Applied Statistics

Mohammed A. Shayib

Mohammed A. Shayib

Applied Statistics

Applied Statistics
1st edition

ISBN 978-87-403-0493-0

Contents

Preface

It is already encrypted on the American currency "IN GOD WE TRUST". This was the first part of a statement that I heard from the Quality Manager at the Lubbock, TX, site of Texas Instrument while I was there (1992–1998) as a quality engineer and Statistical Process Control Trainer. The second part says "EVERYTHING ELSE NEEDS DATA". It is the 21st century and the technology age, and data is abandunt 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 on Applied Statistics.

Applied Statistics is a compendium, an elementary introduction to the growing field of statistics. In this consize 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.

The material had been presented in such a way that only College Algebra can be a prerequisite for the course that covers the whole text in one semester. The material is presented in 6 chapters.

Chapter 1 is about collecting, and organizing qualitative and quantitative data, as well as summarizing the data graphically or numericall regardless if the data were discrete or continuous.

Chapter 2 introduces the notion of probability, its axioms, its rules, and applications. In addition to that, Chapter 2 contains material on probability distributions and their characteristics for discrete and continuous cases.

Chapter 3 covers the first main part of inferential statistics; namely estimation in its two branches: point and interval estimation by introducing the sample statistics as estimators for the population parameters.

Chapter 4 is concerned about the second part of inferential statistics, namely hypothesis testing about one parameter of a population, or two parameters of two populations. In this chapter there is an outline, and procedure on how to implement the steps in hypothsis testing when using the two methods; the classical or the traditional method and the p-value method. The detailed exposure of the two methods will enable the reader, and the student to reach a comprehensive and sound conclusion by using information found in data.

Up till chapter 5, the discussions were with data based on one variable aspect of the population that of interest. **Chapter 5** will introduce how to deal with two related variables and find a simple linear relationship, if it exists, between them. In other words, chapter 5 is restricted to simple linear regression between an explanatory variable and a response variable related to it. Moeover, the correlation concept and definition are introduced in order to measure the strength of that liear relationship which was found earlier.

Chapter 6 deals with another route of checking on data of one variable, or factor, and on the independence or dependence between two factors. Chapter 6 goes on to introduce other tests in statistics and to check on tests that make a decision whether more than two means are the same or not. Chapter 6 introduces the One-Way Analysis of Variance, or ANOVA.

The treatise is wrapped up with an appendix that has 8 tables for use when reading the textbook.

Mohammed A. Shayib
Friday, May 31, 2013

1 Descriptive Statistics

Outline

1.1 Introduction

The subject of statistics may be presented at various levels of mathematical difficulty, and it may be directed toward applications in various fields of inquiry. In this sequel we will keep the mathematical background to a minimal. Now the question: What is statistics can be raised.

Statistics is that branch of science which deals with:

1. Collection of data,
2. Organizing and summarizing the data,
3. Analysis of data, and
4. Making inferences, or decisions and predictions.

Step 4 is the objective of statistics, i.e., making inferences about a population based on information contained in a representative sample taken from that population.

We have some concepts that need to be well defined and established. These concepts are: A **population** is a set of objects called elements that share a certain property, and it is the entire group to be studied. For our purposes here, most populations will be sets of numbers that are of interest to us. For instance, if we talk about students, the population could be all the students you can think of.

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.

❖

Since we cannot, and sometimes it is impossible to deal with all the elements in a population, a smaller and representative part, or a subset, of that population is considered. Thus that representing subset of the population is called **a sample**.

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.

Levels of Measurements Data may be classified in 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) ourse grades in any course: A, B, C, D, F, P , W, I
d) Student gender: male, female

Solution

a) The years 1776 to 1876 represent interval data. There is no natural zero (No "year zero"). 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 income represents 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.

❖

Data will be collected on individuals from the population and termed as variables. Thus variables are 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.**

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.

❖

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, and variables, as qualitative or quantitative; quantitative variables and data as well, can be further classified as: **Discrete or Continuous.**

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.

❖ ----------

A Continuous Variable is a quantitative variable that has an infinite 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.

1.2 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:

1. **Graphically or**
2. **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 date include:

1. **The Bar graph,**
2. **The Pie Chart, and**
3. **The Pareto Chart.**

For describing a quantitative data graphically we use:

1. **the Dot plot,**
2. **The stem and leaf display, and**
3. **The histogram.**

In the subsequent sections we will discuss all those methods of presenting the data graphically.

1.3 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

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

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

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:

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 frequency 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, 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 column. The other way of modifying a frequency distribution is presenting it as a cumulative relative frequency distribution by adding the relative frequencies as we go down the classes, and this will generate the **Ogive** line, **see Figure 5C.**

1.4 Graphical Presentation

1.4.1 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 = max – 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**).

On the 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.**

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**).

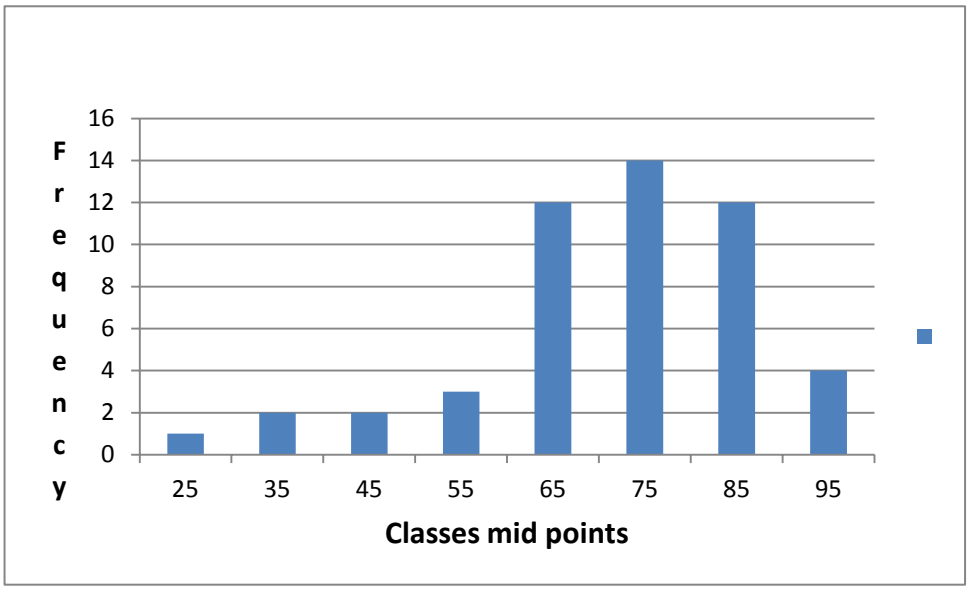


Figure 1A

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 that data had been given in, or classified based upon. 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

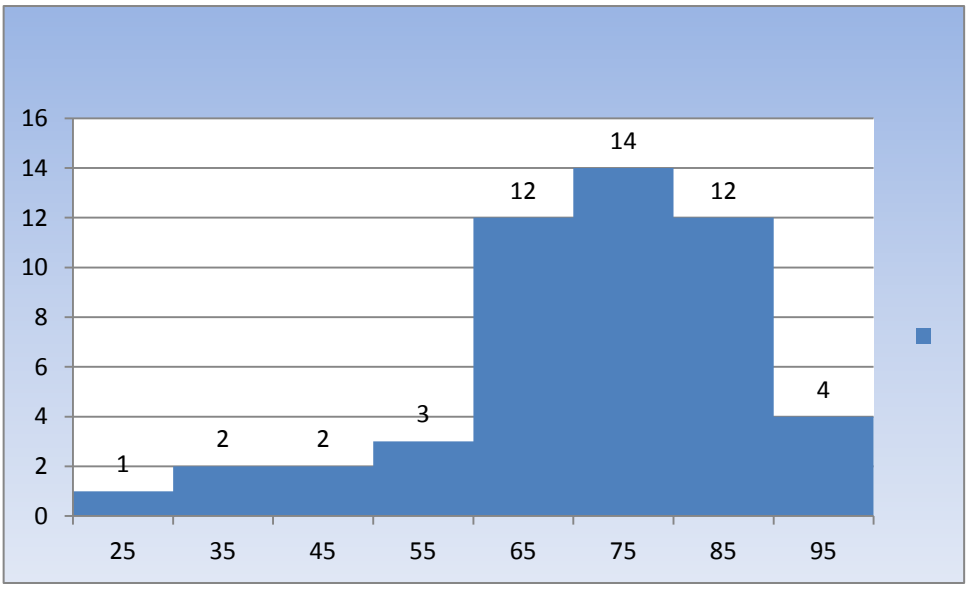


Figure 1B

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. 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 < Pulse Rate ≤ 41, and (41–48] is for the range of values given by 48 < Pulse rate ≤ 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	Frequen
(34 – 41]	2
(41- 48]	2
(48 – 55]	4
(55 – 62]	19
(62 – 69]	40
(69 – 76]	53
(76 -83]	30
(83 – 90]	27
(90 – 97]	10
(97 – 104]	5

Table 3

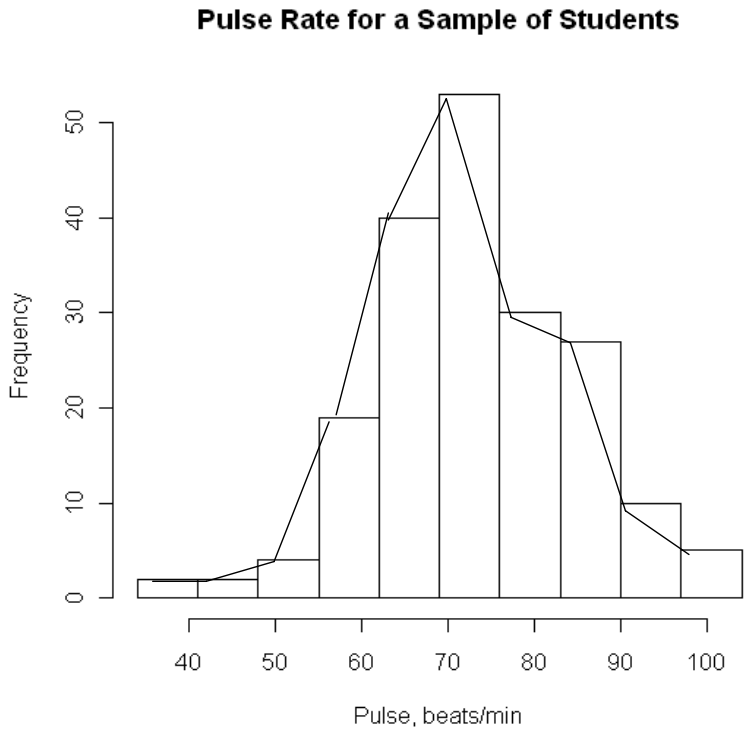


Figure 2

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, in **Table 4**, 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 units digit. Thus we have the following stem-and-leaf diagram

Stem	Leaf	freq
1	1 1 2 3 3 5 5 6 7 7 7 8 8 8 8 9 9	17
2	0 1 2 2 3 3 4 4 4 6 7 7 9 9	14
3	0 1 2 2 6 7 8	7
4	6 8	2

Legend: 3|2 --→32

Figure 3

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	freq
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

❖

1.4.2 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 used to classify the data. Frequency, cumulative frequency, relative frequency distributions can be drawn to present the data graphically. A graph called the **Ogive**, 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

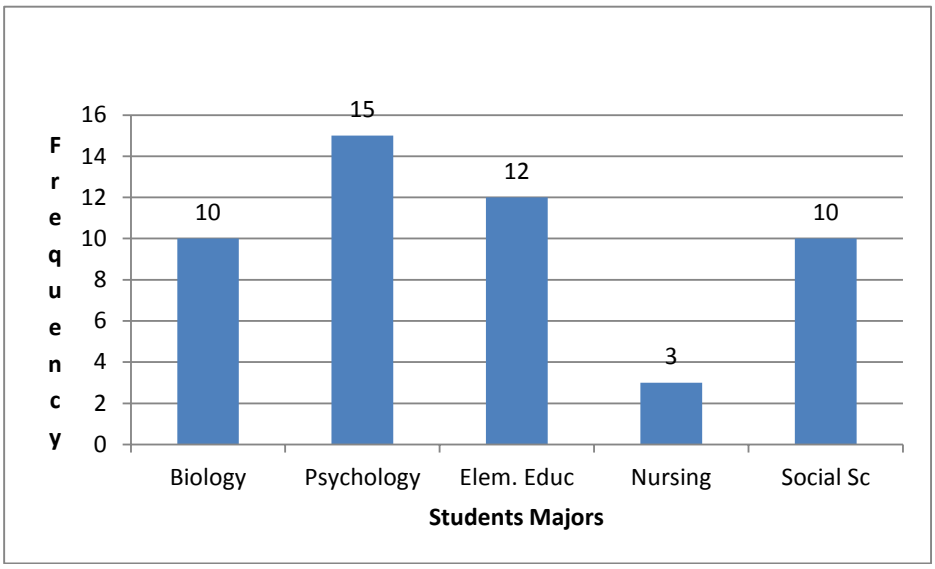


Figure 5A Frequency Distribution

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 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.

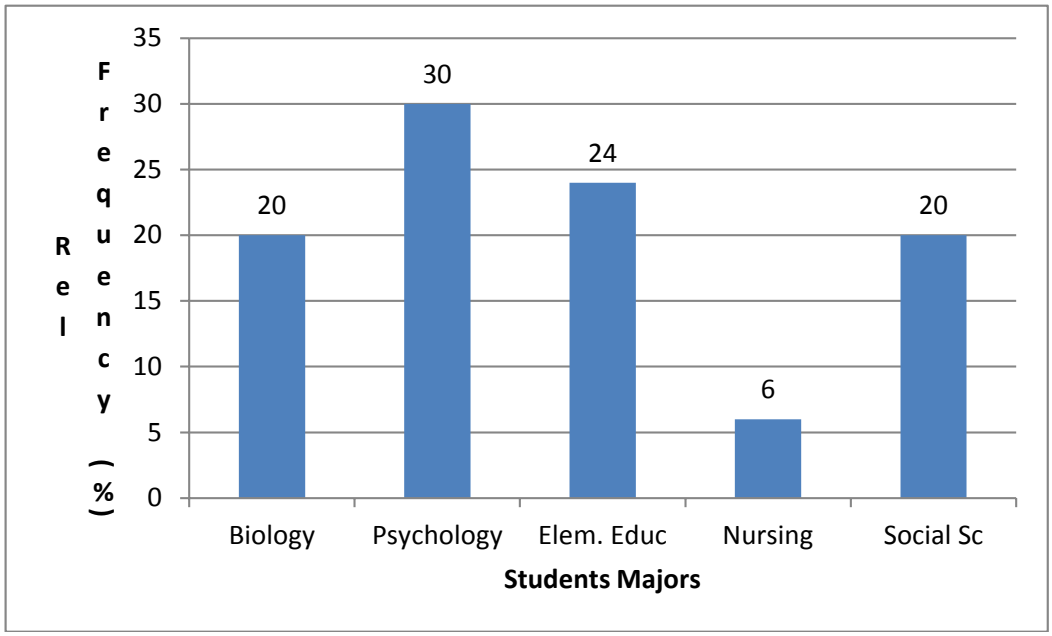


Figure 5B Relative Frequency Distribution

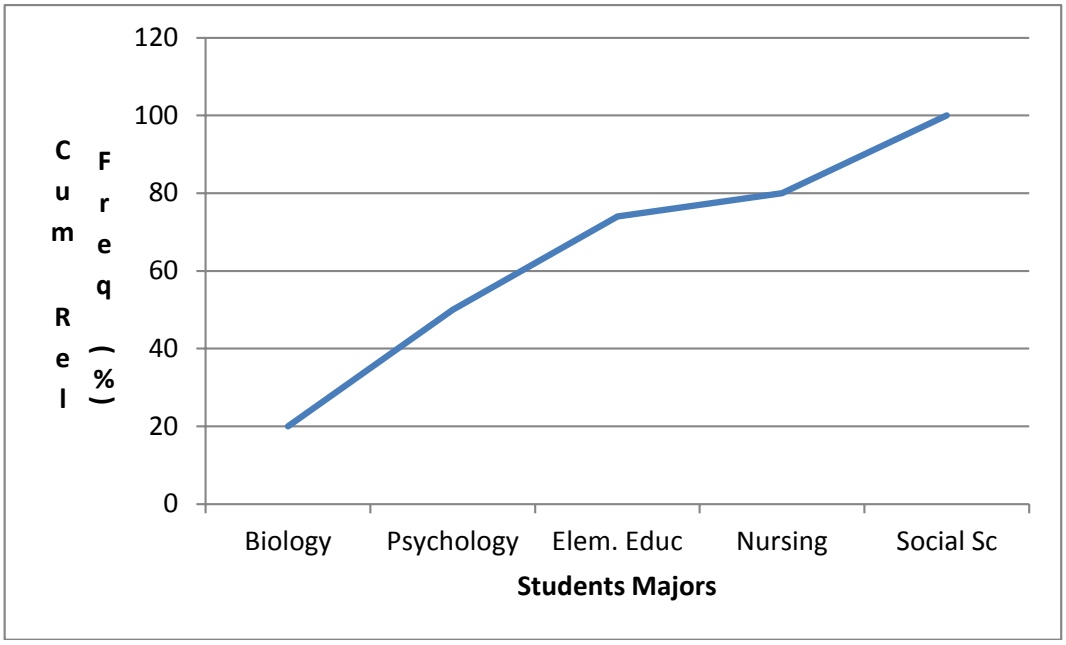


Figure 5C Cumulative Relative Frequency Distribution (Ogive)

