider two males with same height, their weights will be different not only because we cannot measure the weights exactly correct, but also because the two males will have slight different weights. Thus, no unique relationship between the actual weight and height can be written. But we can notice that the **Average Observed Weight**, for a given observed height, increases as height increases. Whether a relationship is exact or non-exact, in so far as average values are concerned, it will be useful especially for prediction purposes.

It is not always clear which variable should be considered the response variable and which the explanatory variable. For example, does high school GPA predict a student's SAT score, or can the SAT score be used to predict the student's GPA in college? The investigator

STUDY FOR YOUR MASTER'S DEGREE
IN THE CRADLE OF SWEDISH ENGINEERING

Chalmers University of Technology conducts research and education in engineering and natural sciences, architecture, technology-related mathematical sciences and nautical sciences. Behind all that Chalmers accomplishes, the aim persists for contributing to a sustainable future – both nationally and globally.

Visit us on **Chalmers.se** or **Next Stop Chalmers** on facebook.



CHALMERS
UNIVERSITY OF TECHNOLOGY

should determine which variable plays the role of the explanatory variable based on the question that needs to be answered.

Scatter Diagrams show the type of the relation that exists between two variables. By plotting the explanatory variable along the horizontal axis and the response variable along the vertical axis, our goal is to distinguish which scatter diagrams will imply a linear relation from those that will imply a nonlinear relation and those that imply no relation. **Figure 2** and **Figure 3** show some scatter diagrams and the type of the relationship implied.

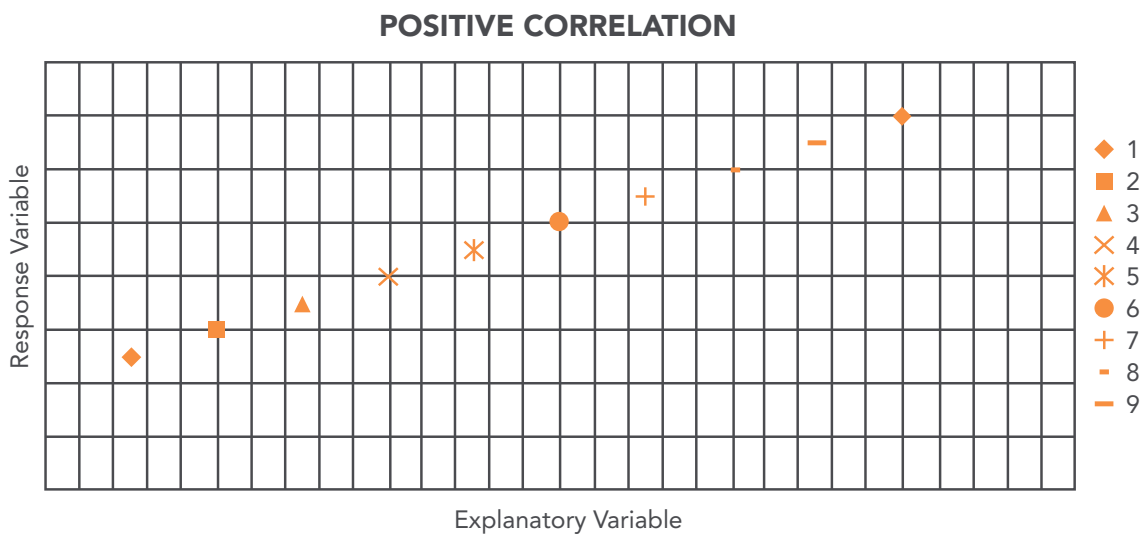


Figure 2 Perfect Positive Correlations

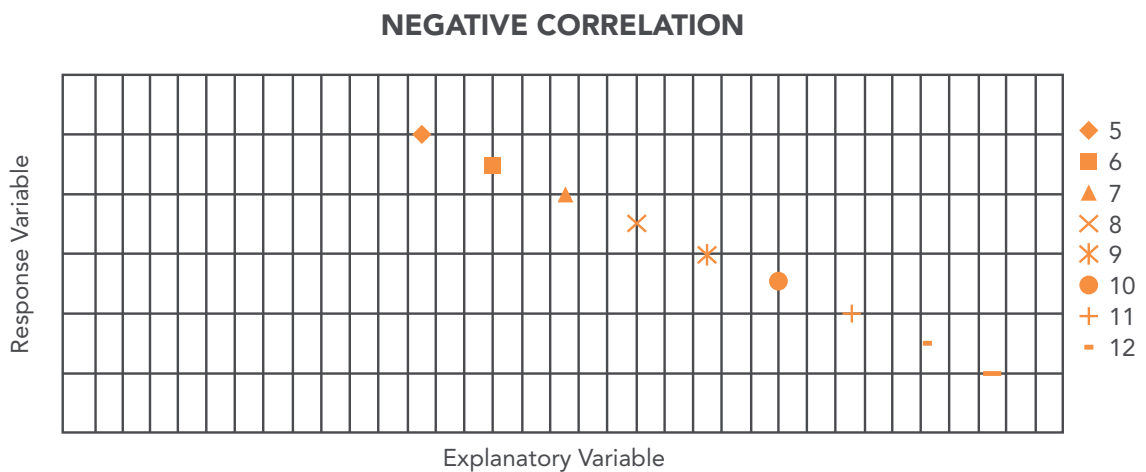


Figure 3 Perfect Negative Correlation.

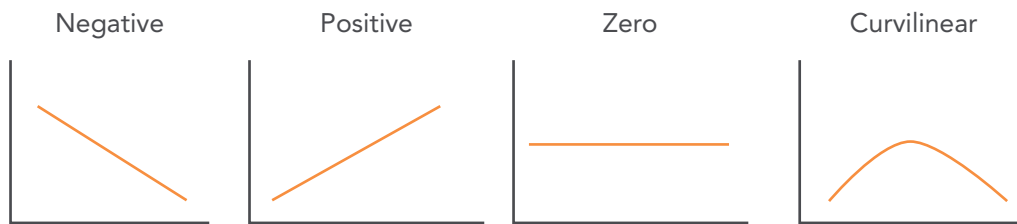


Figure 4 Types of Relationship between Two Variables

Figure 4, above, shows the types of linear and nonlinear relationships between two variables.

In **Figure 4A**, we see a perfect negative linear relationship, and we say that the two variables are negatively associated. In other words, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

On the other hand, **Figure 4B** shows that there are two variables that are linearly related and positively associated. In the same vein, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable increases.

The situation is quite different in **Figure 4C**. There is the case where we see a horizontal line, although it is linear, but the response variable, Y , along the y -axis, is not affected by the change in the explanatory variable, x , along the x -axis. Thus $Y = a$ constant.

Figure 4D shows a nonlinear relationship between the two variables x and Y .

Figure 4 displays ideal cases for positively and negatively associated variables based on a linear relationship between the two. In section 9.4 we will check on the strength of that linear relationship between the explanatory variable x and the response variable Y .

7.2 REGRESSION MODELS

The simplest linear regression model is of the form

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where the response variable, Y , depends on one regressor, or explanatory variable, x , and the equation is a polynomial of the first degree in x . This model is called the simple first degree model. β_0 & β_1 as unknown constants, are called the parameters of the model. Specifically, β_0 is the y -intercept and β_1 is called the regression coefficient, or the slope of the line. The symbol ε denotes a random error (the model error), which will be assumed

to have a normal distribution with mean 0 and variance σ^2 . The first order model (or the straight-line relationship between the response and the explanatory variables) can be valuable in many situations. The relationship may actually be a straight line as in the example on Ohm's law. Even if the relationship is not actually of the first order (or linear), it may be approximated by a straight line at least over some range of the input data. In **Figure 5** below, the relationship between x and Y is obviously nonlinear over the range $0 < x < 100$. However, if we were interested primarily in the range $0 < x < 25$, a straight-line relationship evaluated, on observations in this range, might provide a perfectly adequate representation of the function. Hence the relationship that just got fitted would not apply to values of x beyond the restricted range, and it could not be used for predictive purposes outside that range.



Scholarships

Lnu.se

Open your mind to new opportunities

With 31,000 students, Linnaeus University is one of the larger universities in Sweden. We are a modern university, known for our strong international profile. Every year more than 1,600 international students from all over the world choose to enjoy the friendly atmosphere and active student life at Linnaeus University. Welcome to join us!

Linnæus University
Sweden

Bachelor programmes in
Business & Economics | Computer Science/IT | Design | Mathematics

Master programmes in
Business & Economics | Behavioural Sciences | Computer Science/IT | Cultural Studies & Social Sciences | Design | Mathematics | Natural Sciences | Technology & Engineering

Summer Academy courses



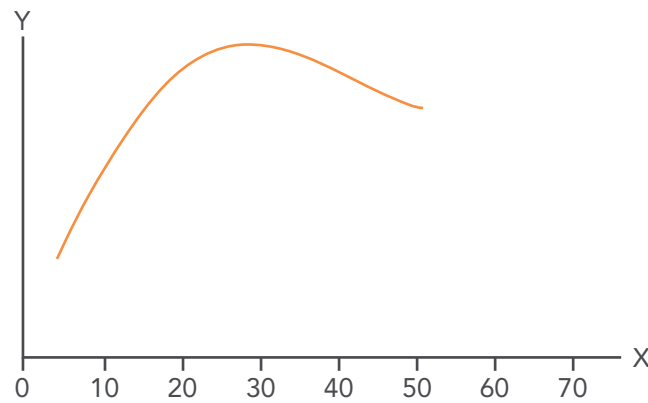


Figure 5 Linear on a part but not on all range of X

7.3 FITTING A STRAIGHT LINE (FIRST ORDER MODEL)

The first order model given by

$$Y = \beta_0 + \beta_1 x + \mathcal{E},$$

have the unknown parameters β_0 and β_1 , that need to be estimated in order to use the relationship for prediction purposes? Thus a sample of size n of ordered observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, will be used to estimate those two parameters. The estimators of those two parameters will be denoted by b_0 and b_1 respectively, and hence we will have the following equation

$$\hat{y} = b_0 + b_1 x,$$

which is to be called the prediction equation, with the understanding that $b_0 = \beta_0$, and $b_1 = \beta_1$.

The symbol of \hat{y} is used in the least squares regression line to serve as the predicted value of Y for a given value of x . It is worth noting that the least squares line contains the point (\bar{x}, \bar{y}) , where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Each observation (x_i, y_i) , $i = 1, 2, \dots, n$, in the sample, satisfies the equation

$$Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i,$$

where \mathcal{E}_i is the value assumed by E_i when Y_i takes on the value y_i . \mathcal{E}_i , $i = 1, 2, \dots, n$, are the unknown error components due to the measurements on Y . These are unobservable random variables, which we assume that they are independently and distributed with mean zero and variance σ^2 ; in other words we have $\mathcal{E}_i \sim N(0, \sigma^2)$. The above equation can be viewed as the model for a single observation y_i . Similarly, using the estimated, or fitted regression line

$$\hat{y} = b_0 + b_1x.$$

Each pair of observations: (x_i, y_i) , $i = 1, 2, \dots, n$, satisfies the relation

$$Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i,$$

where $E_i = y_i - \hat{y}_i$ is called the **Residual**. It describes the error in the fit of the model at the i^{th} data point or observation.