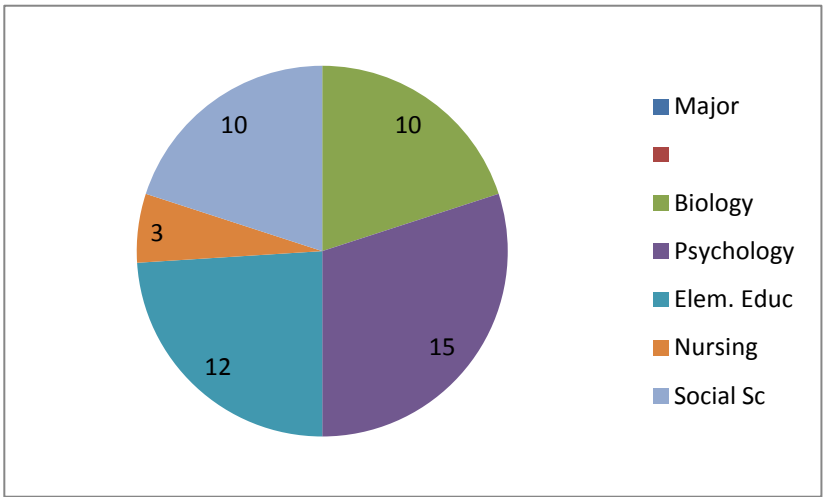


Figure 6 Pie Chart for the Majors in the Class

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**

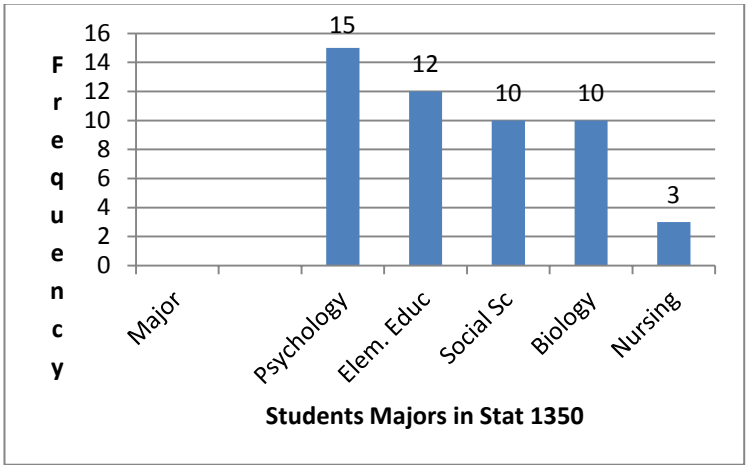


Figure 7 Pareto Chart for Enrollemt by major in Stat 1350

As mentioned above this tool is a quality-tool and it can list the sources of scrap at a factory, so the quality engineer will know where is the major source for scrap and got on 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 make more money then.

1.5 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

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

❖

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.

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

2. The sum of a constatnt times a variable (cX_i) is equal to the constant times the sum of that variable.

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

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

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

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

Some common misconceptions are considered when they should be confused.

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

Let us give an example.

EXAMPLE 1.13

Consider the data in Example 1.10, 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. From Example 1.10 we have

$$\sum_{i=1}^{3} X_i = X1 + X_2 + X_3 = 2+5+6 = 13.$$

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

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

❖ ----------

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 $\sum_{i=1}^{3} X_i^2 \neq \left(\sum_{i=1}^{3} X_i\right)^2$.

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

Clearly the left hand side above can be expressed as

$$\sum_{i=1}^{n} X_i Y_i = X_1Y_1 + X_2Y_2 + \ldots + X_nY_n,$$

While the right hand side is given by

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

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:

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

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

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

❖

1.6 Numerical Methods for Summarizing Quantitative Data

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

1. **Measures of Central Tendency (Measures of Center)**
2. **Measures of Variation (Measures of Dispersion)**
3. **Measures of Position, and**
4. **Measure of Quality and Outliers.**

1.6.1 Measures of Central Tendency

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

1. **The mean**, or the arithmetic average,
2. **The mode**, i.e., the most frequent observation(s),
3. **The median**, and
4. **The mid-range.**

Let us give the definition for each of those measures.

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 $x_1, x_2, ..., x_n$, the sample mean, $\bar{x}$, is calculated as:

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

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

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

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 frequently called and used, is a simple and familiar measure, the following are some of its noteworthy properties:

1. It can be calculated for any set of data, so it always exists.
2. A data set of numerical values has one and only one mean. Thus it is unique.
3. It lends itself to further statistical treatment, as we will see later.
4. 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.
5. It takes into account every item of the data set.
6. 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:
The Geometric Mean, $\bar{G}$ Gbar, which is given by

$$\bar{G} = (x_1.x_2 \ldots . x_n)^{1/n}, \text{ or}$$

$$\bar{G} = (\prod_{1}^{n} X_i)^{1/n}.$$

The Harmonic mean, Hbar $\overline{H}$

$$\overline{H} = = n / \sum_{1}^{n} (1/X_i).$$

EXAMPLE 1.15

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 μ = 16.538.

EXAMPLE 1.16

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

a) The geometric mean,
b) The harmonic mean.
c) Compare the above three means: $\bar{x}$, $\overline{G}$, and $\overline{H}$.

Solution:

a) The geometric mean is given by $\overline{G} = (x_1, x_2 \ldots. x_n)^{1/n}$, 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, $\overline{G}$ we see that with a graphing calculator that $\overline{G} = (5*8*12*15*20)^{1/5} = (144000)^{1/5} = 10.7565$.

b) The Harmonic mean is given by $\overline{H} = = n / \sum_{1}^{n} (1/X_i)$. From the data, and by using a graphing calculator we find that $\overline{H}$ = 5/[1/5+1/8+1/12+1/15+1/20] = 9.524.

c) 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 $\overline{H} < \overline{G} < \bar{x}$.

❖

2. 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 measure. 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 1.17

a) Find the mode for the data in Example 1.15
b) Consider the following Data as presented in classes

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

Solution:

a) The data in Example 1.15 has no mode. Each data point has appeared once.
b) 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.

❖

3. **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 $[(n+1)/2]^{th}$ observation. When n is even, there are two middle values in the array, namely $[n/2]^{th}$ and $[n/2 +1]^{th}$ observations. The median in this case is the average of those observations.

EXAMPLE 1.18

What is the median for the data in Example 1.15?

Solution:

Let us find the median for the data in Example 1.15. 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.

❖

4. 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., Midrange = (Max + Min)/2.

EXAMPLE 1.19

What is the Mid-Range of the data in Example 1.15?

Solution:

Looking at the array in Example 1.15, we see that the minimum is 7 and the maximum is 34. Therefore the mid range is: MR = (7+34)/2 = 20.5.

❖

1.6.2 The Measures of Variation

In section 1.6.1 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,

$$\text{Range} = R = \text{Largest data value} - \text{smallest data value}$$
$$= \text{Maximum} - \text{minimum}.$$

EXAMPLE 1.20

What is the range for the data in Example 1.15?

Solution:

From the array in Example 1.15, 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

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 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 a sample on our hands, the deviation from the mean of the sample $\overline{X}$ can be found as $Y_i - \overline{X}$. Moreover, based on the definition of the sample mean above, we can see that "The sum of the deviations of the sample data points from their mean is also zero", (see exercise 1.8).

We now turn to how to compute the variance and the standard deviation of a finite population. The population variance σ^2 is defined as the average of all the squared deviations of the observations about their mean, i.e.

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

Thus, the standard deviation, σ will be given by

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

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.

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

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

Thus the standard d4viation will be given by the positive square root of the variance by

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

EXAMPLE 1.21

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 $\mu = (2+4+6+8+10)/5 = 6 = \overline{X}$, 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 $S = \sqrt{10}$, and 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. These formulas for the population are as follows:

$$^2 = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}, \text{ with } \quad = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}}.$$

Similarly, for the sample we have

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}, \text{ with } S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}}.$$

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 below Figure 3), 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 rule of thumb states that

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

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

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

(Check the Empirical rule below).

1.7 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,…, n be the original data, and $X_i = Y_i + b$, where b is a constant. It is easily seen that $\overline{X} = \overline{Y} + b$. 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.
2. If data were multiplied by a constant then the mean of the new data will be $\overline{X} = b\overline{Y}$. Also in this case, the mode, the median and the mid-range will change. Each will be multiplied by that constant.
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, See exercise 1.23.
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. See Exercise 1.23.

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 $\overline{X}$ mean replacing μ and the sample standard deviation s replacing σ in the above inequalities. We thus have the Figure 6, 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 $(1\text{-}1/k^2)100\%$ of the data will lie between μ - kσ and μ+ kσ. We can also use Tchebyshev's Inequality based on a sample data.

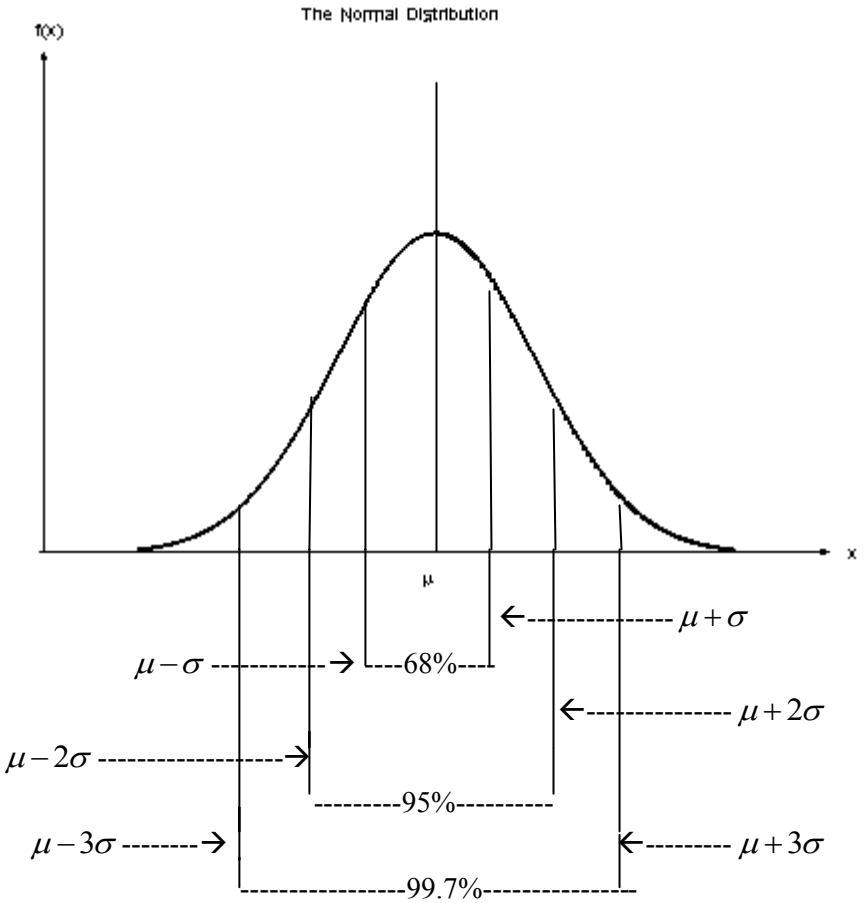


Figure 8

1.8 Other Measures for Quantitative Data

In section 1.6 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-score,**
2. **The Percentiles,**
3. **The Deciles, and**
4. **The Quartiles.**

The Z-score: It represents the distance that a data point is 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-score. There are two Z-scores, a Z-score for the population and another one for the sample. Their formulas are respectively given by

Population Z- Score: $Z = \frac{x -}{}$

Sample Z- score: $Z = \frac{\left(Y - \bar{Y}\right)}{S}$

The Z-score has mean 0 and standard deviation 1.

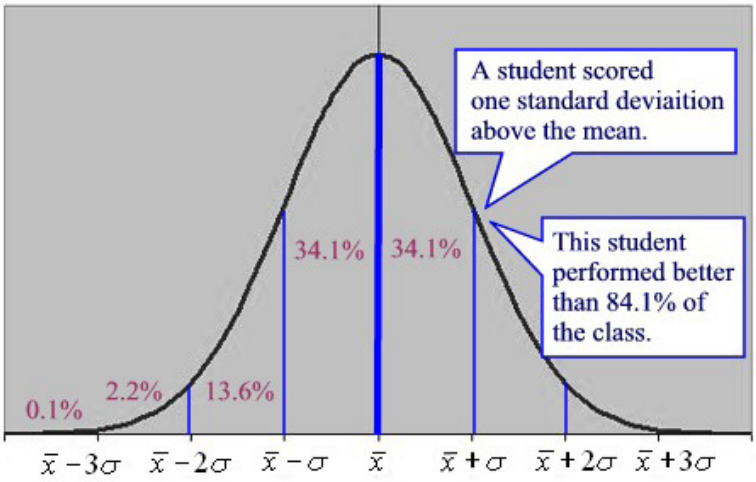


Figure 9
(Source: Internet, Normal curve images)

Figure 9, above, shows the grades of students on the z-Scale. **Figure 10** displays the Empirical rule using Z-scores.

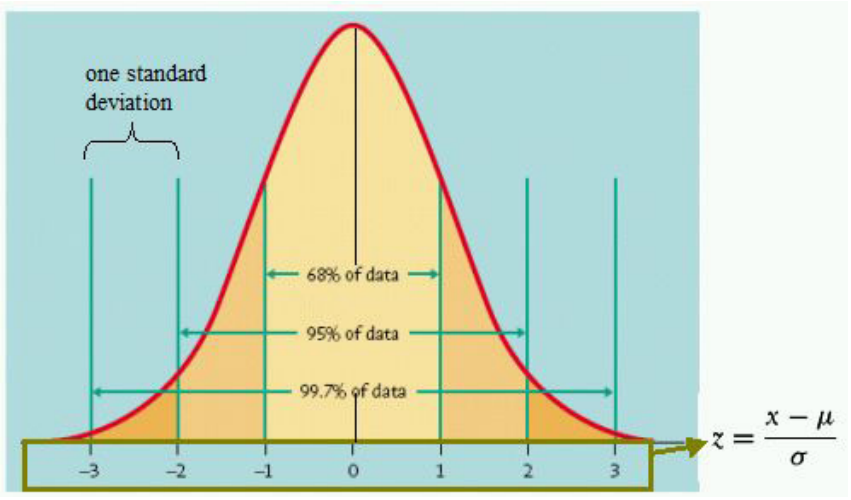
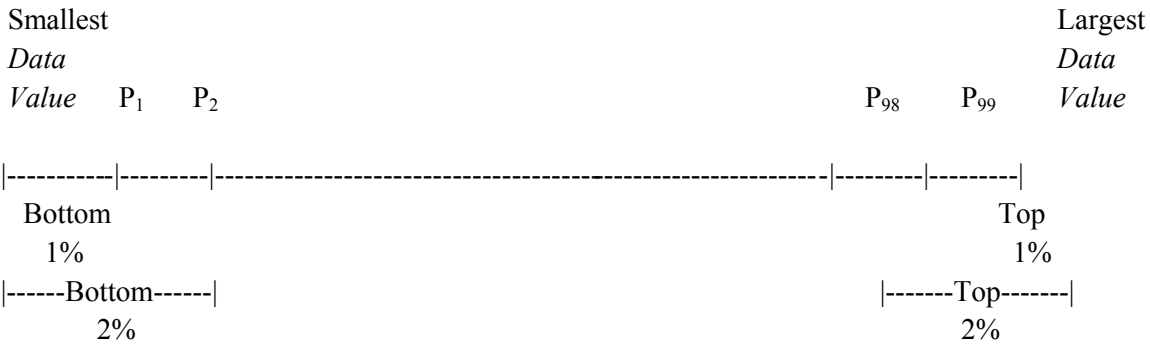


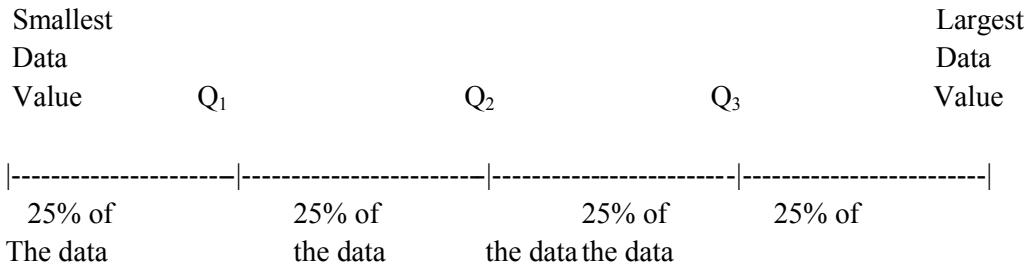
Figure 10
(Source: Internet, Normal curve images)

The **K**th **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 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 is the median and it is equal to P50, the 50th percentile.

The most common used percentiles are the **Quartiles.** Quartiles divide the 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 extreme values. Quartiles, on the other hand, are resistant to extreme values. Because of this property, we can define a measure of dispersion that is based on quartiles, namely the Interquartile Range, **IQR.** The Interquartile Range, **IQR**, is the range of the middle 50% of the observations in the 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

$$\mathbf{IQR} = \mathbf{Q}_3 - \mathbf{Q}_1.$$

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

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

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

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 $\overline{Y}$ 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 of the sample. Our hope, of course, is to use computational procedures for the sample characteristics (e.g. $\overline{Y}$ 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.

> **A Parameter**: is a characteristic of a population set of observations; e.g. N, μ, σ and σ^2.
> **A Statistic:** is a characteristic of a sample of observations, e.g., n, $\overline{Y}$, S, and S^2.

1.9 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.

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. q. r……., different ways.

EXAMPLE: 1.22

> 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*24*23 = 13800 different committee structures.

❖

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! = 3.2.1 = 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! (n factorial).

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

$${}_nP_r = n!/(n-r)!.$$

EXAMPLE 1.23

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 $5! \cdot 4! = 2880$

❖

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!) (1!)] = 3 distinct permutations. Hence we have the following Rule:

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

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

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

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) object that are left. The number of such combinations is denoted, for short, by $\binom{n}{r}$. The number of combinations of n distinct objects taken r at a time is given by

$$\text{nCr} = \binom{n}{r} = \frac{n!}{(n-r)! \cdot r!}.$$

EXAMPLE 1.24

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.q = (6) (3) = 18 committees.

❖

1.10 Description of Grouped Data

It is often that data is already summarized in a frequency table. When it is given in terms of classes' limits and 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 the mean of the data values, in that class, is equal to the class midpoint. We then multiply the class midpoint by 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 are given by

Population mean is $= \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i}$.

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:

Population variance is $^2 = \frac{\sum_{i=1}^{n} (X_i - \ )^2 f_i}{\sum_{i=1}^{n} f_i}$.

Sample variance is $S^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2 f_i}{(\sum_{i=1}^{n} f_i) - 1}$.

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

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

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

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

EXAMPLE 1.25

Let us consider the data in Example 1.1, and look at it in a different way. Consider the following as the mid points off the classes in Example 1.1

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

$$\bar{x} = \frac{\sum_{i=1}^{8} X_i f_i}{\sum_{i=1}^{8} f_i} = [1.(25) + 2.(35) + 2.(45) + 3.(55) + 12.(65) + 14.(75) + 12.(85) + 4.(95)]/50$$

$$= 3580/50 = 71.6$$

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

$$= 247.3878.$$

Hence S = 15.7286.

❖

EXAMPLE 1.26

Let us have another look at the data in Example 1.1. 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:

1-var Stats

$\overline{X}$ = 71.26

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

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

S_x =14.90214339

σ_x = 14.7523693

n = 50

Minx = 29

Q_1 = 63

Med = 73.5

Q_3 = 82

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

S^2 = (14.90214339)2 = 222.0738776

σ^2 = (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.

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.

Range = 97 – 29 = 68

IQR = 82 – 63 = 19

Mode = 63.

It appeared 3 times more than any other data point.

Lower Fence = Q_1 – 1.5*IQR = 63 – 1.5*19 = 34.5

Upper Fence = Q_3 + 1.5*IQR = 82 + 1.5*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: Min, Q_1, Med, Q_3, and Max, displayed on the box plot.

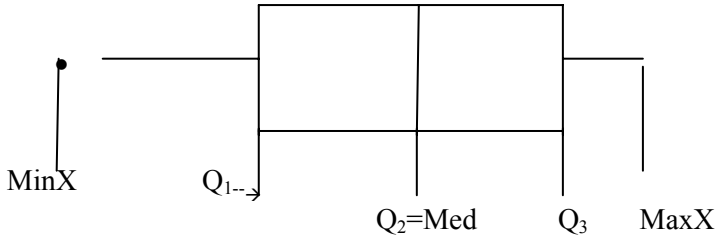
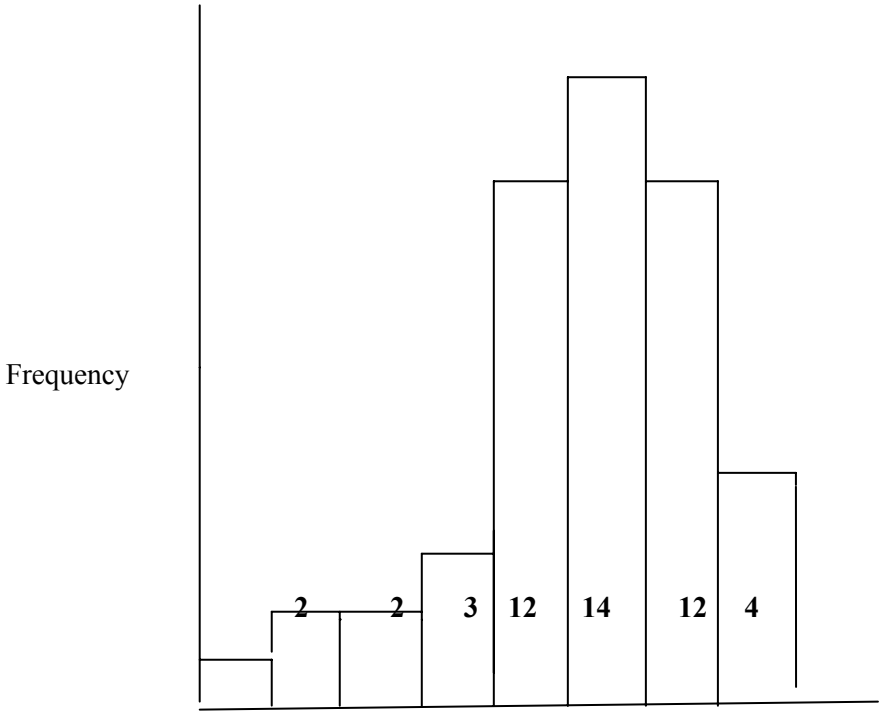


Figure 11 BOXPLOT

❖

The histogram for the data, in Table 2, 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.



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 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										

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

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

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

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

a) $x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4 + x_5 f_5 + x_6 f_6$
b) $y_1^2 + y_2^2 + y_3^2 + y_4^2$
c) $(z_2 + 3) + (z_3 + 3) + (z_4 + 3) + (z_5 + 3) + (z_6 + 3)$

1.6 Given: X_1=2, X_2=3, X_3=4, X_4= -2;Y1= 5,Y_2= -3,Y_3= 2, and Y_4=-1. Find:

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

a) $\sum_{i=1}^{3} Xij$ separately for j = 1, 2, 3, and 4.

b) $\sum_{j=1}^{4} Xij$ separately for I = 1, 2, 3, and 4.

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

1.9 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.

1.10 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

1.11 See Exercise 1.1, Find: a) The mean b) The median c) The Modal Class.

1.12 See Exercise 1.2, Find: a) The mean b) The median c) The Modal Class.

1.13 Consider the following sample: 2, 4, 7, 8, and 9, find the following:

a) The mean b) The Mode c) The Midrange

1.14 Consider the following sample: 7, 6, 10, 7, 5, 9, 3, 7, 5, and 13, find the following:

a) The mean b) The Mode c) The Midrange

1.15 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

1.16 Consider the sample in exercise 1.13. Find:

a) The Range b) The Variance c) The Standard Deviation d) The Coefficient of Variation

1.17 Compute the Coefficient of Variation for an exercise of your choice. (Pick an exercise that has data)

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

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

a) 20! = 20.19.18.17! b) 3! + 4! = 7! c) 4! .3! = 12! d) 16! =17! /17

e) 1/2! + 1/2! =1 f) 9! / (7! .2!) = 72 g) 4! + 0! = 25

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

1.21 Find the mean of the grouped data in Exercise 1.1.

1.22 Find the mean of the grouped data in Exercise 1.2.

1.23 Prove each of the properties listed in section 1.7.

1.24 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?

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

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	3358
6	12

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

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

1.27 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.

1.28 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.

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

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.

1.30 Nation of origin.

1.31 Number of friends

1.32 Eye color

1.33 Grams of sugar in a meal

1.34 Number of left turns you made while driving home today.

1.35 The value of your car

1.36 Your phone number

1.37 Your student ID on any campus

Classify the quantitative variable in the following exercises, as discrete or continuous.

1.38 The distance from your house to school

1.39 The time to run a marathon

1.40 The number of questions you get wrong on a multiple choice exam

1.41 The number of seats in a classroom

1.42 The time you take to finish a pop quiz

1.43 The amount of gas in the tank of your car

1.44 The number of cylinders in the engine of your car

Indentify the type of sampling used in each of the following exercises.

1.45 To determine customer opinion and satisfaction of their boarding policy, an airline company randomly selects 60 flights during a certain week and surveys all passengers on the flights.

1.46 A radio station asks its listeners to call in their opinion regarding the advertising time on the station.

1.47 To estimate the percentage of defects in a recently made batch, a quality-control manager selects every 8th chip that comes off the assembly line starting with the 3rd until he obtains a sample of 140 chips.

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".
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 2^{nd} Y = to access Stat-Plots menu. Select 1: plot1.
3. Place the curser on "ON" and press ENTER.
4. Place the cursor on the histogram icon (check your calculator) and press ENTER. Press 2^{nd} 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

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 curser 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.

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 2^{nd} 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 2^{nd} then 1 to insert L1 on the **HOME** screen. Then press the comma and press 2^{nd} 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.

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 2^{nd} Y = to access Stat-Plots menu. Select 1: plot 1.
3. Place the curser 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 indentify 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.

2 Random Variables and Probability Distributions

Outline

2.1 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 find the following 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 had 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 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. The next two sections 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 and random variables. 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.

2.2 Probability

We start this section with some definitions.

An Experiment is a process by which an observation (or a measurement) is obtained.
A sample Space, S, for an experiment, is the set of all possible outcomes of that experiment.

EXAMPLE 2.1:

Determine the sample space of the following experiments:

1. Tossing a normal coin is an experiment.
2. Taking a test, as a student in any course, is an experiment.

Solution:

1. The Sample Space, S, is given by S = {Head, Tail} = {H, T}
2. The sample space, S, consists of the following grades, S = {A, B, C, D, F}.

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.

EXAMPLE 2.2:

A probability experiment consists of rolling a single fair die.

1. Identify the outcomes of this experiment.
2. Determine the sample space.
3. Define the event E = "roll an even number."

Solution:

1. The outcomes from rolling a single fair die are x_1 = "rolling a one" = {1}, x_2 = "rolling a two" = {2}, x_3 ="rolling a three" = {3}, x_4 ="rolling a four" = {4}, x_5 ="rolling a five" = {5}, and x_6 = "rolling a six" = {6}.
2. The sample space, S, has six outcomes, as it appeared from part 1. Thus S = {1, 2, 3, 4, 5, 6}.
3. The event E = "roll an even number" = {2, 4, 6}.

❖

EXAMPLE 2.3:

1. In tossing a fair normal coin, in example 2.1, identify the simple events.
2. In rolling a fair die, in Example 2.2, identify the compound event.

Solution:

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

❖

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: $x_1, x_2, \ldots, x_n$ then we can assign the number p_i for the probability of the outcome x_i, and we write P (x_i) = p_i, on the condition that

1. 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 $0 \leq p_i \leq 1$,
2. The sum of the probabilities of all the outcomes, in a probability experiment must be equal to 1. That is if the sample space S is given as S = { $x_1, x_2, \ldots, x_n$ }, then P(S) = $\sum_{1}^{n} p_i = 1$.

A **Probability Model** lists the possible outcomes of the experiment and the associated probability of each outcome displayed in a table, by a graph, or by a formula. The probabilities, in any probability model, should satisfy the above two 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. Hence if E = $\{Y_1, Y_2, \ldots, Y_n\}$, then $p(E) = p(y_1) + P(y_2) + \ldots + P(y_n)$. The void, or 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 the event is 0. If an event is a certainty, then its probability should be 1. By our setting, we see that $P(\phi) = 0$, and $P(S) = 1$. The sample space is a certainty. Any outcome will be there in S, and S, as an event, is always occurring.

2.2.1 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 that probability of an event E which can be approximated by the ratio of the number of times it occurred to the number of times that the experiment has been carried out. Thus

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

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

$$P(E) = \frac{Number\ of\ ways\ that\ E\ occurs}{number\ of\ possible\ outcomes\ in\ the\ experiment} = \frac{m}{n}.$$

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

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

It is to be realized that using the subjective probabilities are completely acceptable and legitimate, and possibly the only method of assigning likelihood to an outcome.

2.3 Operations and Probability calculation on Events

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

The **union** of the two events, or sets, written A ∪ B, is defined to be the set

A ∪ B = {x| x ∈A or x∈B},

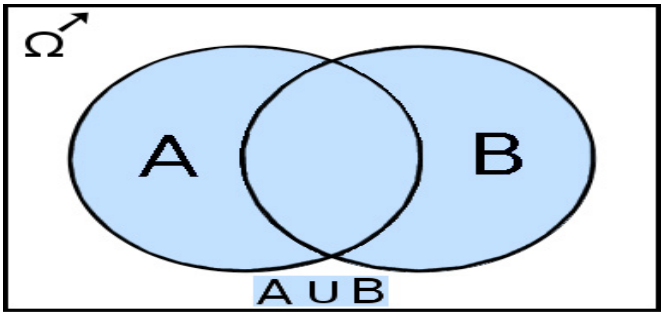


Figure 1

where x ∈ 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 the two sets.

The **intersection** of two events, or sets, written as $A \cap B$, is defined to be the set

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

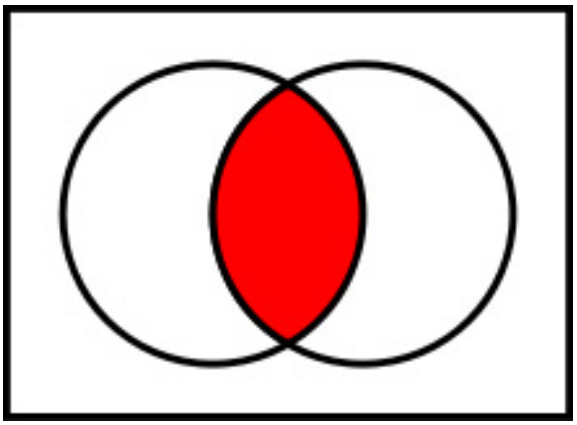


Figure 2

Thus the intersection of two sets is the set of all the elements that are common between the two given sets.

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

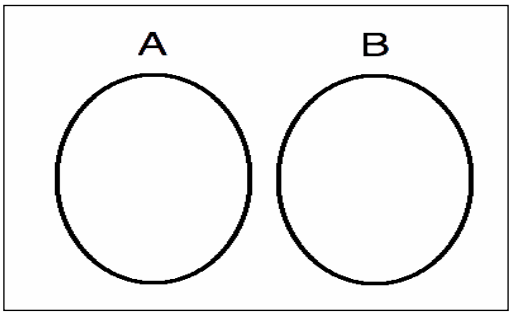


Figure 3

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 in the sample space, See **Figure 3** when A and B have nothing in common) then

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

The general addition Rule states thate

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

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.