We shall find b_0 and b_1 , the estimates of β_0 and β_1 , so that the sum of the squares of the residuals is a minimum. The residual sum of squares is often called the sum of squares of the errors about the regression line and it is denoted by SSE. This minimization procedure for estimating the parameters is called the **Method of Least-Squares**. Thus we shall find b_0 and b_1 so as to minimize

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^n (\text{Observed response} - \text{Predicted response})^2, \\ \text{SSE} &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2. \end{aligned}$$

(The interested reader, who is aware of differential calculus, can check the procedure in R.E. Walpole, and R.H. Myers, 1989.) Based on that, let us have some basic notation for later reference, since those will be used quite often, as displayed below:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n X_i, & \bar{y} &= \frac{1}{n} \sum_{i=1}^n Y_i, \\ S_{xx} &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2, \\ S_{yy} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2, \text{ and} \\ S_{xy} &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{1}{n} \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right). \end{aligned}$$

The least-squares regression line is the line that minimizes the sum of the square of the vertical distances between the observed values of Y and those predicted by the line, \hat{y} . Thus

solving the normal equations resulting from the differentiation of the SSE, we see that the estimates of the linear coefficients are given by

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}.$$

EXAMPLE 7.1: The following measurements of the specific heat of a certain chemical were made in order to investigate the variation in specific heat with temperature:

Temp °C	0	10	20	30	40
Specific Heat	0.51	0.55	0.57	0.59	0.63

Estimate the regression line of specific heat on temperature, and predict the value of the specific heat when the temperature is 25°C.

Solution: From the data we see that: $n = 5$, $\sum_{i=1}^n x_i = 100$, $\bar{x} = 20.0$, $\sum_{i=1}^n y_i = 2.85$, $\bar{y} = 0.57$, and $\sum_{i=1}^n x_i y_i = 59.8$. Thus, by applying the above formulas for b_0 and b_1 , we have $b_0 = 0.514$, and $b_1 = 0.0028$. Hence the fitted equation will be given by

e-learning for kids

- The number 1 MOOC for Primary Education
- Free Digital Learning for Children 5-12
- 15 Million Children Reached

About e-Learning for Kids Established in 2004, e-Learning for Kids is a global nonprofit foundation dedicated to fun and free learning on the Internet for children ages 5 - 12 with courses in math, science, language arts, computers, health and environmental skills. Since 2005, more than 15 million children in over 190 countries have benefitted from eLessons provided by EFKI. An all-volunteer staff consists of education and e-learning experts and business professionals from around the world committed to making difference. eLearning for Kids is actively seeking funding, volunteers, sponsors and courseware developers; get involved! For more information, please visit www.e-learningforkids.org.

$$\hat{y} = 0.514 + 0.0028x.$$

When the temperature $x = 25^\circ\text{C}$, the predicted specific heat is $0.514 + (0.0028) \cdot 25 = 0.584$.

7.3.1 METHOD OF CODING

When the values of the controlled (independent, explanatory) variable are equally spaced, the calculations of the regression line become considerably simplified by coding the data in integers symmetrically situated about zero to make $\sum_{i=1}^n x_i = 0$, and thus the estimators of the coefficients, in the regression line, become

$$b_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \quad \text{and} \quad b_0 = \bar{y}.$$

If there is an odd number of evenly spaced data values for x , denote them by: ..., -2, -1, 0, 1, 2, ...

If there is an even number of evenly spaced data values for x , denote them by: ..., -3, -1, 1, 3,

EXAMPLE 7.2: Consider the following data that stands for the output of a certain company for the years 1960–1964. Predict the output of the company in the year 1965.

Output (1000 tons)	11.1	12.3	13.7	14.6	15.6
Year	1960	1961	1962	1963	1964

Solution: Based on the above setup we can see that the data can be coded, as shown in the Table below. Hence the estimated coefficients in the predicted equation will have the following values: $b_0 = 13.46$ and $b_1 = 1.13$; and the fitted equation is given by $\hat{y} = 13.46 + 1.13x$, where x is the coded year.

To predict the output for the year 1965, we take $x = 3$ (which is its coded value) in the predicting equation, to obtain $\hat{y} = 13.46 + (1.13) \cdot 3 = 16.85$, or 16850 tons.

Variable	Data					Totals
Output y	11.1	12.3	13.7	14.6	15.6	67.3
Year x	-2	-1	0	1	2	0

X^2	4	1	0	1	4	10
XY	-22.2	-12.3	0	14.6	31.2	11.3

Table 1

7.3.2 NON-LINEAR MODEL

The importance of the linear relationship (or the first order model) goes beyond the cases where the linearity is apparent. Some relationships are not linear to start with, but they can be made linear easily. For example, consider the experimental growth model (or Decay, when the rate is negative):

$$P = Ae^{rt} \cdot u,$$

where P is the population size at time t , A is the population size at time zero (a constant), “ r ” is a constant representing the rate (of growth when $r > 0$ and decay when $r < 0$) and u represents a random error. This model describes situations when the relative growth is constant in the population. It can describe the growth in human or biological populations, and the growth in compound interest, when $u = 1$, and A is the principal amount.

EXAMPLE 7.3: Consider the following data for the population of the USA (in millions) during the years 1860–1900 (based on the USA Census every 10 years):

Year t	1860	1870	1880	1890	1900
Population	31.4	39.8	50.2	63.0	76.0

Solution: The exponential growth model, displayed above, with $u = 1$, can be transformed very easily to a linear relation by taking the logarithm to base e (i.e. \ln) of both sides, to obtain

$$\ln P = \ln A + rt + \ln u,$$

Which can be written as: $y = \beta_0 + \beta_1 x + E$, where $y = \ln P$, $\beta_1 = r$, $x = t$ and $E = \ln u$. Now we have a linear model, we can fit using the data given above. Moreover, by appealing to **EXAMPLE 10.2**, and coding the data, we have the following generated data in Table 2 for manual calculations. Thus we have

$$b_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = 0.222, \text{ and } b_0 = \bar{y} = 3.902.$$

Variable	Data					Totals
X	-2	-1	0	1	2	0
Y= Ln P	3.45	3.68	3.91	4.14	4.33	19.51
XY	-6.9	-3.68	0	4.14	8.66	2.22
X ²	4	1	0	1	4	10

Table 2

Hence the linear model takes the form

$$\hat{y} = 3.902 + 0.222x.$$

By encoding we reach at $\bar{P} = 49.5e^{0.222t}$.

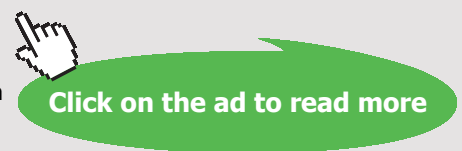


.....Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



7.4 CORRELATION

As it was mentioned above, in this section we will be investigating how strong the linear relationship between the explanatory and response variables.

It is dangerous to use only a scatter diagram to check on the relation between the two variables, see **Figure 1**. For the sake of argument, what if someone else used a different scale for the same data and he depicted the scatter plot, will there be the same conclusion, as it was reached from **Figure 1** earlier? Based on that, we need another tool in order to be sure we reach the same conclusion without a scatter plot dilemma. For that purpose we need to define **The Linear Correlation Coefficient** that will show how strong is that linear relationship between the two variables on hand.

The Linear Correlation Coefficient or **Pearson Product Moment Correlation Coefficient** is a measure for the strength of a linear relation between two quantitative variables. The Greek letter ρ (rho) is used to represent the population correlation coefficient and r will represent the sample correlation coefficient. Recall that ρ is a parameter, and thus can be estimated by a statistic. What is better than r for ρ ?

We will give a simple formula for calculating r using the data on hand. To do that, let us recall how the Z -scores (whether for a population or for a sample) were defined, or calculated. Recall the formula for the Z -scores for a sample data, in words, with S being the standard deviation of the data set, we have

$$\text{Z-score} = \frac{\text{data point} - \text{mean of the data set}}{S}.$$

Since we are dealing with two samples from the explanatory variable x and the response variable Y , then based on the above definition of the Z -scores, we have the following corresponding Z -Scores:

$$Z_x = \frac{x_i - \bar{x}}{S_x}, \text{ and } Z_y = \frac{y_i - \bar{y}}{S_y}.$$

Based on the above Z 's expressions, we can present the sample correlation coefficient, r , as given by

$$r = \frac{\sum_i Z_x Z_y}{n-1}.$$

Without any doubt, the above formula for r is cumbersome and highly error-prone due to the rounding done in the middle steps. An equivalent formula for the linear correlation coefficient of the sample, r , is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}.$$

The above formula is much more accurate for the manual calculations of r than the one given in terms of the Z - scores. In addition to that, any software will give a value for r faster and more accurate to as many decimal places as the investigator likes. **The Pearson Linear Correlation Coefficient** is named in honor of **Karl Pearson** (1857–1936). The following are the properties of the **Linear Correlation Coefficient**.

1. The Linear Correlation coefficient is always between -1 and 1, inclusive i.e. $-1 \leq r \leq 1$.
2. If $r = 1$, then there is a perfect positive linear relation between the two variables. See **Figure 4B**.
3. If $r = -1$, then there is a perfect negative linear relationship between the two variables. See **Figure 4A**.
4. The closer r is to 1, the stronger is the evidence of a positive association between the two variables.
5. The closer r is to -1, the stronger is the evidence of a negative association between the two variables.
6. The closer r is to 0, there is a little or no evidence of a linear relation between the two variables. Never the less the closer r to 0 does not mean no relation, **just no linear relation**, See **Figure 4D**.
7. The linear correlation coefficient is unit less, as it appeared from its definition in terms of the Z - scores, where they are unit less, i.e. there is no unit of measurement attached to it.

To illustrate the notions mentioned above, let us give an example. We will take small values for both x and y just to see how the calculations can be done.

EXAMPLE 7.4: Consider the paired data: (x, y) : (2, 1.4), (4, 1.8), (8, 2.1), (8, 2.3), (9, 2.6). Calculate r .

Solution: Aside from using Technology to find r , faster, more accurate, and less time consuming, let us set the stage for manual calculations by making Table 3.

Variable	Data					Totals
X	2	4	8	8	9	31
Y	1.4	1.8	2.1	2.3	2.6	10.2
XY	2.8	7.2	16.8	18.4	23.4	68.6
X ²	4	16	64	64	81	229
Y ²	1.96	3.24	4.41	5.29	6.76	21.66

Table 3

It is quite clear from Table 3 that all the terms which are needed for the formula, to calculate r , are given. Thus, plugging in those numerical values, we found $r = 0.9572$.

REMARK: It is to be noticed that the expressions giving the coefficient b_1 , and the correlation coefficient r have the same numerator. Moreover, it is needless to say that the denominators in both expressions are positive. Thus, the investigator should be very careful when doing such calculations, that they should have the same sign, i.e. either both are positive or both are negative.

Nido

Luxurious accommodation

Central zone 1 & 2 locations

Meet hundreds of international students

BOOK NOW and get a £100 voucher from voucherexpress

Nido Student Living - London

Visit www.NidoStudentLiving.com/Bookboon for more info.