The **complement** of a set A, written $\overline{A}$ (or A', or A^c), is defined to be, check **Figure 4** below

$$\overline{A} = \{x | \ x \notin A\},$$

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

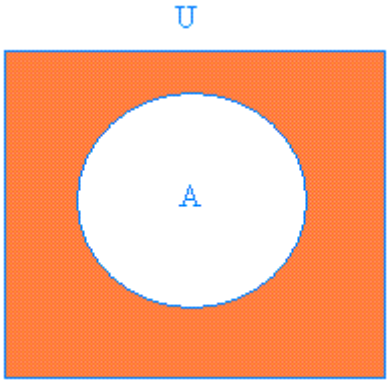


Figure 4

The complement Rule: For any set, or event, E since E ∪ E' = S, and E∪E'=ϕ, then

$$P(E') = 1 - P(E).$$

This is based on P(S) = 1, with the Addition rule we have 1 = P(S) = P (E or E') = P (E) + P (E').

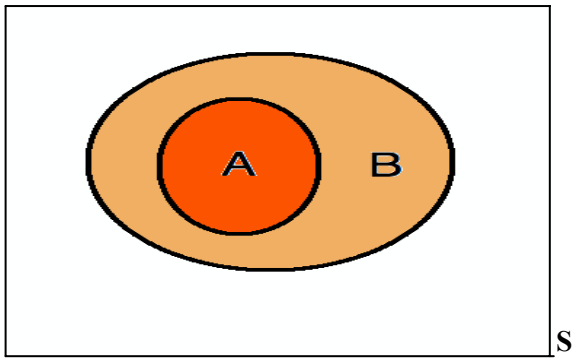


Figure 5

Subsets introduce the notion of membership or contained, 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.

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

EXAMPLE 2.4:

Consider the following sample space S, or as it is called 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∪B, and P (A∪B), ii) A∩B, and P (A∩B), iii) E∩F, and P (E∩F)

iv) E', P (E'), v) P (F∪G).

Solution:

Thus we can see that

1. A∪B = {1, 2, 3, 4, 5, 6, 7, 8, 9}. Hence P (A∪B) = 9/12 =0.75
2. A∩B = {3, 4, 5, 6}, And P (A∩B) = 4/12=1/3.
3. E∩F = {6, 12}, with P (E∩F) = 2/12 = 1/6.
4. While E' = {1, 3, 5, 7, 9, 11}, we have P (E') = 0.5.
5. P (F∪G) = P (F) + P (G) – P (F∩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 and Quotient rules. Really, 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. Each probability rule is related to some kind of events. The addition rule is involved with disjoint or if one will occur than the other. 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)} = ¼, one out of four in the sample space. Remember also, we are tossing a fair coin in such a way that P ({H}) = ½, and ½*1/2 = ¼ that implies P ({HH}) = P ({H})*P ({H}), and hence the product rule:

The product Rule of probabilities: Two events E and F are independent if and only if

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

The above rule can be generalized to more than two events. In the above rule, if neither E nor F is the impossible event, Then E and F cannot be mutually exclusive.

For finding the "quotient" or the general product rule, let us ask the question for a student: 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. It is the probability that event E will occur provided F has occurred already. We conclude with this "quotient" rule:

If A and B are any two events, then

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

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 $P(B) \neq 0$. This is so, since the event has occurred already. It is NOT the impossible event. Hence the general products rule:

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

2.4 Random Variables

The Random Experiment and its Sample Space: A random experiment is any process of measurement or observation, 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 (chance) to occur. Thus tossing a coin once is a random experiment, and the sample space of which is {Head, Tail}. Observing the value of a certain stock in the market is a random experiment, the sample space of which is the set of all possible values of the stock may take. Analyzing a certain chemical to determine the iron content is also a random experiment, Moreover, measuring the height or weight of an object is again a random experiment.

Random Variable: A random variable is any real-valued quantity, or numerical measure, whose value depends on the outcomes of a random experiment. Random variables are typically denoted, or identified, by capital letters such as X. The values that a random variable will take will be denoted by lower case letter x, Thus, in tossing a coin once, the number of heads, X, that may appear is a random variable which may take the values x = 0 or x = 1. In observing a certain stock in the market, the value X of the stock is a random variable which may take any value between \$1 and \$10, i.e. $1 \leq x \leq 10$. In measuring the weight of a person, the weight X is a random variable which may take any value between 105.45 and 300.15 lbs, say, i.e., $105.45 \leq x \leq 300.15$. Based on what had been explained, we see that there are two types of random variables. So, in what is coming next, we like to introduce those types.

A Discrete Random Variable: is that random variable which has either a finite or countable number of values. The values of a discrete Random variable can be plotted on the number line with spaces between them. For example; the number of cousins you have, or the number of friends you know.

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. For example, the distance you drive to work or to school every day.

Any random variable will take a certain values with a certain probability. The function that controls these probabilities is called the **Probability Mass Function**, or pmf, in the discrete case. It is called the **Probability Density Function** in the continuous case, or pdf. In the discrete case the value of the pmf at a certain value of the random variable will give the probability that the random variable will take such a value, for example $f(x) = P(X = x) = p(x)$. On the other hand, the value of the pdf at a certain value of the random variable will give a point on the graph of the pdf for that variable. For finding the probability in the continuous case, we find that area under the curve of the pdf between two points on the range of the random variable. In addition to the pmf and pdf there is another important concept, or function, called the **Cumulative Distribution Function**, or cdf. The cdf is defined all over the real line regardless whether the variable is discrete or continuous, and it is given by:

$$F(k) = P(X \leq k) = \sum_{all\ values\ of\ x}^{k} p(x), \text{ in the discrete case.}$$

While

$$F(x) = \int_{-\infty}^{x} f(t)dt \text{, in the continuous case.}$$

In case the student is not yet familiar with integration, there should be no worry, since we are not going to use that notion very much in this text at this level. The cdf and the pmf of a random variable are related as shown above.

A probability Model, or a probability distribution, is a table, in the case of a discrete random variable, that depicts the values of the random variable and their associated probabilities. In the case of a continuous random variable, the probability model is a formula that gives the probability of that random variable over a defined set of values on the real line.

EXAMPLE 2.5:

Consider the experiment for counting the number of heads in tossing a fair coin twice.

Solution:

Thus the sample space has the following outcomes: HH, HT, TH, TT. Counting the heads in each outcome we find that our discrete random variable takes 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 above for each being non-negative and add up to one.

❖

2.5 Expected Value and Variance of a Random Variable

Given a set of measurements, we sometimes 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 a substance, and the variability in the readings of some apparatus.

In statistics, 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.

The mean or the expected value, of a discrete random variable is given by

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

where x is the value of the random variable and p(x) is the associated probability with that value of X. On the other hand, in the continuous case, the mean is given by

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

It is seen that E(X) can be considered as a **weighted average** of the values of X with the associated probabilities of those values being the weights.

The variance of a random variable denoted by V(X), or Var(X), or by σ_X^2, in either case, is given by

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

Where μ_x is as defined above for either random variable, and $E(X^2)$ is given by

$$E(X^2) = \sum_{\text{all values of } x} [x^2 . p(x)], \text{ in the discrete case, and}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 . f(x) dx, \text{ in the continuous case.}$$

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

It is 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. However, the variance, as it could be checked, is expressed in square units of the measurements. 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 definition of the standard deviation, denoted by σ_x, as the positive square root of the variance, i.e.

$$\text{The Standard Deviation of a random variable} = \sigma_x = + \sqrt{\ ^2_X} .$$

We will not get involved in finding the mean and the standard deviation of a continuous random variable at this time. This is left for a higher course in statistics for those who are interested and pursuing more courses in statistics.

As it was discussed, and for the discrete case, the following rules apply

Let P(x) denote the probability that a random variable X equals x; P(X = x) = P(x), then

1. $0 \leq P(x) \leq 1$,
2. $\sum_{\text{all values of } x} P(x) = 1$, the sum of the probabilities of all the outcomes, in a probability experiment must be equal to 1.

EXAMPLE 2.6:

Which is of the following tables represent a discrete probability distribution?

(a) X	p(x)	(b) x	p(x)	(c) x	p(x)
1	0.2	1	0.2	1	0.2
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) This is not a discrete probability distribution since P (5) = -0.07, which is less than 0.
b) This is not a discrete probability distribution since

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

c) This is 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 2.7:

Based on Example 2.6, part (c), above that was shown to be a discrete probability distribution, Find

a) The mean,
b) The Variance, and
c) The standard deviation.

Solution:

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

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

$E(X^2) = \sum_{\text{all values of } x} [x^2.p(x)] = 1(0.2) +4(0.25) +3(0.10) +16(0.15) +25(0.30) = 12.$

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

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

❖ ----------

2.6 Some 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 five probability distributions that are quite referred to very often. Those distributions are: The Binomial, The Poisson, and The Geometric, The Hyper Geometric, and The Negative Binomial distributions. We start with the Binomial distribution.

2.6.1 Binomial Probability Distribution

A Bernoulli trial is that 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. 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 is constant throughout the experiment, as well as that of a failure, based on 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 the values of x = 0, 1, 2, …,n.
4. The probability of obtaining X successes, in those n independent Bernoulli trials, is given by the following pmf

$$P(X = x) = P(x) = {}_nC_x\, p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, 2, \ldots, n.$$

With ${}_nC_x$ 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.

The Values for the above pmf, based on different values for n and p, are tabulated as a cdf values in Table II, in the Appendix.

EXAMPLE 2.8:

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 is the number of mice contracting the disease. Then X ~ 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 the Table, with n = 10, p = 0.2, and x =3.
b) $P(X = 5) = P(5) = P(X \leq 5) - P(X \leq 4) = 0.9936 - 0.9672 = 0.0264$.

❖

2.6.1 Mean and Variance for a Binomial Random Variable

We discussed finding the mean (or expected value) and standard deviation of a discrete random variable in Section 2.4. Those formulas can be used to find the mean (or expected value) and standard deviation for the binomial random variable as well.

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

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

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

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

(For the mathematically interested student, look-up the formula for how to expand a binomial term, and you can derive the above formulas.)

EXAMPLE 2.9:

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

$_X$ = np =300(0.75) = 225, and

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

❖ ----------

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.

EXAMPLE 2.10:

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 that the values that the random variable will take in each of the case 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)	0.0000	0.0000	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

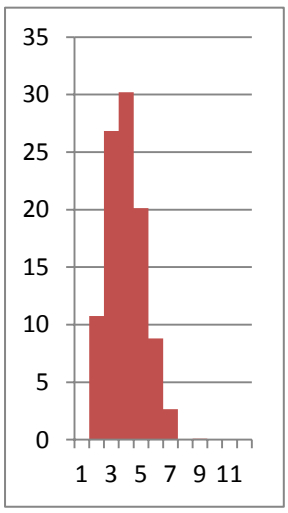


Figure 6A

b) Histogram

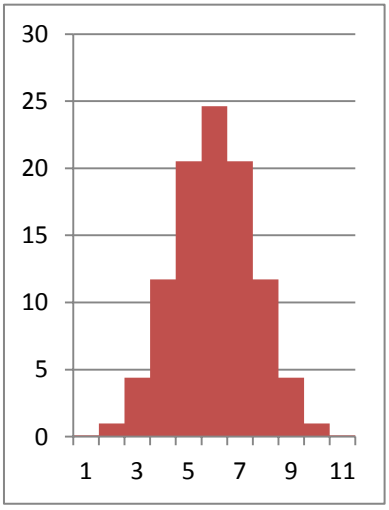
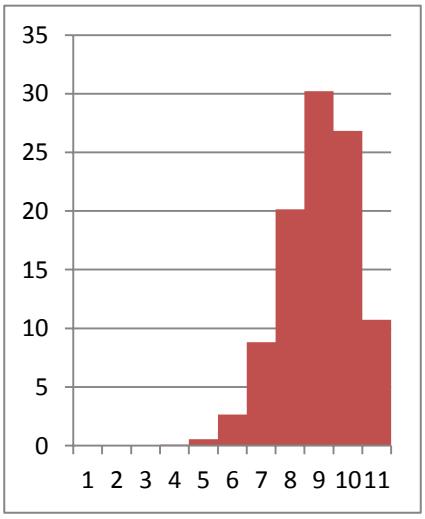


Figure 6B

c) Histogram



Figures 6 C

❖

Using the results in Example 2.10, 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, and the histogram for the probability distribution of the random variable X becomes approximately bell-shaped. As a rule of thumb, if np (1-p) ≥10, the probability distribution for a binomial random variable will be approximately bell-shaped.

2.6.2 Poisson Distribution

Consider a sequence of random events such as: radioactive disintegration during a unit time interval, incoming g calls at a telephone switchboard during lunch hour, the number of traffic accidents at a certain intersection during a week, the number of massages you sent to friends per week, and the number of misprints per page in a book. All of the above events display a counting measure, X, of a random phenomena over a continuous medium, and this counts to define a Poisson distribution. This distribution has the following pmf,

$$P(X = x) = p(x) = {}^{x}e^{-x} / x!, \qquad x = 0, 1, 2, \ldots$$
$$= \text{zero, otherwise.}$$

The constant, $\lambda > 0$, 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 III has the probabilities for the Poisson distribution.

(It can be easily shown, if the student is familiar with calculus, that the above p(x) satisfies the conditions for a pmf)

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 2.9:

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 III, 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 (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 III as P(X = 4) = P($X \leq 4$) – P($X \leq 3$) = 0.629 – 0.433 = 0.196.

❖

Something EXTRA

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

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

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

$$E(X^2) = E[X(X-1)] + E(X)$$

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1).\frac{e^{-} \; . \; ^{x}}{x!} = \quad ^{2}\sum_{x=2}^{n} \frac{e^{-} \; . \; ^{x-2}}{(x-2)!} = \quad ^{2}$$

$$^{2} = E(X^2) - [E(X)]^2 = \quad ^{2} + \quad - \quad ^{2} = \quad .$$

2.6.3 Geometric Distributions

Let us consider an experiment in which the conditions are like those for a binomial one. In other words, let us have an experiment where there are two outcomes only, namely success and failure, with P (success) = p, and that of a failure P (F) = q, with p + q = 1. 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 a geometric distribution;

$$P(X = x) = p^x.q^{x-1}, \qquad x = 1, 2, 3, \ldots$$

$$= 0, \text{ otherwise.}$$

As it was with the Poisson distribution, the above function satisfies the conditions for a pmf.

EXAMPLE 2.10:

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 2.11:

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 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: Due to their importance and use, the mean and variance of a random variable, following the geometric distribution, are given by:

$$\mu = 1/p \text{ and } \sigma^2 = (1 - p)/p.$$

Again the proof is left to the interested reader who is already equipped with some calculus.

2.6.4 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, 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

$$P(X = x) = p(x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, 2 \ldots n;\ 0 \le x \le R,\ 0 \le n - x \le N$$

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 2.12:

A class in Statistics 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

$$P(X=x) = p(x) = \frac{\binom{10}{x}\binom{15}{5-x}}{\binom{25}{5}}, \quad x = 0, 1, 2, \ldots 5.$$

The above pmf for X can be expressed in a tabular form as follows:

X	0	1	2	3	4	5
P(x)	0.0565	0.2569	0.3854	0.2372	0.0593	0.0047

❖

Theorem: The mean and the variance for the Hyper geometric Random variable X are

$$= nR / N$$

And

$$^2 = \frac{N-n}{N-1}.n.\frac{R}{N}.(1-\frac{R}{N}).$$

(For the proof, see Walpole, R.E. and Myers, R.H.; Probability and statistics, 4th Edition, MacMillan 1989)

2.6.5 Negative Binomial Distribution

As it was the case with the Binomial distribution, we are 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. The pmf for X is called the negative Binomial distribution. If we are 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 are r – 1 success, 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

$$p^{r-1}q^{x-r}p = p^r q^{x-r}.$$

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 negative binomial distribution to be given by

$$P(X = x) = p(x) = \binom{x-1}{r-1} . p^r q^{x-r} , x = r, r+1, r+2, \ldots$$

EXAMPLE 2.13:

Find the probability that a person flipping a 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) Using the negative binomial distribution with x =4, r = 1, and p = 0.5, we find that

$$P\ (4) = \binom{3}{0}(0.5)^4 = 1/16.$$

❖ ----------

2.7 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 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 7.**

The normal distribution is no doubt the most important distribution in statistics, and the most widely used continuous probability distribution. There are 4 basic reasons why the normal distribution occupies a prominent place in statistics.

1. The normal distribution comes close to fitting the actual observed frequency distributions of many phenomena:

a) Human characteristics such as weights, heights, and IQs.

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.

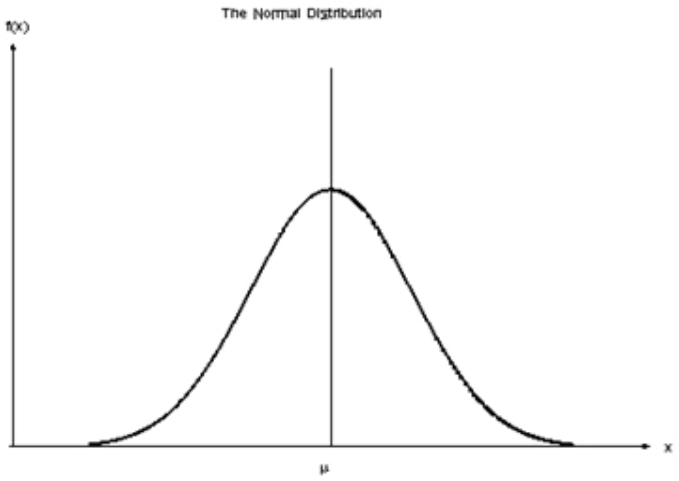


Figure 7 the Bell-Shaped Curve

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 and statistical inferences. This property is clear in the field because the distribution of the mean and the 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 normal data.

The probability density function (pdf) for a normally distributed random variable X, with mean μ and variance σ^2, in short; X ~ N (μ, σ^2), is given by

$$f(x) = \frac{1}{\sqrt{2}}\exp[-(x-)^2/(2\ ^2)],\ -\infty< x <\infty,\ -\infty < \mu <\infty, \text{ and } 0 < \sigma<\infty.$$

In the above notation, for the pdf of the normal random variable X, μ is the mean, or the location parameter, while σ is the standard deviation, the shape, or scale parameter. Due to the extensive use of the above pdf for finding probabilities, and to the exhaustive and wide range of the values for the mean and the standard deviation of X, a unique table has been introduced based on transforming the above general R.V. to the standard normal R.V. Z, where Z = (X–μ)/σ, and thus Z ~ N (0, 1), and it is called the standard normal distribution, **Figure 8.**

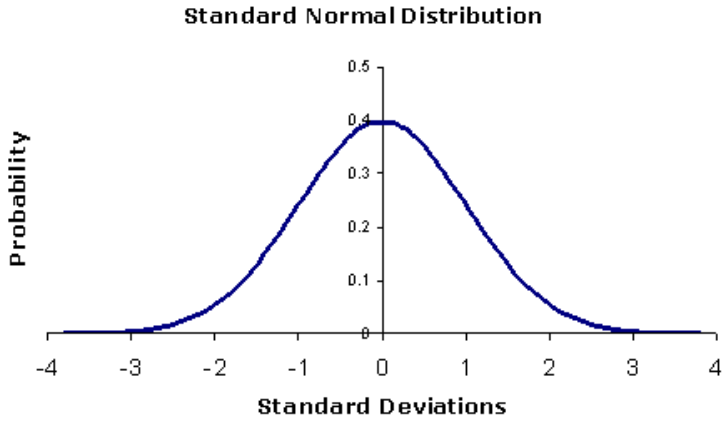


Figure 8

This transformation has tremendously reduced the volumes of the tables that will correspond to the different values of μ & σ, into one single table. Some basic properties of the probability density function for normal random variable X are

1. f(x) is nonnegative all over the real line.
2. The graph of f(x) is symmetric around the value μ, and it is bell-shaped, **Figure 9**.

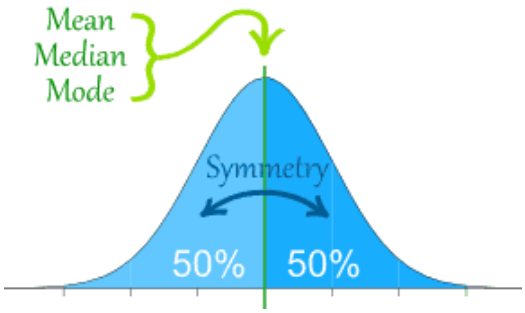


Figure 9 (internet)

3. The integral of f(x) over the real line is 1, i.e $\int_{-\infty}^{\infty} f(x)dx = 1$ i.e. 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. 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 ≤x ≤b, in other words,

$$P(a \le x \le b) = \int_{a}^{b} f(x)dx$$, as shown in **Figure 10**.

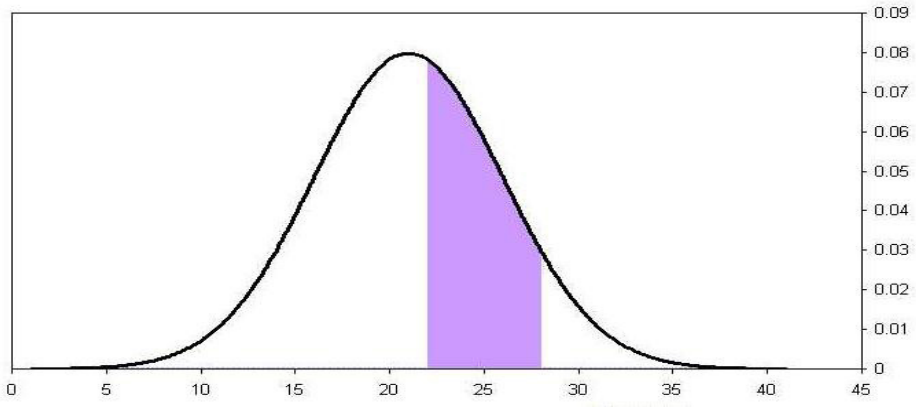


Figure 10 (Internet)

6. It is to be noted that P(a< x < b) = P(a < x ≤b) = P(a ≤ x < b) = P(a ≤ x ≤ b), and this is due to the fact that

$$\int_{a}^{a} f(x)dx = 0.$$

In other words, there is no area under a point.

7. The Empirical Rule: or 68 – 95 – 99.7 rule, is the statistical rule for a normal distribution determined by the mean and the standard deviation. Approximately 68% of the area under the normal curve is between X = $\mu - \sigma$ and x = $\mu + \sigma$, and 95% of the area under the normal curve is between X = $\mu - 2\sigma$ and x = $\mu + 2\sigma$, while 99.7% of the area under the normal curve is between X = $\mu - 3\sigma$ and x = $\mu + 3\sigma$, check **Figure 11A.**

   In terms of probability and mathematical notation the above facts can be expressed as follows:

   $$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6827,$$

   $$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545, \text{ and}$$

   $$P(\mu - 3\sigma \leq x \leq \mu + 3\ \sigma^2\sigma) \approx 0.9973$$

EXAMPLE 2.14:

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.

a) What percentage of seniors scored between 390 and 590 on this SAT test?
b) One student scored 795 on this test. How did this student do compared to the rest of the scores?
c) 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?

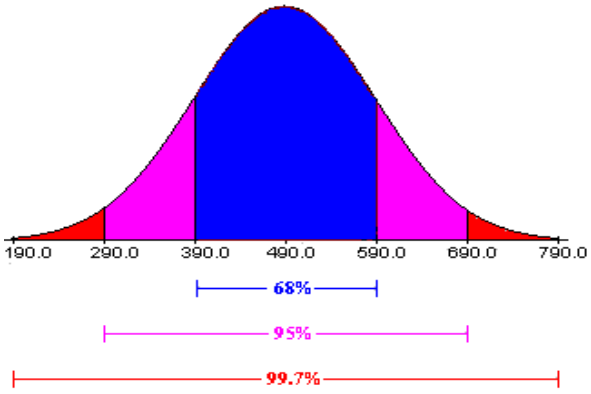


Figure 11A (Internet)

For the example above we have X ~N (490, 100^2), Figure 6A displays the areas noted above.

Solution:

The data being described are the verbal SAT scores for all seniors taking the test 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 below.

a) From the **Figure 11A** above, about 68% of seniors scored between 390 and 590 on this SAT test.
b) 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.
c) 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 11B**, below, shows these percentages.

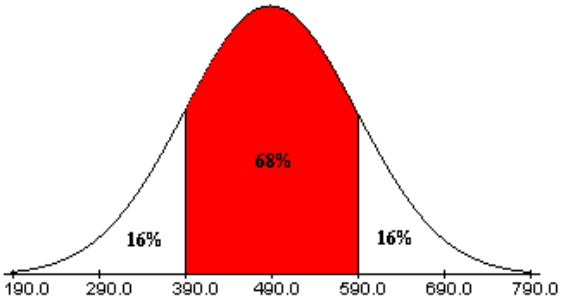


Figure 11B (Internet)

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 2.15:

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 1.50 kg.
b) The probability that a chicken weighs between 0.90 kg and 1.20 kg.
c) The probability that a chicken weighs more than 1.60 kg.
d) The percentage of the chickens that weigh between 0.890 kg and 1.50 kg.
e) Among a group of 300 chickens how many will weigh between 0.80 and 1.50 kg?

Solution:

Let X be the weight of a chicken, then X has a normal distribution with μ =1.0, and σ=0.2, i.e. X ~ N (1.0, 0.04). By using Z = (X -μ)/σ, we have

a) P(X < 1.5) = P[(X -μ)/σ < (1.5 -μ)/σ] = 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 ≤1.6) =1 – P (Z ≤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* 300 = 250.53 ≈251.

❖

For the normal distribution, and revisiting its pdf

$$f(x) = \frac{1}{\sqrt{2}}\exp[-(x- \quad)^2/(2 \quad ^2)],\ -\infty< x <\infty,\ -\infty< \mu <\infty, \text{ and } 0 < \sigma<\infty.$$

We find a large number of values for the mean, for the variance, and also for the variable itself. There is no way that a table, or tables, will be made for any combination of those values. Having another look at the standardizing formula, namely

$$Z = (X - \mu)/\sigma,$$

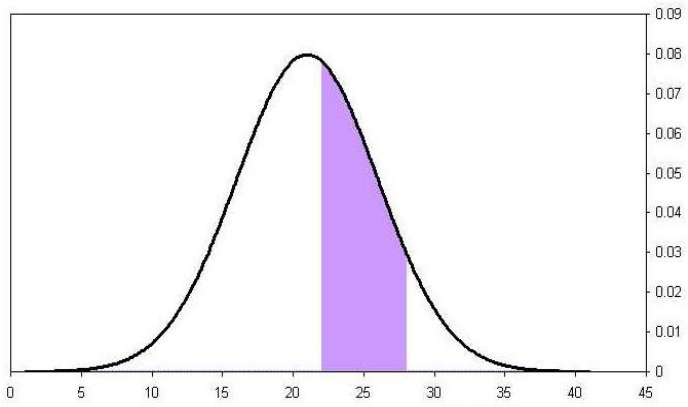
We see that just one table is needed. That table is the Standard Normal Table for the Random Variable Z which is distributed as $Z \sim N(0, 1)$, Check **Figure 8.**

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

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b).$$

$$P(a \le x \le b) = \int_a^b f(x)dx$$, as shown in **Figure 10**.



Is it one probability or four different ones? All are equal, whether we include the end points, or exclude them, or include one and exclude the other. This is based on the concept, in calculus; there is no area above a point in the continuous case of a random variable,

$$P(X = a) = \int_a^a f(x)dx = 0$$

2. Finding the area to the left of a value for the random variable: P(x < c), **check Figure 12**

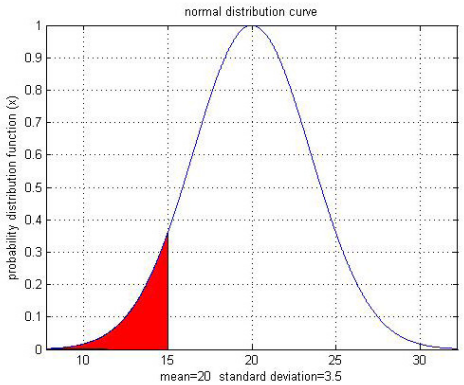


Figure 12 (Internet)

3. Finding the area to the right of a value for the random variable: P(x > d), the un-shaded area in **Figure 13**

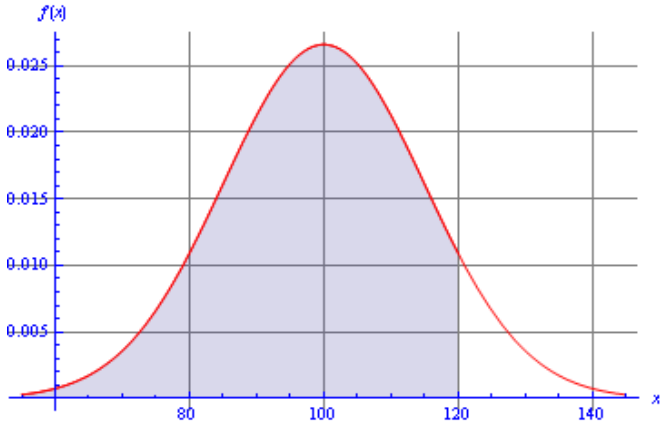


Figure 13 (Internet)

With no doubt that 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 save 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

$$Z = (X-\mu)/\sigma,$$

the above three cases for finding the probabilities can be calculated by using the standard normal table. The equivalent case, in terms oz will look like the following

1. $P(z_1 \le Z \le z_2)$ = P (a ≤ x ≤ b).
2. $P(Z \le z_3)$ = P(x < c).
3. $P(Z \ge z_4)$ = P(x > d).

The standard normal Table gives the area to the left of any point, to find the probability for

Part 1, we have

$$P(z_1 \le Z \le z_2) = P(Z \le z_2) - P(Z \le z_1).$$

To find the probability for part 2, it is straight forward from the table.

For part 3, since the total area under the curve is 1 and the table lists the area to the left we find ourselves doing the following for part 3.

$$P(Z \ge z_4) = 1 - P(Z \le z_4).$$

EXAMPLE 2.16

Let $X \sim N(70,100)$, where $\mu = 70$, and $\sigma = 10$. Find

1. P (56.5 < X < 90.1), 2. P (X < 73.2), and 3. 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:

1. 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.

Second decimal place for z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0		.9778								

We read .9778 for 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				

Similarly, we read 0.0885 for -1.3 under z and under .05, to get 0.0885. The difference is the answer, as it is seen above.

2. 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							

Similarly, we read 0.6255 for 0.3 under z and under .02, to get 0.6255

3. $P(X > 86.8) = 1 - P(X < 86.8) = 1- P\{(86.8-70)/10] = 1- P(Z < 1.68) = 1-0.9535 = 0.0465$.

Second decimal place for z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6									.9535	

Similarly, we read 0.9535 for 1.6 under z and under .08, to get 0.9535.

❖

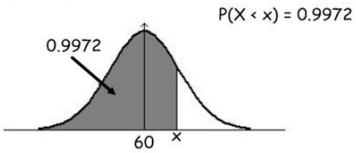
For the figure below the way is going backwards. It is a two way street. Now we are given the area, which is standing for probability, we need to find the cutting point, whether on the X-axis or the Z-axis. Finding the cutting point on one of the axes and using the transformation, below, will get you the other cutting point.

$Z = (X - \mu)/\sigma$.

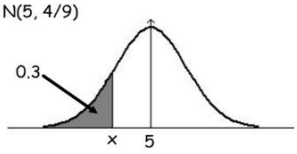
EXAMPLE 2.17

Find the value of x in each of the following diagrams:

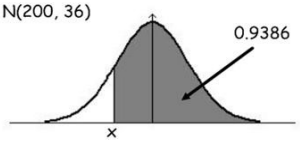
(a) X ~ N(60, 25)



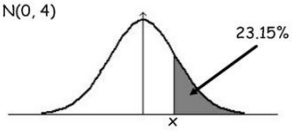
(b) X ~ N(5, 4/9)



(c) X ~ N(200, 36)



(d) X ~ N(0, 4)



(a)

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.7								.9972		

Reading the value in the Standard Normal Table, we found that the 0.9972 is 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

(b)

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.5			.3015	.2981						

Reading the value in the Standard Normal Table, we found that the 0.3000 is between the two values cited along -0.5 under Z and under 0.02 and 0.03 for the second place. Since 0.2981 is closer to 0.3 than .3015, we can take z to be -0.53. Using the transformation based on the distribution of X, we have x = (2/3) (-0.53) + 5 = 4.65.

(c)

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.5				.0630						

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 = 0.0632. Based on that, we now read 0.0632 to be found closer to 0.0630, which is cited along -1.5 under Z and under 0.03 for the second decimal place. We can take z to be -1.53. Using the transformation based on the distribution of X, we have x = 6(-1.53) + 200 = 190.82.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.5				.0630						

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 in 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 0.7685 to be found closer to 0.7673, which is cited along 0.7 under Z and under 0.03 for the second decimal place. We can take Z to be 0.73. Using the transformation based on the distribution of X, we have x =2(.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.

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.

CHAPTER 2 EXERCISES

2.1 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.

2.2 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.

2.3 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

2.4 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

2.5 Consider the exercise 2.4, after finding the missing probability, find

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

2.6 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 = λ.) 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.