+44 (0)20 3102 1060



CHAPTER 7 EXERCISES

- 7.1 What does it mean to say that two variables are positively associated?
- 7.2 What does it mean to say that the linear correlation coefficient between two variables equals 1? What would the scatter diagram look like?
- 7.3 What does it mean if $r = 0$

7.4 For the data set

x	0	2	3	5	6	6
y	5.8	5.7	5.2	2.8	1.9	2.2

- a) Draw a scatter diagram.
 - b) Comment on the type of relation that appears to exist between x and y.
 - c) Determine the least squares regression line.
 - d) Graph the least squares regression line on the scatter diagram drawn in part a), and compare your comments to those made in b).
- 7.5 A study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

- a) Assuming the yield is given by $Y = \beta_0 + \beta_1 x + E$, what are the least squares estimates of β_0 and β_1 ? What is the prediction equation?
 - b) Construct the analysis of variance table and test the hypothesis: $H_0: \beta_0 = 0$, with $\alpha = 0.05$.
 - c) What are the confidence limits for β_1 , with $\alpha = 0.05$.
 - d) What are the confidence limits (with $\alpha = 0.05$) for the true mean value of y when $x = 3$?
- 7.6 Twelve specimens of Cu-Ni alloys, each with specific iron content, were tested in corrosion wheel setup. The wheel was rotated in salt sea water at 30 ft/sec for 60 days. The corrosion was measured in weight loss in mg/square decimeter/day, MDD. The following data were collected: (Take X for Fe and Y for MDD)

x	0.01	0.71	0.95	1.19	1.01	0.48	1.44	0.71	1.96	0.01	1.44	1.9
y	127.6	110.8	103.9	101.5	130.1	122.0	92.3	113.1	83.7	128.0	91.4	86.2

Determine if the effect of iron content, on the corrosion resistance of Cu-Ni alloys in sea water, can be justifiably represented by a straight-line model. Assume $\alpha = 0.05$.

- 7.7 The effect of temperature of the deodorizing process on the color of the finished product was determined exponentially. The data collected were as follows, where X stands for Temperature and Y for color:

x	460	450	440	430	420	410	450	440	430	420	410
	400	420	410	400							
y	0.3	0.3	0.4	0.4	0.6	0.5	0.5	0.6	0.6	0.6	0.7
	0.6	0.6	0.6	0.6							

- Fit the model $Y = \beta_0 + \beta_1x + E$.
- Test for significance regression, with $\alpha = 0.05$.
- Obtain a 95% confidence interval for β_0 and β_1 .

- 7.8 The normal stress on a specimen is known to be functionally related to the shear resistance. The following is a set of coded experimental data on the two variables: X for Normal Stress and Y for Shear resistance;

x	26.8	25.4	28.9	23.6	27.7	23.9	24.7	28.1	26.9	27.4	22.6	25.6
y	26.5	27.3	24.2	27.1	23.6	25.9	26.3	22.5	21.7	21.4	25.8	24.9

- Estimate the regression line $\mu_{y|x} = \beta_0 + \beta_1x$.
- Estimate the shear resistance for a normal stress of 24.5 pounds per square inch.

- 7.9 a) Compute and interpret the correlation coefficient for the following grades of six students selected at random:

Math Grade:	70	92	80	74	65	83
English Grade:	74	84	63	87	78	90

- Test for significant correlation at 5% level.

7.10 Compute and interpret the correlation coefficient for the following data selected at random:

x	4	5	9	14	718	22	24
y	16	22	11	16	7	3	17


7.11 The pressure P of a gas corresponding to various Volumes V was recorded as follows:

V (cm ³)	50	60	70	90	100
P (kg/cm ²)	64.7	51.3	40.5	25.9	7.8


The ideal gas law is given by the equation $PV^\gamma = C$, where γ and C are constants.

- Following the suggested procedure in Example 9.3, find the least square estimates of γ and C .
- Estimate P when $V = 80$ cubic centimeters.

SIMPLY CLEVER




WE WILL TURN YOUR CV INTO AN OPPORTUNITY OF A LIFETIME



Do you like cars? Would you like to be a part of a successful brand?
As a constructor at ŠKODA AUTO you will put great things in motion. Things that will ease everyday lives of people all around Send us your CV. We will give it an entirely new new dimension.

Send us your CV on
www.employerforlife.com



7.12 Consider the following data on the heights of fathers' and sons'

x Father's ht.	70.3	67.2	70.9	66.9	72.9	70.3	71.7	71.0	69.9	70.8
x Father's ht.	70.1	70.4	72.4							
y Son's ht.	74.2	69.3	66.9	69.2	67.8	70.1	70.4	69.3	75.8	72.2
y Son's ht.	69.5	68.7	73.8							

a) Find the least square regression line of the son's ht based on the father's ht. Predict the son's ht, when the father is 77 in. tall.

In the following Exercises: 7.13–7.16, you are given the regression equation:

- Calculate the predicted values
- Calculate the residuals, and the sum of their squares
- Construct a scatter plot of the residuals versus the predicted values
- Construct a normal probability plot of the residuals using technology
- Verify that the regression assumptions are valid

7.13 $\hat{y} = 2.5x + 13.5$

x	1	2	3	4	5
y	15	20	20	25	25

7.14 $\hat{y} = 3.2x + 8$

x	-5	-4	-3	-2	-1
y	0	8	8	16	1

7.15 $\hat{y} = -2x + 8$

x	1	2	2	2	3
y	6	5	4	3	2

7.16 $\hat{y} = -0.5x + 104$

x	10	20	30	40	50
y	100	95	85	85	80

TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP Drawing Scatter Diagrams and Determining the Correlation Coefficient

TI-83/84 Plus

Scatter Diagram

1. Enter the explanatory variable in L1 and the response variable in L2.
2. Press 2^{nd} Y = to access Stat-Plots menu. Select 1: plot 1.
3. Place the cursor on "ON" and press ENTER, to turn the plots on
4. Highlight the scatter diagram icon, and press ENTER. Make sure that Xlist is L1 and Ylist is L2.
5. Press ZOOM, AND SELECT 9: ZoomStat.

Correlation Coefficient

1. Turn the diagnostics on by selecting the catalog (2^{nd} 0). Scroll down and select Diagnostics On. Hit ENTER twice to activate diagnostics.
2. With the explanatory variable in L1 and the response variable in L2, press STAT, highlight CALC and select 4: LinReg (ax + b). With LinReg on the HOME screen, press ENTER.

Excel

Scatter Diagram

1. Enter the explanatory variable in column A, and the response variable in column B.
2. Highlight both sets of data and select the Chart Wizard icon.
3. Select XY (Scatter).
4. Click Finish.

Correlation Coefficient

1. Be sure the Data Analysis Tool Pak is activated by selecting TOOLS menu and highlight Add-Ins.... Check the box for the Analysis Tool Pak and select OK.
2. Select TOOLS and highlight Data Analysis.... Highlight Correlation and select OK.
3. With the cursor in the Input Range, highlight the data. Select OK.

TECHNOLOGY STEP-BY-STEP Determining the Least-Squares Regression Line

TI-83/84 Plus

Use the same steps that were followed to obtain the correlation coefficient.

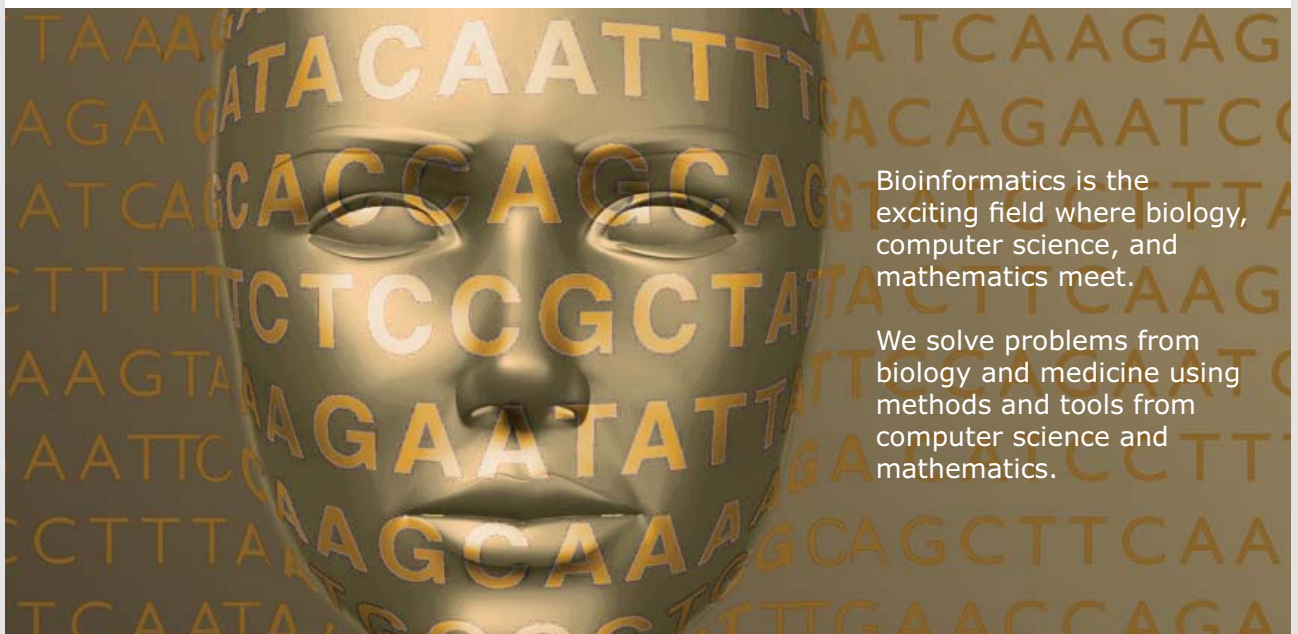
Excel

10. Be sure the Data Analysis Tool Pak is activated by selecting **TOOLS** menu and highlight Add-Ins.... Check the box for the Analysis Tool Pak and select OK.
11. Enter the explanatory variable in column A, and the response variable in column B.
12. Select **TOOLS** and highlight **Data Analysis...**
13. Select **Regression** option.
14. With the cursor in the Y-Range cell, highlight the column that contains the response variable. With the cursor in the X-Range cell, highlight the column that contains the explanatory variable. Select the output range. Press **OK**.



UPPSALA
UNIVERSITET

Develop the tools we need for Life Science Masters Degree in Bioinformatics



Bioinformatics is the exciting field where biology, computer science, and mathematics meet.

We solve problems from biology and medicine using methods and tools from computer science and mathematics.

Read more about this and our other international masters degree programmes at www.uu.se/master



TECHNOLOGY STEP-BY-STEP Determining R^2

TI-83/84 Plus

Use the same steps that were followed to obtain the correlation coefficient to obtain R^2 . Diagnostics must be on.

Excel

This is provided in the standard regression output.

TECHNOLOGY STEP-BY-STEP Testing the Least-Squares Regression Linear Model

TI-83/84 Plus

1. Enter the explanatory variable in L_1 and the response variable in L_2 .
2. Press **STAT**, highlight **TESTS**, and select E: **LinRegTTest**
3. Make sure that XList is L_1 , Ylist is L_2 and Freq is set to 1.
4. Select the direction of the alternative hypothesis.
5. Place the cursor on calculate and press **ENTER**

Excel

1. Be sure the **Data Analysis** Tool Pak is activated by selecting **TOOLS** menu and highlight **Add-Ins...** Check the box for the Analysis Tool Pak and select **OK**.
2. Enter the explanatory variable in column **A**, and the response variable in column **B**.
3. Select **TOOLS** and highlight **Data Analysis...**
4. Select **Regression** option.

With the cursor in the **Y-Range** cell, highlight the column that contains the response variable. With the cursor in the **X-Range** cell, highlight the column that contains the explanatory variable. Select the output range. Press **OK**.