2.7 Find the probability of the indicated event if P (E) = 0.25 and P (F) = 0.45

a) Find P (E or F) if P (E and F) =0.15.
b) Find P (E and F) if P (E or F) 0.6.
c) Find P (E and F) if E and F are Mutually exclusive.
d) Find P (E or F) if E and F are Mutually exclusive.

2.8 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.

a) What percentage of fish will weigh less than 4 lbs?
b) What percentage of the fish will weigh within one lb. of the average weight?
c) What is the chance that one fish will weigh more than 5 lbs.?

2.9 The inside diameter of a piston ring is normally distributed with mean of 4 inches and a standard deviation of 0.01 inches.

a) What percentage of the rings will have an inside diameter exceeding 4.025 inches?
b) 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?

2.10 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 the specifications

a) Cover 95% of the components
b) Cover 90% of the components
c) Cover 99.7% of the components

2.11 A sample space consists of five simple events, E1, E2, E3, E4, and E5.

a) 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).
b) 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.

2.12 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	0.3	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

2.13 Consider the Tables in 2.9, 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.

2.14 Suppose in families with 4 children, 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

a) The expected number of boys in a family of four children.
b) The variability in the number of boys in a family of 4 children.

2.15 Consider a binomial probability distribution with parameters n = 5 and p = 0.2.

a) Construct a binomial probability distribution with these parameters.
b) Computer the mean and standard deviation of the distribution.
c) Draw a probability histogram and comment on the shape, and label the mean on the histogram.

2.16 The random variable X follows a Poisson process with $\lambda = 4$. Find each of the following:

a) P (3), b) P(X<3), c) P(X ≤ 3), d) P(X ≥ 4), e) P (3≤X≤5).
f) What are the mean, the variance, and the standard deviation?

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)

4. Press 2nd VARS to access the probability distribution menu.
5. Highlight 0: binompdf (and hit ENTER.
6. 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 ≤ x)

1. Press 2nd VARS to access the probability distribution menu.
2. Highlight A: binomcdf (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
   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 ≤ 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 ≤ 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.

Excel

Computing P(x)

1. Enter the desired values of the random variable X in column A.
2. With cursor in cell B1, select the Formulas tab, Select more Formulas. Highlight Statistical, and then highlight POISSON in the function name menu.
3. 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 ≤ 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 *lowerbound, upperbound, 0, 1*). 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 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 μ = 40 and σ = 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 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 μ = 40 and σ = 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

TI-83/84 Plus

1. Enter the raw data into L1.
2. Press 2^{nd} 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.

3 Estimation

Outline

3.1 Sampling

The purpose of sampling is to enable us to make inferences about a population after inspecting only a portion (a sample) of that population. Such factors as cost, time, destructive testing, and infinite populations make sampling preferable to making a complete inspection (census, complete enumeration) of a population. Usually, there are some numerical characteristics about the population which the investigator wants to know. Such numerical facts are called **parameters**; e.g., the population size, population mean, a proportion of some attribute, and variability in the population, etc.

A Parameter is a numrical value, or a characteristc, of a population

The parameters cannot be determined exactly, but can only be estimated from a sample by quantities called **statistics.**

A Statistic is a numerical summary, or a characteristc, of a sample.

Then a major issue is accuracy. How close the estimators (the statistics) are to the true values (the parameters)? Logically, we will get an accurate estimate if the sample is **representative** of the population from which it was drawn. How to get a sample, and make sure we have a fairly large representative sample. There are two categories of sampling methods.

3.1.1 Non-Probability Sampling

a) **Convenience Probability Sampling:** The sample is restricted to a part of the population that is readily accessible. A sample of coal from an open wagon may be taken from the top 6 to 9 inches.
b) **Haphazard sampling:** The sample is selected haphazardly by picking a sample of ten rabbits from a large cage in a laboratory. The investigator may take those that his hands rest on, without conscious to planning.
c) **Judgment (or subjective) sampling:** With a small but heterogeneous population, the investigator selects a small sample of "typical" units – that are most representative of the population according to his judgment.
d) **Volunteer sampling:** The sample consists of essentialy volunteers because the measuring process is unpleasant or troublesome to the person being measured. This, particulary, occurs in medical research, and clinical trials.

Under the right conditions, any of the above methods, of non-probability sampling, may give a representative sample of the population. However, there are two disadvantages:

1. There may be a systematic tendency on the part of the sampling procedure to exclude one kind of a unit or another from the sample. This is called **selection bias**.
2. Non-probability sampling is not amenable to the development of sample theory, since no element of random selection is employed. That is, we cannot compute the precision, or any quantity measuring the error of the estimators.

3.1.2 Probability Sampling Methods

There is a definite procedure, for selecting the sampling units, that involves the **planned** use of chance. In a probability sampling procedure, the laws of probability can be applied to compute the chance of selecting the different samples and the individual units. Also, the frequency distribution and the precision of the estimators can be computed, at least theoretically. In probability sampling, the investigator has no discretion at all as to which units to select. There is no selection bias.

A very important and basic type of probability sampling is the **Simple Random Sampling.** Variations of simple random sampling include **systematic, stratified, and cluster sampling.** These sampling methods, in addition to the non-probability **convenience** sampling, make the methods of sampling in statistics. Let us give a formal definition for each of the four mentioned sampling methods.

Simple random sampling: Simple random sampling selects samples, of size n, by methods that allow each possible sample of size n to have an equally likely chance or **equal probability** of being considered. In addition, each unit, in the entire population, has an equal chance of being included in the sample in each single drawing.

EXAMPLE 3.1

A small community college employs 58 full-time faculty members. To gain the faculty's opinion about upcoming contract negotiations, the president of the faculty union wishes to obtain a simple random sample that will consist of 5 faculty members. How will you help in selecting the committee members by a simple random sampling?

Solution:

Using Table I, the Random Number Table, from the Appendix, let the prsident close her eyes and drop a pencil on the Table. It points to row 19, and column 7. Let this be the starting point. Since the numbers of the faculty are two digits, so 1 will be representd by 01, and so on till we get all faculty identified by their numbers. Based on that, we need to consider columns 7 and 8, for the two digits. So the following sample will occur: 29, 32, 09, 58, and 12. those faculty members with the corresponding numbers will be on the committee.

❖

A Systematic Sample is obtained by selecting every k^{th} individual from the population. The first individual selected correspond to a random number between 1 and k.

Steps in Systematic sampling

1. Make sure that you have a finite population of size N, and enumerate each individual.
2. Decide on how large your sample will be, n.
3. Calculate N/n, and round, up or down, to an integer. Let that integer be k.
4. Randomly select a number between 1 and k, call that number j.
5. Thus the sample will consist of the following enumerated individuals:
   j, j+k, j+2k, …j+(n-1)k.

EXAMPLE 3.2

The human resources department at a certain company wants to conduct a survey regarding workers benefits. The department has an alphabetical list of all 2949 employees at the company and wants to conduct a systematic sampling of size 30.

A) What is k B) Determine the individuals who will be administering the survey. Choose j, and suppose that we randomly selected 20, who will be in the survey?

Solution:

K = 2949/30 = 98. From the information we have j = 20. Thus the units in the sample will be: 20, 118, 216, …, 2862.

❖

A Stratified Sample is obtained by dividing the population into separate homogeneous categories, or groups, that do not overlap. These are called Strata (the singular is Stratum) and then take a sample by simple random sampling from each stratum. These chosen subsamples will form the stratified sample needed.

EXAMPLE 3.3

There is a class of 400 students enrolled in a multi section course Stat 1350. You like to have a committee representing the class, and every section in the class. The groups here are the different sections of students.

Solution:

The strata here are the different sections in the course. Thus the course divided, and based on how many will be on the committee, you choose at random from that section, and choose from that stratum at random, and repeat until you get your members.

❖

A Cluster Sample is obtained like the stratified sample by dividing the popultion into groups, obtain a simple random sample from the proups, as a whole, and select all the individuals within the randomly selcted group or stratum.

EXAMPLE 3.4:

The above example can be redone by clustering. Here once the section, or stratum was chosen, then every one in that "section" will be taken into the sample.

❖

Recall that technology is being used to do almost any kind of sampling. For example, you can use **MINITAB** for obtaining a sample.

In the rest of the text we will confine ourselves to probability sampling, especially to Simple Random Sampling (SRS). For more detailed discussion on sampling techniques, the interested reader is referred to Cochran, 1978.

3.2 Point Estimation

Estimation stands as the first part of inferential statistics, while the second part is the hypothesis testing. There are two types of estimation: **point estimation** and **interval estimation**. We will address the point estimation in this section. The interval estimation will be addressed in the next section.

A point Estimate is that value of a statistic, which has been calculated from a sample, that estimates a parameter of the population.

In this section, we will consider the point estimation of proportions, population's mean and population's variance.

3.2.1 Methods of Estimation (Optional)

let us outline the procedures by which we can find the point estimators of a parameter. The procedures to be used here are: (1) the method of moments, and (2) the maximum liklihood method.

A) The Method of Moments

Let $X_1, X_2, \ldots, X_n$ be a random sample of a random variable X. The average value of the kth power of $X_1, X_2, \ldots, X_n$.

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

is called the kth sample moment, for k = 1, 2, 3, … . while E $[X^k]$ is called kth population moment, for k = 1, 2, 3, … . Thus the method of moments estimators of the parameters is given by setting the sample moments equal to population moments and solving the resulting equations simlutaneously, for the parameters of the population.

Maximum likelihood estimators

The essential feature of the principle of maximum likelihood estimation, as it applies to the problem of estimation, is that it requires the investigator to choose as an estimate of the parameter that value of the parameter for which there is the apriori probability of obtaining the sample point actually observed, is as large as possible. This probability will in general depend on the parameter, which is then given that value for which this probability is as large as possible.

Suppose that the population random variable X has a probability function which depends on some parameter θ =Pr [X=x] = f (x; θ). We suppose that the form of the function f is known, but not the value of θ. The joint probability function of the sample random variables, evaluated at the sample point ($x_1, x_2, \ldots, x_n$), is

$$L(\theta) = f(X_1, X_2, \ldots, X_n; \theta) = \prod_1^n f(x_i, \theta).$$

This function is also known as the likelihood function of the sample. We are considering it as a function of θ when the sample values $x_1, x_2, \ldots, x_n$ are fixed.

The principle of maximum likelihood requires us to choose as an estimate of the unknown parameter that value of θ for which the likelihood function assumes its maximum value.

3.2.2 Statistics as Estimators for Parameters

It is clearly visible that we use statistics to estimate parameters due to the lack of time, energy, resources, and infinite populations. Statistics, from the sample, can be listed as: proportions, Arithmatic averages, ranges, quartiles, deciles, percentiles, variances, and standard deviations. It will become clear enough what each one means and what it will stand for.

A) The sample proportion

Suppose that there is a population in which each individual either does or does not have a certain characteristc. We like to consider a random sample out of this population. Selecting an individual from this population is an experiment that has ONLY two outcomes: either that individual has the characteristic we are intersted in or does not.This kind of an experimant is called a Bernoulli expriment that has two outcomes, from now on labelled, a success or a failure. Let a random sample of size n be taken from this population. The sample proportion, denoted by $\hat{p}$ (read "p-hat"), is given by

$$\hat{p} = \frac{x}{n}.$$

Here x is the number of individuals in the sample with the specified characteristic. The sample proportion $\hat{p}$ is a statistic that estimates the population proportion, p, which is a parameter.

EXAMPLE 3.5

On the first day of the semester, in Stat 1350, asking a class of 30 students the following question: Who has a calculator? 12 students raised their hands by showing that they have a calculator. Estimate the percentage of the students, in this class, that have a calculator.

Solution:

By our notation we see that n = 30, and x = 12. Therefore $\hat{p} = \frac{x}{n}$ = 12/30 =0.40 = 40.00%.

❖ ----------

B) The sample mean and the sample variance

Let there be a population with unknown mean "μ", (lowecase Greek mu) and unknown standard deviation "σ" (lower case Greek sigma). To estimate μ and σ, we draw a simple random sample of size n: x_1, x_2, …, x_n. Then compute the **sample mean**

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

and compute the **sample variance**

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

Then, μ is estimated by $\overline{X}$, i.e, $\hat{} = \overline{X}$, and σ is estimated by s, $\hat{} = s$, called the sample standard deviation.

NOTE: S^2 can be computed from

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

This formula will give more accurate values, since we are rounding twice in this formula compared to (n+2) times of rounding in the formula for the definition of the sample variance S^2.

EXAMPLE 3.6:

In a certain experiment it was required to estimate the nitrogen content of the blood plasma of a certain colony of rats at their 37th day of age. A sample of 9 rats was taken at random and the following data was obtained (grams per 100 c.c. of plasma): 0.98, .83, .99, .86, .90, .81, .94, .92, and 0.87. Find the estimates for the average nitrogen content and for the variation in nitrogen content in the colony.

Solution:

Let μ be the true average content of nitrogen and let σ be the true variation in the nitrogen content. Then, μ is estimated by $\overline{X}$ and σ is estimated by S. Using thedata on hand, we find that

$\overline{X}$ = 8.10/9 = 0.90 grams per 100 cc. and
$S^2 = 1/8 \{ 7.3220 - (8.10)^2/9\} = 0.004$ (grams)^2 per 100 cc.
Hence S = 0.063 grams per 100 cc.

❖

It is clear that different possible samples could be drawn from the same population. Each of those samples yields a different value for $\overline{X}$, and aother different value for s^2. One may inquire about the average value of all possible $\overline{X}$s, their standard deviation and their distribution. Likewise for s^2.

The following properties are of theoretical nature and are very essential to the theory of statistical inference. (For proofs and more details see, e.g., Hogg & Craig, 1978).

3.2.3 Properties of the Estimators and their Sampling Distributions

In the frst part of this section, we have estimated the fraction of defectives p of the binomial distribution, the mean and the standard deviation of a population by using the observations $x_1, x_2, \ldots, x_n$ that were obtained bvy sampling from the distribution of interest. With no doubt that those estimators do not equal the true values of the estimated parameters of the interesting distribution. For example, if other samples were taken from the sample distribution, we find that we can have other values for the same parameters. Since the estimates vary from one sample to another, it becomes essential and important to investigate the sampling distributions of those estimators. **What is a sampling distribution**? Depending on the parameter we are esimating, then the sampling distribution of that estimator, for a given sample size n, consists of the collection of all the estimators of that parameter, of all possible samples of size n from that population. Although the derivation of the following result is beyond the scope of our text, we should be convinced that the following result is reasonable.

Sampling Distribution of $\hat{p}$

For a simple random sample of size n such that n is at most 0.05N (that is the sample size is less than 5% of the population size) (check chapter 2)

- The shape of the sampling distribution of $\hat{p}$ is approximately normal provided np(1- p) is at least 10,
- The mean of the sampling distribution of p is $_{\hat{p}}$ = p
- The standard deviation of the sampling distribution of $\hat{p}$ is $_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

The condition that the sample size is no more than 5% of the population size is needed so that result obtained from one individual is independent of the result obtained from any other individual in the sample. The Condition of np(1-p) is at least 10 for normality and help for making inferences later on.

Sampling Distribution of $\overline{X}$

To describe the sampling distribution of the sample mean $\overline{X}$, we need the following definition or assumption. It is clearly understood and makes more sense that the more information we get from the population the closer the statistics values will be to their corresponding parameters. This led us to define the following concept of Law of Large Numbers.

Suppose that a simple random sample of size n is drawn from a large population with mean μ and variance σ^2. The sampling distribution of $\overline{X}$ will have a mean that is equal to the population mean μ, and a variance which is given by σ^2/n.

The Law of Large Numbers: As additional observations are included in the sample, the difference between the statistic $\overline{X}$ (the sample mean) and the parameter μ (the population mean) approaches 0.

In addition to the above, when n increases the sample mean become very close to the population mean, and we see that the standard deviation of the sampling distribution of $\overline{X}$ decreases as n increases. Hence we have the theorem that referred to as the CLT, **the Central Limit Theorem.**

Regardless of the shape of the population, whether it was normal or not, the sampling distribution of $\overline{X}$ becomes approximately normal as n increases with mean and variance as specified above.

The interested reader might consult the above reference for the proof of the stated result (For proofs and more details see, e.g., Hogg & Craig, 1978).

3.3 Interval Estimation

An Interval Estimation is a range of values, calculated based on the information in the sample, that the parameter in a population will be within that range with some degree of confidence.

Suppose that a scientist wishes to weigh an ore sample with high precision. Because of a random error in measurements, he takes several readings and averages them by computing $\overline{X}$. He may report his estimate of the true weight as:

1. As point estimate, $\overline{X}$ or
2. As an interval, $\overline{X} \pm$ E, where E stands for the error. It is the margin of error.