APPENDIX A

TABLES

Row	Column									
	01-05	06-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
1	00467	93671	74438	38690	25956	84084	69732	40508	09980	93017
2	97141	74197	96225	95694	73772	47501	03811	66921	5243	57051
3	44690	04429	81692	48434	90603	80705	58951	38740	26288	46603
4	23980	21232	31803	02214	01698	80449	81601	78817	36040	47455
5	84592	59109	88679	46584	29328	84106	68158	08264	00648	64181
6	89392	93458	42116	26909	09914	26651	27896	09160	61548	00467
7	23212	55212	33306	68157	68773	99813	73213	31887	38779	79141
8	74483	25906	64807	20037	87423	40397	189984	08763	47050	44960
9	36590	66494	32533	83668	31847	02957	88499	54158	78242	23890
10	25956	96327	50727	11577	82126	65189	28894	00377	63432	02398
11	36544	17093	30181	00483	49666	66628	85262	31043	71117	84259
12	68518	51075	90605	14791	94555	14786	86547	28822	30588	40907
13	40805	30664	36525	90398	62426	15910	81324	06626	94683	17255
14	09980	55744	30153	26552	73934	79743	31457	98477	33802	18351
15	61458	18416	24661	95851	83846	89370	62869	89783	07617	00817
16	17639	15980	80100	17684	45868	47460	85581	36329	30604	17498
17	96252	20609	98370	65115	33468	19191	96635	01315	15987	23798
18	95649	45590	17638	82209	16093	26480	82182	02084	28945	16696
19	73727	89817	05403	46491	29775	33912	28906	48565	76149	80417
20	33912	58542	86186	18610	30357	36544	40603	84756	80357	50824

Table I Random Numbers

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	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3694	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	.3.50	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.4969	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681

UNIVERSITY OF COPENHAGEN



Copenhagen Master of Excellence

Copenhagen Master of Excellence are two-year master degrees taught in English at one of Europe's leading universities

Come to Copenhagen - *and aspire!*

Apply now at
www.come.ku.dk



cultural studies

religious studies

science

1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002

Table II Standard Normal Distribution: $P(Z > z)$

		$F(x) = P(X \leq x)$									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	0.9025	0.8100	0.7226	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
	2										1.0000
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.9928	0.9720	0.9392	0.8960	0.8438	0.7840	0.7182	0.6480	0.5748	0.5000
	2	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
	3										1.0000
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2410	0.1780	0.1296	0.0915	0.0625
	1	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3		0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
	4										1.0000
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3		0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4			0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
	5										1.0000
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.9672	0.8857	0.7765	0.6553	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438
	3	0.9990	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6562
	4		0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5				0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
	6										1.0000

n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4		0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5			0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6					0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
7											1

n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	0.9996	0.995	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.477	0.3633
	4		0.9996	0.9971	0.9896	0.9727	0.942	0.8939	0.8263	0.7396	0.6367
	5			0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6				0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7						0.9999	0.9998	0.9993	0.9983	0.9961
8											1

n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	#####	0.0195
	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4		0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5		0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6				0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7					0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8							0.9999	0.9997	0.9992	0.9980
9											1

n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5		0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6			0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7				0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8						0.9999	0.9995	0.9983	0.9955	0.9893
	9								0.9999	0.9997	0.9990
	10										1.0000

n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
11	0	0.5688	0.3183	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005
	1	0.8981	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0020	0.0139	0.0059
	2	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
	3	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
	4	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
	5		0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7536	0.6331	0.5000
	6			0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.9262	0.7256
	7				0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
	8					0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
	9							0.9998	0.9993	0.9978	0.9941
	10									0.9998	0.9995
	11										1.0000

n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	0	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
	2	0.9683	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
	6		0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
	7			0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000
	8			0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964
	9				0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
	10					0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
	11						0.9999	0.9995	0.9981	0.9937	0.9824
	12							0.9999	0.9987	0.9989	0.9963
	13								0.9999	0.9995	
	14										1.0000
	15										1

Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

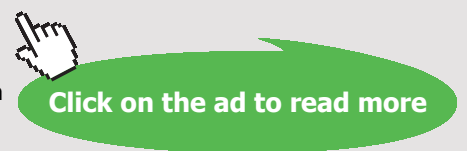
Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF



n	x	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
20	0	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
	1	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000
	2	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002
	3	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013
	4	0.9974	0.9568	0.8298	0.6296	0.4148	0.2357	0.1182	0.0510	0.0189	0.0059
	5	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207
	6		0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577
	7		0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316
	8		0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517
	9			0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119
	10				0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881
	11				0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483
	12					0.9998	0.9987	0.9940	0.9790	0.9420	0.8684
	13						0.9997	0.9985	0.9935	0.9786	0.9423
	14							0.9997	0.9984	0.9936	0.9793
	15								0.9997	0.9985	0.9941
	16									0.9997	0.9987
	17										0.9998
	18										1.0000
	19										1.0000
	20										1

TABLE III BINOMIAL DISTRIBUTION

T-Table Part A $df = 1-25$, Area is to the left
Probability less than the critical value $T_{1-\alpha, v}$

df	0.90	0.95	0.975	0.99	0.995	0.999
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782

8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
	1.282	1.645	1.960	2.326	2.576	3.090

T-Table Part B df = 26-50. Area is to the left
Probability less than the critical value $T_{1-\alpha, v}$

df	0.90	0.95	0.975	0.99	0.995	0.999
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385
31.	1.309	1.696	2.040	2.453	2.744	3.375
32.	1.309	1.694	2.037	2.449	2.738	3.365

33.	1.308	1.692	2.035	2.445	2.733	3.356
34.	1.307	1.691	2.032	2.441	2.728	3.348
35.	1.306	1.690	2.030	2.438	2.724	3.340
36.	1.306	1.688	2.028	2.434	2.719	3.333
37.	1.305	1.687	2.026	2.431	2.715	3.326
38.	1.304	1.686	2.024	2.429	2.712	3.319
39.	1.304	1.685	2.023	2.426	2.708	3.313
40.	1.303	1.684	2.021	2.423	2.704	3.307
41.	1.303	1.683	2.020	2.421	2.701	3.301
42.	1.302	1.682	2.018	2.418	2.698	3.296
43.	1.302	1.681	2.017	2.416	2.695	3.291
44.	1.301	1.680	2.015	2.414	2.692	3.286
45.	1.301	1.679	2.014	2.412	2.690	3.281
46.	1.300	1.679	2.013	2.410	2.687	3.277

Trust and responsibility

NNE and Pharmaplan have joined forces to create NNE Pharmaplan, the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries.

Inés Aréizaga Esteva (Spain), 25 years old
Education: Chemical Engineer

– You have to be proactive and open-minded as a newcomer and make it clear to your colleagues what you are able to cope. The pharmaceutical field is new to me. But busy as they are, most of my colleagues find the time to teach me, and they also trust me. Even though it was a bit hard at first, I can feel over time that I am beginning to be taken seriously and that my contribution is appreciated.



NNE Pharmaplan is the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries. We employ more than 1500 people worldwide and offer global reach and local knowledge along with our all-encompassing list of services.
nnepharmaplan.com

nne pharmaplan®



47.	1.300	1.678	2.012	2.408	2.685	3.273
48.	1.299	1.677	2.011	2.407	2.682	3.269
49.	1.299	1.677	2.010	2.405	2.680	3.265
50.	1.299	1.676	2.009	2.403	2.678	3.261

T-Table Part C df = 51-75. Area is to the left
Probability less than the critical value $T_{1-\alpha, v}$

df	0.90	0.95	0.975	0.99	0.995	0.999
51.	1.298	1.675	2.008	2.402	2.676	3.258
52.	1.298	1.675	2.007	2.400	2.674	3.255
53.	1.298	1.674	2.006	2.399	2.672	3.251
54.	1.297	1.674	2.005	2.397	2.670	3.248
55.	1.297	1.673	2.004	2.396	2.668	3.245
56.	1.297	1.673	2.003	2.395	2.667	3.242
57.	1.297	1.672	2.002	2.394	2.665	3.239
58.	1.296	1.672	2.002	2.392	2.663	3.237
59.	1.296	1.671	2.001	2.391	2.662	3.234
60.	1.296	1.671	2.000	2.390	2.660	3.232
61.	1.296	1.670	2.000	2.389	2.659	3.229
62.	1.295	1.670	1.999	2.388	2.657	3.227
63.	1.295	1.669	1.998	2.387	2.656	3.225
64.	1.295	1.669	1.998	2.386	2.655	3.223
65.	1.295	1.669	1.997	2.385	2.654	3.220
66.	1.295	1.668	1.997	2.384	2.652	3.218
67.	1.294	1.668	1.996	2.383	2.651	3.216
68.	1.294	1.668	1.995	2.382	2.650	3.214
69.	1.294	1.667	1.995	2.382	2.649	3.213
70.	1.294	1.667	1.994	2.381	2.648	3.211
71.	1.294	1.667	1.994	2.380	2.647	3.209

72.	1.293	1.666	1.993	2.379	2.646	3.207
73.	1.293	1.666	1.993	2.379	2.645	3.206
74.	1.293	1.666	1.993	2.378	2.644	3.204
75.	1.293	1.665	1.992	2.377	2.643	3.202

TABLE IV t-DISTRIBUTION

I	$\lambda = E(x)$									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	0.905	0.819	0.741	0.67	0.607	0.549	0.497	0.449	0.407	0.368
1	0.995	0.982	0.963	0.938	0.607	0.878	844	8099	0.772	0.736
2	1	0.999	0.996	0.992	0.91	0.977	0.966	0.953	0.937	0.92
3		1	1	0.999	0.986	0.977	0.994	0.991	0.987	0.981
4				1	1	1	0.999	0.999	0.998	0.996
5							1	1	1	0.999
6										1

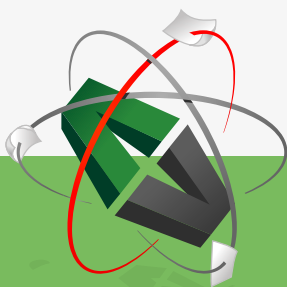
II	$\lambda = E(x)$									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
0	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.15	0.135
1	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.9	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.995	0.992	0.989	0.986	0.981	0.976	0.97	0.964	0.956	0.947
5	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.99	0.987	0.983
6	1	1	1	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7				1	1	1	1	0.999	0.999	0.999
8								1	1	1

III	$\lambda = E(x)$									
x	2.2	2	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0	0.111	0.091	0.074	0.061	0.05	0.041	0.033	0.027	0.022	0.018
1	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126	0.107	0.092
2	0.623	0.57	0.518	0.469	0.423	0.38	0.34	0.303	0.269	0.238
3	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515	0.473	0.433
4	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706	0.668	0.629
5	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844	0.816	0.785

6	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927	0.909	0.889
7	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969	0.96	0.949
8	1	0.999	0.999	0.998	0.996	0.994	0.992	0.988	0.984	0.979
9		1	1	0.999	0.999	0.998	0.997	0.996	0.994	0.992
10				1	1	1	0.999	0.999	0.998	0.997
11							1	1	0.999	0.999
12									1	1

	$\lambda = E(x)$									
x	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
0	0.015	0.012	0.01	0.008	0.007	0.006	0.005	0.004	0.003	0.002
1	0.078	0.066	0.056	0.048	0.04	0.034	0.029	0.024	0.021	0.017
2	0.21	0.185	0.163	0.143	0.125	0.109	0.095	0.082	0.072	0.062
3	0.395	0.359	0.326	0.294	0.265	0.238	0.213	0.191	0.17	0.151
4	0.59	0.551	0.513	0.478	0.44	0.406	0.373	0.342	0.313	0.285

This e-book
is made with
SetaPDF



PDF components for PHP developers

www.setasign.com



5	0.753	0.72	0.686	0.651	0.616	0.581	0.546	0.512	0.478	0.446
6	0.867	0.844	0.818	0.791	0.762	0.732	0.702	0.67	0.638	0.606
7	0.936	0.921	0.905	0.887	0.867	0.845	0.822	0.797	0.771	0.744
8	0.972	0.964	0.955	0.944	0.932	0.918	0.903	0.886	0.867	0.847
9	0.989	0.985	0.98	0.975	0.968	0.96	0.951	0.941	0.929	0.916
10	0.996	0.994	0.992	0.99	0.986	0.982	0.977	0.972	0.965	0.957
11	0.999	0.998	0.997	0.996	0.995	0.993	0.99	0.988	0.984	0.98
12	1	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.991
13		1	1	1	0.999	0.999	0.999	0.998	0.997	0.996
14					1	1	0.999	0.999	0.999	0.999
15							1	1	1	0.999
										1

	$\lambda = E(X)$									
x	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11
0	0.002	0.001	0.001	0	0	0	0	0	0	0
1	0.011	0.007	0.005	0.003	0.002	0.001	0.001	0	0	0
2	0.043	0.03	0.02	0.014	0.009	0.006	0.004	0.003	0.002	0.001
3	0.112	0.082	0.059	0.042	0.03	0.021	0.015	0.01	0.007	0.005
4	0.224	0.173	0.132	0.1	0.074	0.055	0.04	0.029	0.021	0.014
5	0.369	0.301	0.241	0.191	0.15	0.116	0.089	0.067	0.05	0.038
6	0.527	0.45	0.378	0.313	0.256	0.207	0.165	0.13	0.102	0.079
7	0.673	0.599	0.525	0.453	0.386	0.324	0.269	0.22	0.179	0.143
8	0.792	0.729	0.662	0.593	0.523	0.456	0.392	0.333	0.279	0.232
9	0.877	0.83	0.776	0.717	0.665	0.617	0.572	0.528	0.485	0.441
10	0.933	0.901	0.862	0.816	0.763	0.706	0.645	0.583	0.521	0.46
11	0.966	0.947	0.921	0.888	0.849	0.803	0.752	0.697	0.639	0.579
12	0.984	0.973	0.957	0.936	0.909	0.876	0.836	0.792	0.742	0.689
13	0.993	0.987	0.978	0.966	0.949	0.926	0.898	0.864	0.825	0.781
14	0.997	0.994	0.99	0.983	0.973	0.959	0.94	0.917	0.888	0.854
15	0.999	0.998	0.995	0.992	0.986	0.978	0.967	0.951	0.932	0.907
16	1	0.999	0.998	0.996	0.993	0.989	0.982	0.973	0.96	0.944
17		1	0.999	0.998	0.997	0.995	0.991	0.986	0.978	0.968
18			1	0.999	0.999	0.998	0.996	0.993	0.988	0.982
19				1	0.999	0.999	0.998	0.997	0.994	0.991

20		1	1	0.999	0.998	0.997	0.995
21				1	0.999	0.999	0.998
22					1	1	0.999
23							1

x	$\lambda = E(X)$									
	11.5	12	12.5	13	13.5	14	14.5	15	15.5	16
0	0	0	0	0						
1	0	0	0	0						
2	0.001	0.001	0	0						
3	0.003	0.002	0.002	0.001	0.001	0	0	0	0	0
4	0.011	0.008	0.005	0.004	0.003	0.002	0.001	0.001	0.001	0
5	0.028	0.02	0.015	0.011	0.008	0.006	0.004	0.003	0.002	0.001
6	0.06	0.046	0.035	0.026	0.019	0.014	0.01	0.008	0.006	0.004
7	0.114	0.09	0.07	0.054	0.041	0.032	0.024	0.018	0.013	0.01
8	0.191	0.155	0.125	0.1	0.079	0.062	0.048	0.037	0.029	0.022
9	0.289	0.242	0.201	0.166	0.135	0.109	0.088	0.07	0.055	0.043
10	0.402	0.347	0.297	0.252	0.211	0.176	0.145	0.118	0.096	0.077
11	0.52	0.462	0.406	0.353	0.304	0.26	0.22	0.185	0.124	0.127
12	0.633	0.576	0.519	0.463	0.409	0.358	0.311	0.268	0.228	0.193
13	0.733	0.682	0.628	0.573	0.518	0.464	0.413	0.363	0.317	0.275
14	0.815	0.772	0.725	0.675	0.623	0.57	0.518	0.466	0.415	0.368
15	0.878	0.844	0.806	0.764	0.718	0.669	0.619	0.568	0.517	0.467
16	0.92	0.899	0.869	0.835	0.798	0.756	0.711	0.664	0.615	0.566
17	0.954	0.937	0.916	0.89	0.861	0.827	0.79	0.749	0.705	0.659
18	0.974	0.963	0.948	0.93	0.908	0.883	0.853	0.819	0.782	0.742
19	0.986	0.979	0.969	0.957	0.942	0.923	0.901	0.875	0.846	0.812
20	0.992	0.988	0.983	0.975	0.965	0.952	0.963	0.917	0.894	0.868
21	0.996	0.994	0.991	0.986	0.98	0.971	0.96	0.947	0.93	0.911
22	0.998	0.997	0.995	0.992	0.989	0.983	0.976	0.967	0.956	0.942
23	0.999	0.999	0.998	0.996	0.994	0.991	0.986	0.981	0.973	0.963
24	1	0.999	0.999	0.998	0.997	0.995	0.992	0.989	0.984	0.978
25		1	0.999	0.999	0.998	0.997	0.996	0.994	0.991	0.987

26	1	1	0.999	0.999	0.998	0.997	0.995	0.993
27			1	0.999	0.999	0.998	0.997	0.996
28				1	0.999	0.999	0.999	0.998
29					1	1	0.999	0.999
30							1	0.999
31								1

Table V Poisson distribution

r	Chi-Square							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
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	9.21
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28



Sharp Minds - Bright Ideas!

Employees at FOSS Analytical A/S are living proof of the company value - First - using new inventions to make dedicated solutions for our customers. With sharp minds and cross functional teamwork, we constantly strive to develop new unique products - Would you like to join our team?

FOSS works diligently with innovation and development as basis for its growth. It is reflected in the fact that more than 200 of the 1200 employees in FOSS work with Research & Development in Scandinavia and USA. Engineers at FOSS work in production, development and marketing, within a wide range of different fields, i.e. Chemistry, Electronics, Mechanics, Software, Optics, Microbiology, Chemometrics.

We offer
A challenging job in an international and innovative company that is leading in its field. You will get the opportunity to work with the most advanced technology together with highly skilled colleagues.

Read more about FOSS at www.foss.dk - or go directly to our student site www.foss.dk/sharpminds where you can learn more about your possibilities of working together with us on projects, your thesis etc.

The Family owned FOSS group is the world leader as supplier of dedicated, high-tech analytical solutions which measure and control the quality and production of agricultural, food, pharmaceutical and chemical products. Main activities are initiated from Denmark, Sweden and USA with headquarters domiciled in Hillerød, DK. The products are marketed globally by 23 sales companies and an extensive net of distributors. In line with the corevalue to be 'First', the company intends to expand its market position.



Dedicated Analytical Solutions

FOSS
 Slangerupgade 69
 3400 Hillerød
 Tel. +45 70103370
www.foss.dk



5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21
11	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72
12	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22
13	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69
14	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14
15	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58
16	5.812	6.908	7.962	9.312	23.54	26.30	28.84	32.00
17	6.408	7.564	8.672	10.08	24.77	27.59	30.19	33.41
18	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.80
19	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19
20	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57
21	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93
22	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31
26	12.20	13.84	15.38	17.29	35.56	38.88	41.92	45.64
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
40	22.16	24.43	26.51	29.05	51.80	55.76	59.34	63.69

Table VI Chi-Square distribution

Df of		Df of Num									
P(F<f)	Den.	1	2	3	4	5	6	7	8	9	10
0.95	1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9
0.975		647.8	799.5	864.2	899.6	921.9	937.1	948.2	956.7	936.3	968.6
0.99		4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
0.95	2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4
0.975		38.51	39	39.17	39.25	39.3	39.33	39.36	39.37	39.39	39.4
0.99		98.5	99	99.17	99.25	99.3	99.33	99.36	99.37	99.39	99.4
0.95	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.975		17.44	16.04	15.44	15.1	14.88	17.73	14.62	14.54	14.47	14.42
0.99		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
0.95	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96
0.975		12.22	10.65	9.98	9.6	9.36	9.2	9.07	8.98	5.9	8.84
0.99		21.2	18	26.69	15.98	15.52	15.21	14.98	14.8	14.66	14.55
0.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.975		10.01	8.43	7.76	7.039	7.15	6.98	6.85	6.76	6.68	6.62
0.99		16.26	13.27	12.06	11.39	1097	10.67	10.46	10.29	10.16	10.05
0.95	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06
0.975		8.81	7.26	6.6	6.23	5.99	5.82	5.7	5.6	5.52	5.46
0.99		13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.1	7.98	7.87
0.95	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
0.975		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.9	4.82	4.76
0.99		12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
0.95	8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35
0.975		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.3
0.99		12.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
0.95	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
0.975		7.21	5.71	5.08	4.72	4.48	4.32	4.2	4.1	4.03	3.96
0.99		10.56	8.02	6.99	6.42	6.06	5.8	5.61	5.47	5.35	5.26
0.95	10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
0.975		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
0.99		10.04	7.56	6.55	5.99	5.64	5.39	5.2	5.06	4.94	4.85

Table VII F-distribution

Critical values for Correlation Coefficient

n	C.V.	n	C.V.
3	0.997	17	0.482
4	0.95	18	0.468
5	0.878	19	0.456
6	0.811	20	0.444
7	0.754	21	0.433
8	0.707	22	0.423
9	0.666	23	0.413
10	0.632	24	0.404
11	0.602	25	0.396
12	0.576	26	0.388
13	0.553	27	0.381

"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



14	0.532	28	0.374
15	0.514	29	0.367
16	0.497	30	0.361

Table VIII Critical values for Correlation Coefficient

Critical Values for the Signed-Rank Test

n	One-Tailed 0.01	One-Tailed 0.025	One-Tailed 0.05
	Two-Tailed .02	Two-Tailed 0.05	Two-Tailed 0.10
5			1
6		1	2
7	0	2	4
8	2	4	6
9	3	6	8
10	5	8	11
11	7	11	14
12	10	14	17
13	13	17	21
14	16	21	26
15	20	25	30
16	24	30	36
17	28	35	41
18	33	40	47
19	38	46	54
20	43	52	60
21	49	59	68
22	56	66	75
23	62	73	83

24	69	81	92
25	77	90	101
26	85	98	110
27	93	107	120
28	102	117	130
29	111	127	141
30	120	137	152

TABLE IX Critical Values for the Signed-Rank Test

Critical Values for the Rank-Sum Test

n1	n2																
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																	
2					0	0	0	0	1	1	1	1	1	2	2	2	2
3		0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5		2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7				8	10	12	14	16	18	20	22	24	26	28	30	32	34
8					13	15	17	19	22	24	26	29	31	34	36	38	41
9						17	20	23	26	28	31	34	37	39	42	45	48
10							23	26	29	33	36	39	42	45	48	52	55
11								30	33	37	40	44	47	51	55	58	62
12									37	41	45	49	53	57	61	65	69
13										45	50	54	59	63	67	72	76
14											55	59	64	67	74	78	83
15												64	70	75	80	85	90
16													75	81	86	92	98
17														87	93	99	105
18															99	106	112
19																113	119
20																	127

TABLE IX Critical Values for the Rank- Sum Test

APPENDIX B

ANSWERS TO SELECTED EXERCISES

Chapter 1 Describing Data Graphically

1.4 a) $x_1 + x_2 + x_3 + x_4$ d) $X_1Y_1 + X_2Y_2 + \dots + X_7Y_7$.