It is seen that a point estimate is a single value that is used to estimate an unknown population parameter. A point estimate is often insufficient because it is right or wrong. If the scientist reports that the estimated weight is 404 micrograms, (mg), then he does not really mean that the true weight is exactly 404 mg, and he should report how much error is involved. Also, it is insufficient to report that the true mean is within the range of values $\overline{X} \pm 1$S.E. because he does not report how much confidence he has in such a statement.

The best thing that can be done is to report estimates as "interval estimates". An interval estimate (or a confidence interval) is a range of values used to estimate a population parameter with a certain degree of confidence. The confidence interval indicates the error of estimation in two ways: i) by the extent of its range and ii) by the probability of the true population parameter will be lying within that range.

In the following sections, we will introduce how to construct a confidence interval for estimating one or two population parameters, and specify the probability as a level of confidence.

3.4 Confidence Interval about One Parameter

In this section we will address the procedures on how to calculate the confidence interval on one parameter. Those parameters are the proportion, the mean, and the variance of a population.

3.4.1 Confidence Interval about One Proportion

It is of interest to estimate the proportion of employees, who favor a certain type of work, or the proportion of defective items in a certain lot, or the proportion of rats having a certain kind of symptoms. Let P be the true proportion of elements that have attribute A (a certain characteristic of interest) in a population. We draw a simple random sample of size n from this population and let X equal number of elements in the sample that have attribute A. Thus the point estimate of the true proportion p in the population is given by $\hat{p}$ = X/n, where X has a binomial distribution with parameters n and p. Recall that E(X) = np, and Var (X) = n.p. (1-p). Hence we find that E ($\hat{p}$) = p, and Var ($\hat{p}$) = p.(1-p)/n. By the Central Limit Theorem, the random variable given $Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ has a standard normal distribution as n increases. Thus

$$P(-Z_{/2} < Z < Z_{/2}) = 1 - \alpha,$$

where Z is the value of the standard normal variable comprising a probability of alpha on its right. This, with some algebra manipulations we reach the 100(1 – α) % C.I. (Confidence Interval) on p to be given by

$$\hat{p} - Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\ .$$

Based on the above confidence interval, when the two limits are given, we can find the sample proportion by the following equation:

The lower limit + the upper limit = 2(sample proportion).

EXAMPLE 3.7

In a simple random sample of 500 employees, 160 preferred to take training classes in the morning rather than in the afternoon. Construct a 95% C.I. on the true proportion of employees who favor morning training classes.

Solution:

From the information on hand we have; x =160, n = 500, the confidence level is 0.95, and so $\hat{p}$ = 0.32, alpha = 0.05, and $Z_{0.025}$= 1.96. By substitution in the above interval we find that 0.28 < p < 0.36.

This means that the true proportion, of the employees who favor the morning training classes, is between 28% and 36%.

❖ ----------

3.4.2 Confidence Interval about One Mean

Let there be a population with mean μ and variance σ^2. We wish to construct a confidence interval about, μ with 100(1- α) % confidence level, where $0 < \alpha < 1$. There are three cases to be considered here. For the following cases, we will have a simple random sample of size n from the original population. Let that sample be X_i, i =1, 2, …, n.

Case I: The variance of the population is known, In this case the random variable

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

has a N (0, 1) distribution and

$$P(- Z_{/2} < Z < Z_{/2}) = 1 - \alpha.$$

With some algebra manipulations we reach the 100(1- α) % C. I. on μ to be given by

$$\bar{x} - Z_{/2} \quad / \sqrt{n} < \quad < \bar{x} + Z_{/2} \quad / \sqrt{n}$$

The above interval is called a z-interval about the mean μ. It is to be noticed here that we can find the sample mean or the margin of error when the limits of any confidence interval are given. This based on the following two equations:

The Lower limit + the Upper limit = 2(the sample mean), while
The upper limit – the lower limit = 2(the margin of error).

Case II: The variance of the population is unknown, with a large sample size ($n \geq 30$)

Since σ is not known, and we have a random sample, we estimate σ by the standard deviation of the sample on hand. By replacing σ by s in the above z-interval we reach at the following interval that will be the 100(1 – α) % C.I. on μ,

$$\bar{x} - Z_{/2} S / \sqrt{n} < \quad < \bar{x} + Z_{/2} S / \sqrt{n}\,.$$

Remark: in the above two cases for constructing the C.I. about μ, the interval is termed a Z-interval.

Case III: The variance of the population is unknown, with a small sample size (n < 30)

Since we have a small sample, and the variance of the population is not known, we ask ourselves Is the population Normally distributed. If not, we stop here and a non-parametric approach is needed. Non-parametric statistics is a field by itself and it is not discussed in the text. But if the answer is yes, that the population is normally distributed then the following approach will be applied to construct a C.I. about μ. Consider the random variable that is defined below

$$T = \frac{\bar{X} -}{S/\sqrt{n}}$$

This random variable will have a student t-distribution with n-1 degrees of freedom. (In two sentences, the t-distribution behaves like the standard normal distribution, with variance >1. It is characterized by its degrees of freedom, and as n increases the t- distribution tends to be very close to the standard normal distribution.) With the random variable T, as given above we see that

$$P(-t_{/2} < t < t_{/2}) = 1-\alpha$$

As before, a little algebra will get us to the following C.I. on μ,

$$\bar{x} - t_{/2} S/\sqrt{n} < \quad < \bar{x} + t_{/2} S/\sqrt{n}$$

EXAMPLE 3.8

Suppose that an investigator wishes to estimate the content of mercury in the water of a certain area. He collected 16 measurements and obtained the following data (as ppm, parts per million): 409, 400, 406, 399, 402, 406, 401, 403, 401, 403, 398, 403, 407, 402, 410, and 399. The investigator wishes to report the results as a confidence interval with 95% confidence level.

Solution:

Clearly, the variance of the population, and hence the standard deviation, is not known. The sample size is small, n = 16 < 30. Thus we well assume that the population is normal, and proceed as in case III. From the data we can get the following statistics: n=16, $\overline{X}$ = 403.063 ppm and S^2 = 12.996, from which we see that S = 3.605 ppm. With the significance level of 0.05 we have from the t-table, with 15 degrees of freedom, $t_{.025}$ =2.131. Incorporating all of these values we reach at the 95% C.I. on the mean content of mercury in the water to be (401.143, 404.983)

❖

3.4.3 Confidence Interval about One Variance

In studying the precision of measuring instruments, and in studying variability in populations, we face the problem of estimating the population variance, or its standard deviation from a random sample. In this section we will investigate how to construct a confidence interval on either the population variance or the standard deviation. For to have a good confidence and excellent estimation for a 100(1 $-\alpha$) % C.I. on the population variance we need to assume that that population has a normal distribution with some variance σ^2.

From a random sample of size n; X_1, X_2… X_n, taken from that normal population, we can see that

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

is the sample variance. Thus the random variable given by

$$^2 = \frac{(n-1)S^2}{^2}$$

will have a chi-square (kigh-square) distribution with n-1 degrees of freedom. Let $^2_{/2}$ and $\chi^2_{1-\alpha/2}$ be those values on the Chi-square axis such that the area to the right of that value is $\alpha/2$ and 1- $\alpha/2$ respectively. Hence

$$P\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1 - \alpha ,$$

Where

$$^2 = \frac{(n-1)S^2}{^2}.$$

As it was done before, with little algebra, we have the 100(1 – α) % C.I. on σ^2 given by

$$\frac{(n-1)S^2}{^2_{/2}} < \ ^2 < \frac{(n-1)S^2}{^2_{1-\ /2}}.$$

In addition to the C.I. on σ^2, we can get the 100(1- α) % C.I. on σ, and thus it is given by just taking the square root of every term in the above interval.

$$\left(\frac{(n-1)S^2}{^2_{/2}}\right)^{1/2} < \ < \left(\frac{(n-1)S^2}{^2_{1-\ /2}}\right)^{1/2}.$$

EXAMPLE 3.9

Human beings vary in the time it takes them to respond to driving hazards. In one experiment in which 100 healthy adults between age 21 and 30 years were subjected to a certain driving hazard, and the sample variance of the observed times it took them to respond was 0.0196 second squared. Assuming that the times to respond are normally distributed, estimate the variability in the time response of the given age group using a 95% C.I.

Solution:

The confidence level is 0.95, so that $\alpha/2 = 0.025$. Reading the χ^2-Table with 100 -1 = 99 degrees of freedom we find that $^2_{0.025}$ =128.45, $^2_{0.975}$ =128.45. Substituting in the C.I. for σ^2 we obtain the following interval

$0.0151 < \sigma^2 < 0.0265.$

Moreover the 95% C.I. on σ is given by

$0.123 < \sigma < 0.163.$

❖

3.5 Sample Size Determination

Frequently, we wish to determine how large a sample should be in order to ensure that the error in estimating the population mean, or the population proportion, is less than a specified value of the error.

As it was shown in the derivation of the confidence interval for μ and P, the margin of error was given by the following two formulas respectively

$$E = Z_{/2} \cdot \sigma / \sqrt{n}, \text{ and}$$

$$E = Z_{/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

when the confidence level is taken to be $100(1 - \alpha)$ % in both cases.

It is quite clear when the sample is too small; the required precision is not achieved. On the other hand, when the sample size is large, then some resources have been wasted. In order to meet the criterion of a specified margin of error calculations can be made to approximate the sample size needed in both cases of the mean and the proportion. Also, checking the formulas above for E, we find that each has 4 quantities. Thus whenever three of those are given the fourth can be found. In case of finding the sample size, the rounding will be up to the nearest whole number in order to meet the error criterion. Manipulating the above formulae we have

$$n = (Z_{/2} \cdot \sigma / E)^2.$$

This formula will give the sample size when E, the confidence level 100(1 – α) %, and the population standard deviation are given.

In case for finding the needed sample size to meet a certain criterion for estimating the population proportion, we have two cases to consider:

1) If there is an estimated value for that proportion p, and
2) If there is no information about P.

The formulae for n will be, respectively, given by

$$n = (Z_{/2} / E)^2 . \hat{P}(1 - \hat{P}) \text{ , and}$$

$$n = 0.25 . (Z_{/2} / E)^2 .$$

EXAMPLE 3.10

Suppose you want to estimate the average weight of chickens in a laboratory. You like to be 95% certain that the error is at most 0.1lbs. How many chickens you should include in your sample?

Solution:

The error is E = 0.1, the confidence level is 95%, and thus $Z_{\alpha/2}$ = 1.96. We are still missing the standard deviation, in the weight of the population σ. There is away to guess that, or to use an estimated value. Pick up the heaviest and the lightest chickens, and weigh those two. Find the range in the two weights, R. A rough estimate of σ is R/6. Let that range to be 1.5 lbs. and hence σ = 0.25 lbs. Hence the required sample size is n = 25 chickens.

❖ ----------

EXAMPLE 3.11

An economist wants to know if the proportion of the population who commutes to work via carpooling is on the increase due to gas prices. What sample size should be obtained if the economist wants to estimate with 2 percentage points of the true proportion with90% confidence if

a) The economist uses the 2006 estimate of 10.7% obtained from the American Community Survey?
b) The economist does not use any prior estimates?

Solution:

In both cases we have E = 0.02, and $Z_{\alpha/2}$ = 1.645. Hence for

a) We use $n = (Z_{/2} / E)^2 . \hat{P}(1 - \hat{P})$, and find that n = 647.

b) We use $n = 0.25.(Z_{/2} / E)^2$, and find that n = 1692.

We can see the effect of not having a prior estimate of p. In this case, the required sample size more than doubled of the case when we used a prior estimate.

❖

3.6 Confidence Interval about two Parameters

The confidence intervals to be calculated on two parameters will involve: a) two means, b) two proportions, c) two variances, and two standard deviations. In addition to those parameters, we will establish a confidence interval between the means of two matched samples. For the means and proportions the confidence intervals will be established on the difference between the two parameters, while in the case of two variances or two standard deviations, the confidence interval will be on the ratio between the two parameters.

3.6.1 Confidence Interval about the difference between two proportions

Comparisons of proportions, in different groups, are a common practice. A whole-seller compares the proportions of defective items found in two separate sources of supply from which he buys these items. A safety engineer compares the proportions of head injuries sustained in an automobile accident by passengers with seat belts against those without seat belts.

Consider two independent samples of sizes n_1 and n_2 that are drawn from two binomial populations with parameters (i.e. probabilities of successes) p_1 and p_2. A 100(1 – α) % confidence interval will derived on the difference between p_1 and p_2 using the central limit theorem and the normal approximation to the binomial distribution.

Let x_1 and x_2 be the number of successes obtained in sample 1 and sample 2 respectively. We then have $\hat{P}_1 = x_1/n_1$ and $\hat{P}_2 = x_2/n_2$ as the point estimates of p_1 and p_2 respectively. Moreover we have

$E(\hat{P}_1) = \hat{P}_1$, $E(\hat{P}_2) = p_2$, with $Var(\hat{P}_1) = p_1 (1- p1)/n1$ and $Var(\hat{P}_2) = p_2 (1- p_2)/n_2$.

Because of the independence of the two samples, we can write

$$E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2 \text{ with Var } (\hat{P}_1 - \hat{P}_2) = p_1 (1- p_1)/n_1 + p_2 (1- p_2)/n_2.$$

Now by the Central Limit Theorem, the random variable Z given by

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

has approximately a standard normal distribution with mean 0 and variance 1, i.e. $Z \cong N(0, 1)$. Hence

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha,$$

where Z is as given above. With little algebra manipulations we reach at the $100(1- \alpha)$ % confidence interval on the difference between the two proportions, $(p_1 - p_2)$ as given by the two limits

Lower Limit: $(\hat{P}_1 - \hat{P}_2) - Z_{/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$,

Upper Limit: $(\hat{P}_1 - \hat{P}_2) + Z_{/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$.

EXAMPLE 3.12

A certain change in a manufacturing procedure for component parts is being considered. Samples were taken using both the old and the new procedures in order to determine the difference induced by the new procedure. Suppose that 75 out 1500 items, from the existing procedure, were found defective, and 80 out 2000 items, from the new procedure were found to be defective. Find a 90% C.I. on the true difference in the fraction of defectives between the two procedures.

Solution:

Let p_1 and p_2 be the true proportions of defectives in the existing and new procedures, respectively. Thus we have $x_1 = 75$ and $x_2 = 80$, with $n_1 = 1500$ and $n_2 = 2000$. With $1 - \alpha = 0.90$, we see that $Z_{0.05} = 1.645$, using the Standard Normal Table, and with the above limits on the difference between the fraction of defectives between the two procedures we found that the 90% C.I. is given as

$$-0.0017 < p_1 - p_2 < 0.0217$$

❖ ----------

Notice as the interval contains zero, there is no reason to believe that the new procedure reduces the proportion of defectives.

3.6.2 Confidence Interval about the difference between Two Means

It is quite often the following question is raised: Which average of those two means which are under investigation is better, or higher or smaller, or worse? In comparative experiments the investigator wishes to estimate the difference between two processes based on the difference between their means. For example, a chemist likes to compare the effects of two catalysts on the output of some chemical reactions. An agronomist wants to estimate the difference in yield of two varieties of corn. These questions and many others lead us to investigate how to estimate the difference between two means based on finding the confidence interval about that difference.

Assume that there are two populations with their means and variances given $_i$ and ^2_i for i =1, 2 respectively. These populations could be normally distributed or not, as the discussion will reveal the cases below. We will select two simple random samples of sizes n_i, i = 1, 2, and denote them by X_j and Y_j, j = 1,2, …. n_i. Based on the data, from the samples, we can compute the mean and the variance for each sample, which are given by $\overline{X}$, S_1^2 and $\overline{Y}$ and S_2^2. There are four cases to be considered. Each case will be addressed based on the data and the information on hand.

Case I: The two population variances, σ_i^2 for i = 1, 2, are known.

We know from earlier discussion that $\overline{X}$ and $\overline{Y}$ are normally distributed each has a normal distribution with mean and variance given by μ_1, σ_1^2 / n_1 and μ_2, σ_2^2 / n_2 respectively, and thus the random variables

$$Z_1 = \frac{\overline{X} - \mu_1}{\sqrt{\sigma_1^2 / n_1}}, \text{ and } Z_2 = \frac{\overline{Y} - \mu_2}{\sqrt{\sigma_2^2 / n_2}},$$

each will have a standard normal distribution with mean 0 and variance 1. Because of the independence of the two samples we have $\overline{X} - \overline{Y}$ as the point estimate for $\mu_1 - \mu_2$ which it has a normal distribution with mean equal to $\mu_1 - \mu_2$ and variance given by $\sigma_1^2 / n_1 + \sigma_2^2 / n_2$, and thus the random variable

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Has a standard normal distribution with mean 0 and variance 1. Hence we have

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

Now as it was the case when we derived the C.I. on the difference between two proportions, and with little algebra manipulation we reach at the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ to be given by the following two limits

Lower Limit: $(\overline{X}-\overline{Y}) - Z_{\alpha/2}\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$

Upper limit: $(\overline{X}-\overline{Y}) + Z_{\alpha/2}\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$.