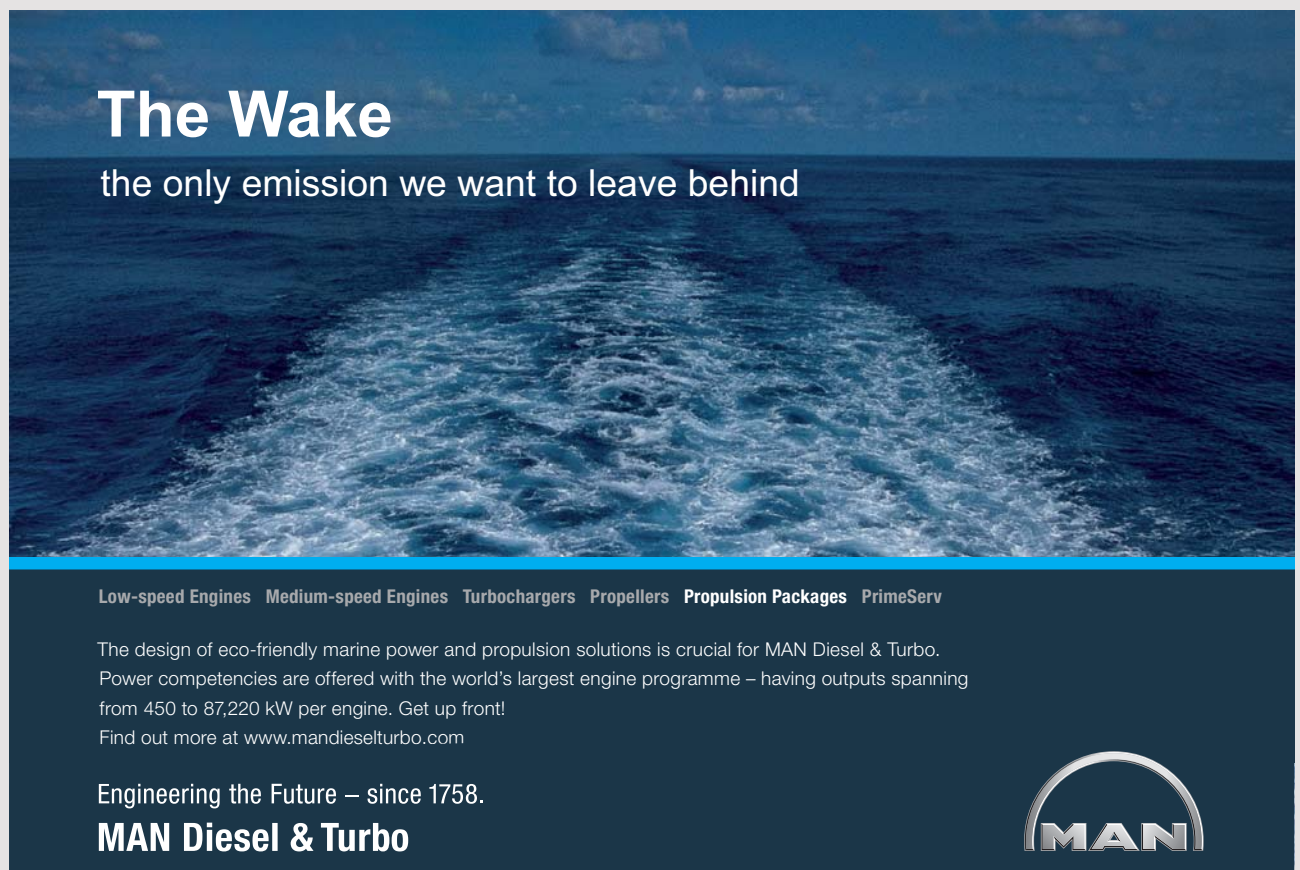
1.5 b) $\sum_{i=1}^4 Y_i^2$

1.6 c) $\sum_{i=1}^4 X_iY_i = 11$

1.7 a) For $j = 1$: $\sum_{i=1}^3 X_{i1} = X_{11} + X_{21} + X_{31} = 7$

1.9 a) True c) False e) True g) True

1.19 Quantitative



The Wake

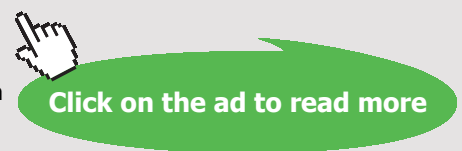

the only emission we want to leave behind

Low-speed Engines Medium-speed Engines Turbochargers Propellers Propulsion Packages PrimeServ

The design of eco-friendly marine power and propulsion solutions is crucial for MAN Diesel & Turbo. Power competencies are offered with the world's largest engine programme – having outputs spanning from 450 to 87,220 kW per engine. Get up front! Find out more at www.mandieselturbo.com

Engineering the Future – since 1758.

MAN Diesel & Turbo



1.22 Qualitative

1.23 Qualitative

1.25 Continuous

1.27 Discrete

1.29 Continuous

1.30 Discrete

Chapter 2 Describing Data Numerically

2.3 22.75

$$2.5 \quad \sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X} = n\bar{X} - n\bar{X} = 0$$

2.7 b) The median is $14.5 = (1/2)[14.4 + 14.6]$.

2.10 a) The mean is 6, b) There is no mode c) The midrange = 5.5, The median = 7

2.11 a) The measures of central tendency are: Mean = 7.2, Median = 7, Mode = 7, Midrange = 8

2.12 b) The median = 7

Chapter 3 Probability

3.1 c) d), and j) cannot be used for a probability of an event.

3.2 i) 3/13 vi) 0 viii) 0

3.4 a) $B = \{3.7, 3.8\}$ b) Sociology major, female and with GPA 3.7.

3.7 d) $0.2 + 0.32 - 0.07 = 0.45$

3.10 b) A has 15 outcomes, almost all except TTTT. Thus $P(A) = 15/16$

c) B has 15 outcomes, almost all except HHHH. Thus $P(B) = 15/16$

3.13 $A = (A \text{ and } B) \text{ or } (A \text{ and } B')$; $P(A) = P(A \text{ and } B) + P(A \text{ and } B')$

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

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

$$0.77 + 0.87 = 2P(A) + P(B) + P(B') - P(A) = 1 + P(A)$$

$$\text{Hence } P(A) = 1.64 - 1 = 0.64.$$

3.15 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$0.7 = 0.4 + 0.5 - P(A \text{ and } B)$ Therefore $P(A \text{ and } B) = 0.2$. So A and B are independent since $P(A) \cdot P(B) = P(A \text{ and } B) = 0.2$

a) $P(A|B) = P(A) = 0.40$, b) $P(B^c|A) = P(B^c) = 1 - P(B) = 1 - 0.50 = 0.5$.

3.20 a) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = P(A) + P(B) - P(A) \cdot P(B)$
 $= 0.6 + 0.22 - 0.6 \times 0.22 = 0.6880$.

3.22 a) $1/3$ b) $2/3$ c) $1/2$

3.23 Hint: put $(B \text{ or } C) = D$, and apply the addition rule twice. And do the substitution.

3.25 b) $P(F_1|S_1) = P((F_1 \text{ and } S_1) / P(S_1) = 39/67$

3.27 a) 0.3 b) 0.6

3.29 Useful hints: $P(A) = 0.37$ $P(C) = 0.48$, and $P(B) = 0.50$

3.31 Part d) in this exercise should read the following: The conditional probability that Box B_2 had been selected, given that a red chip was drawn, i.e. find $P(B_2 | R)$

Chapter 4 Discrete Probability Distributions

4.1 $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - 0.7351 = 0.2649$

4.3 b) it is not a probability distribution since the sum > 1.0

4.4 a) $1 - 0.7 = 0.3$

4.5 a) the mean = 4.1

4.7

X	0	1
P_x	7/18	11/18

4.10 a) 0.55 b) 0.10

4.12 a) The mean = 301

gaiteye[®]
Challenge the way we run

**EXPERIENCE THE POWER OF
FULL ENGAGEMENT...**

**RUN FASTER.
RUN LONGER..
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY
WWW.GAITEYE.COM**

4.13 $\mu = 2$ and $\sigma = 1$

4.15 a) $X \sim \text{Bin}(15, 0.65)$ e) $\mu = 9.75$, and $\sigma = 1.84729$

4.18 a) $P(3) = 0.1954$ f) The mean = 4

4.20 0.08 20 feet 2

4.21 a) 0.1462 b) 0.6160

4.24 b) 0.7358

Chapter 5 Continuous Probability Distributions

5.1 b) 0.9545

5.4 $P(E_5) = 0.10$ and $P(E_4) = 0.2$

5.10 a) 1.833 b) -1.746 c) 2.101 d) -1.706 e) -2.602 f) 2.718

5.12 a) $f(x) = (1/\theta) \exp(-x/\theta)$, $x \geq 0$, $M(t) = (1 - 3t)^{-1}$ means $\theta = 3$

5.15 a) 0.025

5.17 $X \sim \text{Gamma}(12, 2) = \text{Gamma}(\alpha, \beta)$ thus $E(X) = 24$

5.18 $\mu = \frac{d_2}{d_2 - 2}$, $d_2 > 2$.

5.22 a) $f(x) = x/2$, $0 \leq x \leq 2$; c) Yes

5.23 a) $f(x) = 2(1-x)$, $0 \leq x \leq 1$; ($E(X) = 1/3$, $E(Y) = 2/3$, $E(Y^2) = 1/2$).

5.26 a) $c = 8$.

Chapter 6 Sampling Distributions

- 6.2** a) $\mu_{\bar{x}} = 80$ $\sigma_{\bar{x}} = 2$
- 6.4** b) 0.0668 c) 0.7970
- 6.5** a) $\bar{X} \sim N(64, 9)$ c) 0.6560
- 6.6** a) 0.3520 f) 0.9844
- 6.7** c) 0.0205
- 6.8** The samples are: 3, 3 3, 5 3, 7 5, 5 5, 7 7, 7
The Means are 3 4 5 5 6 7
- 6.10** a) $\mu = 5.3$ and $\sigma = 0.9$
- 6.11** $\bar{X} \sim N(80, 1)$ $\bar{Y} \sim (75, \frac{1}{4})$ $\bar{X} - \bar{Y} \sim (5, 1.25)$
- 6.15** c) has the smallest
- 6.17** The distribution is approximately normal i.e. $N(p, p(1-p)/n)$
- 6.20** b) 0.0043

Chapter 7 Simple Linear Regression and Correlation

- 7.3** c) The predicted equation is given by $y = ax + b$, with $a = -0.7136$, $b = 6.55$, $r^2 = 0.8981$, and $r = -0.9477$
- 7.5** b) The predicted equation is given by $y = ax + b$, with $a = 1.43636$ $b = 9.2727$, $r^2 = 0.9144$, and $r = 0.9563$.
- 7.9** $r = 0.2397$
- 7.10** Data needs to be rearranged.
- 7.11** $\ln P + \gamma \ln V = \ln C$ i.e. $\ln P = -\gamma \ln V + \ln C \rightarrow y = ax + b$.

- 7.12 a) $y = ax + b$ where $a = 0.016$, $b = 63.4045$, $r^2 = 0.004329$, and $r = 0.065798$. b) 71.2277 in.
- 7.13 b) The residuals are, respectively: -1, 1.5 -1 1.5 -1.
The sum of their squares is 7.5
- 7.16 b) The residuals are, respectively: 1, 1 -4 1 1.
The sum of their squares is 20

**Technical training on
WHAT you need, *WHEN* you need it**

At IDC Technologies we can tailor our technical and engineering training workshops to suit your needs. We have extensive experience in training technical and engineering staff and have trained people in organisations such as General Motors, Shell, Siemens, BHP and Honeywell to name a few.

Our onsite training is cost effective, convenient and completely customisable to the technical and engineering areas you want covered. Our workshops are all comprehensive hands-on learning experiences with ample time given to practical sessions and demonstrations. We communicate well to ensure that workshop content and timing match the knowledge, skills, and abilities of the participants.

We run onsite training all year round and hold the workshops on your premises or a venue of your choice for your convenience.

**For a no obligation proposal, contact us today
at training@idc-online.com or visit our website
for more information: www.idc-online.com/onsite/**

**OIL & GAS
ENGINEERING**

ELECTRONICS

**AUTOMATION &
PROCESS CONTROL**

**MECHANICAL
ENGINEERING**

**INDUSTRIAL
DATA COMMS**

**ELECTRICAL
POWER**

Phone: +61 8 9321 1702
Email: training@idc-online.com
Website: www.idc-online.com

**IDC
TECHNOLOGIES**

APPENDIX C

FORMULAS

Summarizing Data Numerically

- Sample Arithmetic Mean:

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n = \sum_{i=1}^n X_i / n$$

- Sample Geometric Mean:

$$\bar{G} = (X_1 \cdot X_2 \cdot \dots \cdot X_n)^{1/n}, \quad \bar{G} = \left(\prod_1^n X_i \right)^{1/n}$$

- Sample Harmonic Mean:

$$\bar{H} = n / \sum_1^n (1/X_i)$$

- Finite Population Arithmetic Mean:

$$\mu = (x_1 + x_2 + \dots + x_N) / N = \sum_{i=1}^N X_i / N$$

- Finite Population Geometric Mean: $\bar{G} = (X_1 \cdot X_2 \cdot \dots \cdot X_N)^{1/N} = \left(\prod_1^N X_i \right)^{1/N}$

- Finite Population Harmonic Mean: $\bar{H} = N / \sum_1^N (1/X_i)$

- Range = Max \sqcup Min

- Finite Population Variance:

$$\sigma^2 = \frac{\sum_{i=1}^N (Y_i - \mu)^2}{N} \quad \text{OR} \quad \sigma^2 = \frac{\sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i \right)^2}{N}}{N}$$

- Finite Population Standard Deviation:

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

- Sample Variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \text{ OR } S^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}, \text{ OR}$$

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_i^2}{n(n-1)}$$

- Sample Standard Deviation:

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

For Grouped Data

- Weighted Arithmetic Mean:

$$\bar{x} = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i}$$

where f_i is the weight for the observation X_i , or its frequency, with n different classes

- Finite Population Arithmetic Mean:

$$\mu = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i}, \text{ with } n \text{ classes.}$$

- Sample Arithmetic Mean:

$$\bar{x} = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i}, \text{ with } n \text{ classes}$$

- Finite Population Variance: $\sigma^2 = \frac{\sum_{i=1}^n (X_i)^2 f_i - \frac{\left(\sum_{i=1}^n X_i f_i\right)^2}{\sum_{i=1}^n f_i}}{\sum_{i=1}^n f_i}, n \text{ classes}$

- Sample Variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i)^2 f_i - \frac{(\sum_{i=1}^n X_i f_i)^2}{\sum_{i=1}^n f_i}}{(\sum_{i=1}^n f_i) - 1}, \text{ n classes}$$

- Z-Score for an observation in a Sample: $Z = \frac{(Y - \bar{Y})}{S}$
- Z-Score for an observation in a Finite Population:

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

- Interquartile Range: $\text{IQR} = Q_3 - Q_1$
- Percentile Rank (position of p^{th} percentile)
 $i = (p/100) \cdot n$
- Percentile of an observation, x :
 $x = (\text{Number of points} < x) \cdot (100/n)$
- Determining the k^{th} percentile:
 $i = (k/100) \cdot (n+1)$.

If i is not an integer, find the mean of the observations on either side of i .

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements



MAERSK

- Five-Number Summaries:
Minimum, Q_1 , Median, Q_3 , and Maximum
- Lower Fence:
 $LF = Q_1 - 1.5 \cdot (IQR)$
- Upper Fence:
 $UF = Q_3 + 1.5 \cdot (IQR)$

(Any observation jumping over either fence is termed as an **Outlier**)

Describing the Relationship between Two Variables

- Based on the definition of the Z-Score,

Z-Score = [Data point – Mean of the Data Set] / Standard Deviation of the Data Set, and if

$$Z_x = \frac{x_i - \bar{x}}{S_x}, \text{ and } Z_y = \frac{y_i - \bar{y}}{S_y}, \text{ for } i = 1, 2, \dots, n.$$

- Sample Correlation Coefficient r is given by: $r = \frac{\sum Z_x Z_y}{n-1}$ OR

$$r = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\left(\sqrt{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} \right) \left(\sqrt{\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}} \right)} \quad r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

- The Regression Equation (The Estimated Regression Line):

$$\hat{y} = b_0 + b_1 x,$$

where \hat{y} is the predicted value, and by using the least square estimation method we have:

- Slope of the Regression $b_1 = \frac{S_{xy}}{S_{xx}} = r \cdot \frac{S_y}{S_x}$

- Y-Intercept: $b_0 = \bar{y} - b_1 \bar{x}$

Total Sum of Squares:

$$\sum_{i=1}^n (Y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (Y_i - \hat{y}_i)^2$$

i.e. $SST_{\text{total}} = SSR_{\text{reg}} + SSE_{\text{residuals}}$

- Total Sum of Squares, $SST = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} = S_{yy}$

- Residual Sum of Squares

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

- Estimate of error variance

$$\sigma^2 = SSE / (n-2)$$

- Coefficient of Determination, r^2 , is:

$$r^2 = SSR/SST = \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}}$$

It is expressed as a percentage to represent the percentage in y-variability that was explained by the linear regression including x.

- Exponential Equation of Best Fit is:

$$y = ab^x,$$

Linear: $\log y = \log a + x \cdot \log b$; OR

Linear: $\ln y = \ln a + x \cdot \ln b$; $a, b, y > 0$

- Power Equation of Best Fit is given by:

$$y = ax^b;$$

Linear: $\log y = \log a + b \cdot \log x$; OR

Linear: $\ln y = \ln a + b \cdot \ln x$; $a, x, y > 0$

- Residual = Observed y – Predicted y

$$R = y - \hat{y}$$

Matrix Presentation for Simple Linear Model

- The Simple Linear Regression Model in one Explanatory variable and a response, that is represented by $Y = \beta_0 + \beta_1 X + \varepsilon$, and based on n data pair points, can be put in matrix notation as follows: $Y = X\beta + \varepsilon$, where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \dots \\ \dots & \dots \\ 1 & x_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\text{and } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

In other words, Y is a column vector, i.e. one column and n rows, X is an $n \times 2$ matrix with n rows and 2 columns, β is a column vector, 2×1 matrix, i.e. 2 rows and one column, while ε is an $n \times 1$ matrix, i.e. a column vector.



www.job.oticon.dk

oticon
PEOPLE FIRST



The normal equations for finding the parameters of the model are given by

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'X)^{-1} \cdot X'Y,$$

Such that X' is the transpose of the matrix X (it is the matrix with 2 rows and n columns), $(X'X)^{-1}$ is the inverse matrix of $X'X$, where $X'X$ is the matrix given by

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}, \text{ and}$$

$$S^2 = \frac{[(Y - \mathbf{Xb})^T (Y - \mathbf{Xb})]}{n-p} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-p},$$

is the estimate for the error variance, when there are k predictors and p parameters; $p = k + 1$

Inference on the Least-squares Regression Model and Multiple Regression

- Standard Error of The Estimate

$$S_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}}$$

- Standard Error of b_1

$$S_{b_1} = \frac{S_e}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

- Test Statistic for the slope, b_1

$$t_0 = \frac{b_1 - \beta_1}{S_{b_1}}$$

- **Confidence Interval for the slope of the Regression Line**

A $100(1-\alpha) \%$ C.I. on the slope, β_1 , is given by

$$b_1 \pm t_{\alpha/2} \cdot S_{b_1}$$

Where $t_{\alpha/2}$ are calculated based on $n-2$ degrees of freedom

- **Confidence Interval about the Mean Response of y , \hat{y}**

A $100(1-\alpha)$ % C.I. for the mean response of y , \hat{y} , is given by

$$\hat{y} \pm t_{\alpha/2} \cdot S_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

where x^* is the given value of the explanatory variable, and $t_{\alpha/2}$ are calculated based on $n-2$ degrees of freedom

- **Prediction Interval about an Individual Response, \hat{y}**

A $100(1-\alpha)$ % Prediction Interval for the individual response of y , \hat{y} , is given by

$$\hat{y} \pm t_{\alpha/2} \cdot S_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

where x^* is the given value of the explanatory variable, and $t_{\alpha/2}$ are calculated based on $n-2$ degrees of freedom

Probability

- Empirical Probability

$$P(E) = \frac{\text{Frequency of } E}{\text{Number of Trials of Experiment}}$$

- Classical Probability

$$P(E) = \frac{\text{Number of ways that } E \text{ occurs}}{\text{Number of possible outcomes}}$$

- Probabilities for Complements; where A' or A^c is the complement of A

$$P(A) + P(A') = 1, \text{ or}$$

$$P(A^c) = 1 - P(A),$$

- General Addition Rule:

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Special Addition Rule:

$$P(A \cup B) = P(A) + P(B), \text{ when } A \text{ \& } B \text{ are mutually exclusive, i.e. } (A \cap B) = \phi.$$

- Conditional Probabilities

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

- General Multiplication Rule for two events:

$$P(A \cap B) = P(A) \cdot P(B|A); \text{ or}$$

$$P(A \cap B) = P(B) \cdot P(A|B)$$

- Multiplication Rule: A and B are independent if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$

- Multiplication Rule for n Independent Events

$$P(E \text{ and } F \text{ and } G \dots) = P(E) \cdot P(F) \cdot P(G) \cdot \dots$$

- n Factorial: $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1$, where n is a positive integer. $0! = 1$, by convention

- Permutation of r items chosen from n distinct items:

$${}_n P_r = n! / (n - r)!, \text{ provided } 0 \leq r \leq n$$



Schlumberger

WHY WAIT FOR PROGRESS?

DARE TO DISCOVER

Discovery means many different things at Schlumberger. But it's the spirit that unites every single one of us. It doesn't matter whether they join our business, engineering or technology teams, our trainees push boundaries, break new ground and deliver the exceptional. If that excites you, then we want to hear from you.

careers.slb.com/recentgraduates

- Combination of r items chosen from n distinct items:

$${}_n C_r = \binom{n}{r} = \frac{n!}{(n-r)! \cdot r!}, \text{ with } 0 \leq r \leq n$$

- Permutation with repetition: n_1 of one type, n_2 of a second type, ..., with $n_1 + n_2 + \dots + n_k = n$

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$$

Random Variables

Discrete Random Variables

- Mean of a discrete random variable:

$$\mu = \sum_x X \cdot P(X)$$

- Binomial Probability Mass Function:

$$P(X = x) = P(x) = {}_n C_x \cdot p^x \cdot (1-p)^{n-x},$$

$x = 0, 1, 2, \dots, n$

- Mean of a Binomial Distribution;

$$X \sim \text{Bin}(n, p), \mu = n \cdot p$$

- Poisson Probability Mass Function:

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

$$x = 0, 1, 2, \dots$$

- Mean of a Poisson Random Variable:

$$\mu = \lambda$$

- Variance of a discrete random variable:

$$\sigma^2 = \sum_{i=1}^n (X_i - \mu)^2 \cdot P(X) = \sum_{i=1}^n \{X_i^2 \cdot P(X)\} - \mu^2$$

- Variance of a Binomial Distribution;

$$X \sim \text{Bin}(n, p), \sigma^2 = n \cdot p \cdot (1-p)$$

- Variance of a Poisson Random Variable:

$$\sigma^2 = \lambda$$

- Standard deviation:

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

Continuous Random variables

- Mean of a continuous random variable: $\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$

- Variance of a continuous random variable:

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

- Standard deviation:

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

Probability of a Random Variable over an interval;

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

$$= P(a \leq X \leq b) = \int_a^b f(x) dx.$$

Sampling Distributions

- Sampling Distribution of \bar{X} :

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}}^2 = \sigma^2/n$$

Provided that the population has a mean μ , and with a finite variance σ^2

- CLT for standardizing on the sample mean: $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

- Sampling Distribution of P :

$$P = x/n \quad \mu_p = p \text{ and } \sigma_p^2 = p \cdot (1 - p)/n$$

- CLT for standardizing on P : $Z = \frac{P - p}{\sqrt{\frac{p(1-p)}{n}}} \sqsim N(0,1)$

Confidence Intervals on One Parameter

100(1- α) % Confidence Interval on μ :

- Population Variance is known, and any sample size: $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.
- Sample Size needed to estimate the population mean with a margin of error E:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2,$$

rounded **up** to the next whole number.

- Population Variance is Unknown, large Sample Size, $n \geq 30$: $\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$
- ❖ Population Variance is Unknown, Small Sample Size, $n < 30$:

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

Note: $t_{\alpha/2}$ is found based on $n-1$ degrees of freedom



PREPARE FOR A LEADING ROLE.

English-taught MSc programmes in engineering: Aeronautical, Biomedical, Electronics, Mechanical, Communication systems and Transport systems. No tuition fees.

→ liu.se/master

li.u LINKÖPING UNIVERSITY

100(1- α) % Confidence Interval on p:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{provided } n \cdot \hat{p}(1-\hat{p}) \geq 10$$

- Sample Size, with a previous estimate on p: $n = \hat{p}(1-\hat{p}) \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2$,
Rounded up to the next whole number
- Sample Size, without a previous estimate on p: $n = 0.25 \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2$,
Rounded up to the next whole number

100(1- α) % Confidence Interval on σ^2 :

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}$$

Based on a normal population

- **100(1- α) % C.I. on σ :**

$$\sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}}$$

Based on a normal population

Confidence Intervals for Two Parameters

100(1- α) % C.I. for the Difference of two Means

- **Independent Samples** with populations' variances known, any sample sizes:

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

(It is a two-sample Z-Interval)

- **Independent Samples** with populations' variances unknown and sample sizes ≥ 30 :

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

(It is a two-sample Z-Interval)

- **Independent Samples** with populations' variances unknown and small sample sizes < 30 :

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}. \text{ (It is a two-sample T-Interval)}$$

(For the T-Interval, Check the text for the degrees of freedom for the case under consideration)

- **Dependent Samples, Matched pairs Data**

$$\bar{d} \pm t_{\alpha/2} \cdot \frac{S_d}{\sqrt{n}}, \text{ with } n-1 \text{ df}$$

100(1- α) % C.I. for the Difference of Two Proportions

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

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

as the margin of error.

- Sample Size (if previous estimates on P_1 and P_2), \hat{p}_1 and \hat{p}_2 are available:

$$n = n_1 = n_2 = \lceil [\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)] \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2 \rceil, \text{ rounded up to the next whole number.}$$

- Sample Size, without previous estimates on P_1 and P_2 :

$$n = n_1 = n_2 = 0.25 \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2,$$

Rounded up to the next whole number.

100(1- α) % C.I. on the Ratio of Two Variances

$$\frac{1}{F(\alpha/2, r_1, r_2)} \frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} F(\alpha/2, r_2, r_1)$$

Hypothesis Testing Regarding One Parameter

Testing on the Population Mean μ

- The Test Statistic to be used is:

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

based on a normal population, with known variance. When $n \geq 30$, and σ^2 is not known, S is replacing σ in the above test.

- The Test Statistic to be used is:

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

based on a normal population, with unknown variance, $n < 30$, and with $n-1$ degrees of freedom

Click here to learn more

TAKE THE
RIGHT TRACK

Give your career a head start
by studying with us. Experience the advantages
of our collaboration with major companies like
ABB, Volvo and Ericsson!

Apply by
15 January

World class
research

www.mdh.se

MÄLARDALEN UNIVERSITY
SWEDEN

Testing on the Population Proportion

- The Test Statistic to be used is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

based on a normal population and $n < 5\%N$.

Testing on the Population Variance

- The Test Statistic to be used is:

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

to follow the Chi-Square distribution with $n-1$ degrees of freedom with a normal population

Hypothesis Testing Regarding Two Parameters

- On the Difference between Two Means:** (independent sampling, known variances) Test Statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- On the Difference between Two Means:** (independent sampling, unknown variances, and large samples) Test Statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- On the Difference between Two Means:** (independent sampling, unknown variances, small samples, *No pooling*) Test Statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

(For the above T-Test, check the text for the degrees of freedom)

- **On the Difference between Two Means:** (independent sampling, unknown variances, small samples, *with pooling*) Test Statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

$$\text{and } S_p = \sqrt{S_p^2}$$

is the pooled common standard deviation of the two populations, with df $\nu = n_1 + n_2 - 2$

- **On the Difference between Two Means:** (dependent sampling, unknown variances, & small samples) Test Statistic is $t = \frac{\bar{d}}{S_d / \sqrt{n}}$,
where \bar{d} is the mean and S_d is the standard deviation of the differenced data

- **On the Difference between Two Proportions:**

Test Statistic is:

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

$$\text{where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- **On the Ratio between Two Variances:** Test Statistic is:

$$F = \frac{S_1^2}{S_2^2}$$

with $n_1 - 1$ and $n_2 - 1$ degrees of freedom for the numerator and denominator respectively.

- Finding a Critical F for the left Tail

$$F_{1-\alpha, n_2-1, n_1-1} = \frac{1}{F_{\alpha, n_1-1, n_2-1}}$$

Further Inference Methods

Goodness-of-Fit Test: Chi-Square Test

$$\chi^2 = \sum_1^k \frac{(O_i - E_i)^2}{E_i}$$

Where O_i are the Observed Data, and E_i are the Expected values, K is No. of Classes.

Chi-Square Test for a Contingency Table

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where O_{ij} are the Observed Data, and E_{ij} are the Expected values; in each cell.

$$E(\hat{n}_{ij}) = \frac{r_i \cdot c_j}{n}$$

where r_i and c_j are the totals of the i^{th} row and j^{th} column, respectively.



How will people travel in the future, and how will goods be transported? What resources will we use, and how many will we need? The passenger and freight traffic sector is developing rapidly, and we provide the impetus for innovation and movement. We develop components and systems for internal combustion engines that operate more cleanly and more efficiently than ever before. We are also pushing forward technologies that are bringing hybrid vehicles and alternative drives into a new dimension – for private, corporate, and public use. The challenges are great. We deliver the solutions and offer challenging jobs.

www.schaeffler.com/careers

SCHAEFFLER



ANOVA

One-way-ANOVA

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \text{ Versus}$$

$$H_1: \exists i, j; 1 \leq i, j \leq k; \mu_i \neq \mu_j.$$

ANOVA Setup Table

Treatment	1	2	...	k
	Y_{11}	Y_{12}	...	Y_{1k}
	Y_{21}	Y_{22}	...	Y_{2k}

	$Y_{n1,1}$	$Y_{n2,2}$...	$Y_{nj,k}$
Totals	$Y_{.1}$	$Y_{.2}$...	$Y_{.k}$
Means	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$...	$\bar{Y}_{.k}$
Variiances	$S_{.1}^2$	$S_{.2}^2$...	$S_{.k}^2$

Two-way-ANOVA Sum of Squares

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (X_{ijk} - \bar{X}_{...})^2,$$

$$SS_{ER} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (X_{ijk} - \bar{X}_{ij.})^2,$$

$$SS_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{X}_{i..} - \bar{X}_{...})^2,$$

$$SS_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{X}_{.j.} - \bar{X}_{...})^2, \text{ and}$$

$$SS_{A,B} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2.$$

Therefore, we can write:

$$SS_{TO} = SS_B + SS_A + SS_{A,B} + SS_{ER}.$$

APPENDIX D

REFERENCES

1. Bakir, S. T. and M.A. Shayib, 2013, Applied Statistical Methods, WWW.alibris.com.
2. Conover, J.W., 1999, Practical Nonparametric Statistics, 3rd Edition, Wiley, INC, NY.
3. Fisher, R.A., 1925, "[Applications of "Student's" Distribution](#)". *Metron* **5**: 90–104.
4. Graybill, F.A., 1983, Matrices with Applications in Statistics, Wadsworth International Group.
5. Hogg, R.V. and Tanis, E.A., 2010, Probability and Statistical Inference, Pearson.
6. Johnson, N.L. and Kotz, S., 1969–1972, Distributions in Statistics. (4 Vols.) Wiley, New York.
7. Kinney, J.J. 2014, Probability: An Introduction with Statistical Applications, 2nd Edition, Wiley.
8. Kirk, R.E., 1995. Experimental design: Procedures for the behavioral sciences. Pacific Grove, CA:
9. Montgomery, D.C. (2001) Design and Analysis of Experiments. John Wiley and Sons, New York.
10. Myers, R.H., 1990, Classical and Modern Regression with Applications, 2nd Edition, PWS-KENT Publishing Company, Boston, MA.
11. Pfanzagl, J.; Sheynin, O., 1996, "A forerunner of the t -distribution (Studies in the history of probability and statistics XLIV)". *Biometrika* **83** (4): 891–898.
12. Shayib, M.A., 2013, Applied Statistics; <http://bookboon.com/en/applied-statistics-ebook>
13. Shayib, M.A. 2005, Effects of Parameters and Sample Size on $P(Y < X)$, ASA, JSM Proceedings, Minneapolis, MN, 142–144.
14. Shayib, M.A. and Aly, E.E., 1992, On Some Goodness-of-Fit Tests for the Normal, the Logistic and the Extreme-value Distributions, *Commun. Statist. – Theory and Methods*, 21(5), 1297–1308.
15. Shayib, M.A. and Awad, A.M., 1990, Prediction Intervals for the Difference Between Two Sample Means from the Exponential Population, A Bayesian Treatment, *Pakistan J. Statist.*, 6(1), 1–23, 1990.
16. Shayib, M.A. and Young, D.H., 1989, Modified Goodness-of-Fit Test in Gamma Regression, *Statist. Comput. Simul.* Vol. 33, pp. 125–133.
17. Student, W. S. Gosset, (March 1908). "The probable error of a mean". [Biometrika](#) **6** (1): 1–25.
18. Whittaker, E.T. and Watson, G.N., 1965, A Course in Modern Analysis, Cambridge University Press, 6th Edition.

To see Part II download:
Inferential Statistics – The Basics for Biostatistics: Part II