In other words, the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ is given by

$$(\overline{X}- \overline{Y}) - Z_{\alpha/2}\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2} < \mu_1 - \mu_2 < (\overline{X}- \overline{Y}) + Z_{\alpha/2}\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}\,.$$

It is to be noted here that the above Confidence interval applies for any values for the sample sizes.

EXAMPLE 3.13

Data was collected to compare the wear of two different materials. The summary came up to be

Sample	Mean	Size	population Standard deviation
I	85	12	5
II	81	10	4

Calculate the 95% C. I on the difference between the two means.

Solution:

Using the above formula for the C.I., when the populations variances are know, we see that we get the following interval:

$$(0.238, 7.762).$$

There is clear evidence that $_1 - \ _2 > 0$.

Case II: The two population variances σ_i^2 for i = 1, 2, are unknown, Large Sample Sizes.

In this case the question that will be asked is: What are the sample sizes? For sample sizes of greater than 30 each, by using the Central Limit Theorem, and replacing the population variances by their estimates, from the sample we see that the random variable given by

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2 / n_1 + S_2^2 / n_2}},$$

has a standard normal distribution. Hence the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ will be given by

$$(\bar{X} - \bar{Y}) - Z_{\alpha/2} \sqrt{S_1^2 / n_1 + S_2^2 / n_2} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + Z_{\alpha/2} \sqrt{S_1^2 / n_1 + S_2^2 / n_2}.$$

EXAMPLE 3.14

Consider the above example, EXAMPLE 3.13, but with larger samples, and sample standard deviations as given below:

Sample	Mean	Size	Sample Standard deviation
I	85	32	5
II	81	30	4

Calculate the 95% C. I on the difference between the two means.

Solution:

Since the sample sizes are larger than 30, we will consider the above C.I. interval with the following results for the limits:

(1.753, 6.247)

It is to be noted that the interval shifted to the right and it is shorter than the one before.

❖ ----------

Case III: The two populations variances, σ_i^2 for i = 1, 2, are unknown, Small sample sizes.

We will assume in this case that the populations, from which the two samples are randomly drawn, are normally distributed with means μ_1 and μ_2. Again there is a question to be asked about the two population variances: Are they equal or unequal? When the two population variances are equal, we can pool the samples' variances to estimate the common value. This value is termed **the pooled Variance**, and it is given by

$$S_{pooled}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}.$$

In this case we have a new random variable given by

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{pooled}\sqrt{1/n_1 + 1/n_2}},$$

That has a student t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. Thus the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ will be given by

$$(\bar{X} - \bar{Y}) - t_{\alpha/2} . S_{pooled}\sqrt{1/n_1 + 1/n_2} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + t_{\alpha/2} . S_{pooled}\sqrt{1/n_1 + 1/n_2}.$$

On the other hand, when the two population variances are not equal, we still have a student t-distribution for the random variable T given by

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

In this case the degrees of freedom, ν will be calculated from the following formula, and it is rounded down to the nearest whole number,

$$= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{(n_2 - 1)}}.$$

In this case the 100(1 – α) % C.I., on the difference $\mu_1 - \mu_2$, will be given by

$$(\bar{X} - \bar{Y}) - t_{/2}\sqrt{S_1^2/n_1 + S_2^2/n_2} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + t_{/2}\sqrt{S_1^2/n_1 + S_2^2/n_2}.$$

EXAMPLE 3.15

Consider the above example, EXAMPLE 3.13, but with small samples, and sample standard deviations as given below:

Sample	Mean	Size	Sample Standard deviation
I	85	12	5
II	81	10	4

Calculate the 95% C. I on the difference between the two means, by

a) Pooling for the common variance of the two populations,

b) Not pooling.

Solution:

a) In the pooled case we have this interval

$$(\bar{X} - \bar{Y}) - t_{\alpha/2} . S_{pooled} \sqrt{1/n_1 + 1/n_2} < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + t_{\alpha/2} . S_{pooled} \sqrt{1/n_1 + 1/n_2} .$$

By using the data we get the 95% C.I. on $_1 - _2$ to be given by

(-0.088, 8.088),

With df = 20, the pooled standard deviation = 4.577.

b) In the non –pooled we have to use this interval

$$(\bar{X}-\bar{Y})-t_{/2}\sqrt{S_1^2/n_1+S_2^2/n_2} < {}_1-{}_2 < (\bar{X}-\bar{Y})+t_{/2}\sqrt{S_1^2/n_1+S_2^2/n_2}\,.$$

The degrees of freedom are given by $=\dfrac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{(n_1-1)}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{(n_2-1)}}$ Based on the interval cited and the data provided, we have these limits for the 95% C.I.

(-0.0036, 8.0036).

Note: The above calculations for the 3 examples: 3.13, 3.14, and 3.15, were done using a TI Calculator 83-Plus.

❖ ----------

Remarks on the confidence Interval about two means

When constructing a confidence interval about the difference between two means, we must verify that the samples are large enough ($n \geq 30$) or the samples should come from populations that are normally distributed so that we can use the normal model for the construction of the confidence interval. Fortunately, the procedures for constructing confidence intervals presented above are **robust**, which means that minor departures from normality will not seriously affect the results. Verifying normality assumption for small sample sizes will be checked by drawing normal probability plot and checking for outliers by drawing a box plot.

Case IV: The samples are paired, or dependent, from two populations.

In the previous three cases the 100(1 – α) % C.I., on the difference $_1 - _2$, was constructed on the assumption that we have two independent samples from two different populations. Here we will investigate the case when the two samples are dependent, or related in any way. Sometimes the comparative experiment is conducted in a different way: pairs of similar individuals or objects are selected randomly. A common application occurs in self-pairing in which a single individual is measured on two occasions, usually, before and after some treatment. For example, the blood pressure of a subject might be measured before and after a heavy exercise; the weight of a person is measured before and after being on a certain diet. The analysis of paired samples is treated differently from that of independent samples, and the investigator has to be aware of the samples are paired or independent. The two dependent samples will be reduced to one sample of paired data, i.e. the sample of the differences of the data points between the two samples. An advantage of pairing data is resulted in the assumption of equal population variances is not needed any more. Based on this discussion, we will derive the Confidence interval for the difference between two population means when the samples are dependent.

Let (X_i, Y_i), i = 1,2,…., n form the paired observations of the two random samples. The assumption here is that the X's have come from a population with mean N (μ_1, $_1^2$) and the Y's have come from another population that is N (μ_2, $_2^2$). For the analysis, as mentioned earlier, we form the sample of differences $d_i = X_i - Y_i$, i= 1, 2, …, n. Thus we have

$$_d = E(D) = E(X - Y) = _1 - _2,$$

$$\bar{d} = (1/n)\sum_{i=1}^{n} d_i, \text{ and}$$

$$S_d^2 = \{1/(n-1)\}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

It follows that the random variable given by the formula below will have a student t-distribution with n-1 degrees of freedom.

$$T = \frac{\bar{d} - (_1 - _2)}{S_d/\sqrt{n}}.$$

Hence the 100(1 – α) % C.I., on the difference $_D = _1 - _2$, will be given by

$$\bar{d} - t_{/2} S_d/\sqrt{n} < _d < \bar{d} + t_{/2} S_d/\sqrt{n}.$$

EXAMPLE 3.16

Consider the following set of paired data, where it is assumed that the differences are normally distributed. Compute a 95% confidence interval about the population mean difference $D = {}_2 - {}_1$.

X_i	7.6	7.6	7.4	5.7	8.3	6.6	5.7
Y_i	8.1	6.6	10.7	9.4	7.8	9.0	8.5
d_i	0.5	-1.0	3.3	3.7	-0.5	2.4	2.8

Solution:

We are taking $d_i = Y_i - X_i$, and by applying the above techniques we get $\bar{d} = 1.075$ and $S_D = 1.915$. Here n is 7, $t._{025} = 2.447$, and by applying the above formula for the C.I., we find that it is given by (-0.16, 3.39)

3.6.3 Confidence Interval about the Ratio between Two Variances

It is desired sometimes to compare two processes or procedures based on their variability. A company, for example, sells frozen shrimp may have two methods of packaging. The two procedures give an average content of 12 ounces, but the packaging of one method seems to be more variable than the other. In this case the comparison of the two methods will be based on the variances of each.

It is quite clear that the difference between variances is not suitable to apply here. The procedure to obtain a confidence interval to measure the variability between the two methods will rely on the ratio of the two population variances. Thus we will proceed as follows.

Consider the following two independent random samples X_i , i = 1, 2, …, n_1 , and Y_j , j = 1, 2, …, n_2 that were taken from two normal populations that have μ_1 and μ_2 as their mans with ${}^2_1$ and ${}^2_2$ as their variances respectively. In other words, we have $\mathbf{X}_i \approx N(\mu_1, {}^2_1)$, i = 1, 2, n_1, and $\mathbf{Y}_j \approx N(\mu_2, {}^2_2)$.

It is known that the following variables defined by

$$U = \frac{(n_1 - 1)S_1^2}{{}^2_1}, \text{ and } V = \frac{(n_2 - 1)S_2^2}{{}^2_2},$$

are independent and each has a χ^2 distribution with $r_1 = (n_1 - 1)$ and $r_2 = (n_2 - 1)$ degrees of freedom respectively. Then the random variable defined by

$$F = \frac{U / r_1}{V / r_2},$$

has an F-distribution with r_1 and r_2 degrees of freedom. For brevity, we say that F is F (r_1, r_2). Clearly, the reciprocal of F,

$$\frac{1}{F} = \frac{V / r_2}{U / r_1},$$

Must be F (r_2, r_1), because U and V (and the associated degrees of freedom $r_1$and r_2) have exchanged roles. We can use the F-distribution to find a confidence interval for σ_1^2 / σ_2^2. With the setting above, we see that the random variable given by

$$F = \frac{\sigma_1^2 S_2^2}{\sigma_2^2 S_1^2}$$

is F (r_2, r_1) and therefore

$$P\ [F\ (1- \alpha/2, r_2, r_1) < F < F\ (\alpha/2, r_2, r_1)] = 1 - \alpha.$$

In other words, we have the following interval

$$F(1- \alpha/2, r_2, r_1) \leq \frac{\sigma_1^2 S_2^2}{\sigma_2^2 S_1^2} \leq F(\alpha/2, r_2, r_1).$$

Since F (1- α/2, r_2, r_1) = 1/ F (α/2, r_1, r_2), the inequality can be written as

$$\frac{1}{F(\quad/2, r_1, r_2)}\frac{S_1^2}{S_2^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} F(\quad/2, r_2, r_1).$$

By taking the square root of every term in the above inequality, we can have a confidence interval for the ratio of two standard deviations, as well.

EXAMPLE 3.17

A standard and a new method of teaching statistics are being compared with respect to their variability as measured by the final examination scores produced by the two methods. One class is taught by each method, and both classes take the same final examination. The observed sample variances are $S_1^2 = 100$, and $S_2^2 = 144$, where the first data coming from the standard method and the second is from the new method. If the classes contained 121 and 61 students respectively, find a 95% C.I. for $\frac{\sigma_2^2}{\sigma_1^2}$.

Solution:

Since we are finding the C.I. on $\frac{\sigma_2^2}{\sigma_1^2}$, the above interval becomes as follows

$$\frac{S_2^2}{S_1^2}\frac{1}{F(1-\alpha/2, r_2, r_1)} \le \frac{\sigma_2^2}{\sigma_1^2} \le \frac{S_2^2}{S_1^2} F(1-\alpha/2, r_1, r_2).$$

By the substitution the values for $r_2 = 60$, $r_1 = 120$, $S_1^2 = 100$, $S_2^2 = 144$, with F (0.975, 60, 120) = 1.5299, and F (0.975, 120, 60) = 1.5810, in the above inequality we get

$$0.941 < \frac{\sigma_2^2}{\sigma_1^2} < 2.277.$$

(The values of F-distribution for different values of the degrees of freedom for the numerator and denominator are cumbersome and need a lot of space and selection.)

❖

CHAPTER 3 EXERCISES

3.1 The percentage of copper in a certain chemical element is measured 6 times. The standard deviation of repeated measurements in such an analysis is known to be 2.5%. The sample mean is 14.1%. Construct a 95% C.I. for the true percentage of copper, assuming that the observations are approximately normally distributed.

3.2 25 measurements are made on the speed of light. Those averaged to 300007 with an SD of 10, the units being in Kilometers per second. Report your estimate of the speed of light as a 95% C.I. (1 Km = (5/8) mile).

3.3 A laboratory has a method for measuring lengths, using modern laser technology. The operator's job is to calibrate a yardstick. Measurements were taken 25 times, resulting in an average of 0.910835 meters, with a standard deviation of 45 microns (a micron is one millionth of a meter). Find an approximate 95% C.I. for the exact length of this stick. (Using this modern laser technology, the length can be measured to within one wave length of visible light which is about half a micron.)

3.4 An investigator made 10 measurements of a metric standard and obtained an average of 1.0002 meters, with a standard deviation of 0.0001 meters. Construct a 90% C.I. for the exact length.

3.5 The weight of v7 similar containers of sulfuric acid is: 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 ounces. Find an 85% C.I. for the mean of all such containers assuming an approximate normal distribution.

3.6 An efficiency expert wishes to determine the average time it takes to drill three holes in a certain clamp. How large a sample will he need to be 95% confident that his sample mean will be within 15 seconds of the true mean? Assume that it is known from previous studies that sigma is 40 seconds.

3.7 A random sample of 8 cigarettes of a certain brand has an average nicotine content of 1806 milligram and a standard deviation of 2.4 milligram. Construct a 99% C.I. for the true average of nicotine content of this particular brand of cigarettes.

3.8 A random sample of 100 families from a large city is chosen to estimate the current average annual demand for milk in that city. The mean family demand from the sample is 150 gallons with a standard deviation of 40 gallons.

a) Construct a 95% C.I. for the mean annual demand of milk by all families in the city.
b) If the rage you obtained in a) is larger than you are willing to accept, in what way can you narrow it?

3.9 In a part of a large city in which houses were rented, an economist wishes to estimate the average monthly rent correct to within US$50, a part from a 1-in-20 chance. If he guesses from past experience that sigma is about US$40, how many houses must he include in his sample?

3.10 The yield of alfalfa fro 9 plots were 0.8, 1.3, 1.5, 1.7, 1.7, 1.8, 2.0, 2.0, and 2.2 tons per acre. Set a 95% C.I. for the true average yield.

3.11 A manufacturer of batteries guarantees them to last for a specified period of time and wants to know how much variability there is in the lifetime of the batteries. A sample of 20 batteries was tested for longevity and S2 was found to be 53 hours. Suppose that the lifetimes are normally distributed, estimate the true variability in the life time as a 99% C.I.

3.12 Determine the point estimate of the population mean and margin of error for each of the following confidence intervals: a) Lower bound: 18, upper bound: 24, b) Lower bound: 20, upper bound: 30; c) Lower bound: 5, upper bound: 23, d) Lower bound : 15, upper bound: 35.

3.13 A simple random sample of size n is drawn from a population whose standard deviation, sigma, is known to be 3.8. The sample mean is determined to be 59.2.

a) Compute a 90% C.I. for the true population mean if the sample size n, is 45.
b) Compute a 98% C.I. for the true population mean if the sample size n, is 45. Compare the results to those obtained in part a). How does increasing the level of confidence affect the size of the margin of error E?
c) Compute a 90% C.I. for the true population mean if the sample size n, is 55. Compare the results to those obtained in part a). How does increasing the sample size affect the margin of error E?
d) Can we compute a C.I. for the true population mean based on the information given if the sample size is n=15? Why? If the sample size is n=15, what must be true regarding the population from which the sample was drawn?

3.14 An electrical engineer wishes to estimate the variation in the amount of heat generated by a certain type of electronic component in order to design an appropriate heat dissipater for it. He took a sample of 16 components and observed the following units of heat generated: Find a 95% confidence interval for the true variation in heat generation.

4.260, 3.882, 4.741, 3.897, 4.925, 4.021, 4.822, 4.113, 4.628, 4.013, 4.728, 4.224, 4.171, 4.585, 4.509, 4.419

3.15 The following are the weights, below, in ounces, of 10 packages of grass seeds distributed by a certain company: Find a 95% confidence interval for the variation in all such packages distributed by this company.

16.9, 15.2, 16.0, 16.4, 16.1, 15.8, 17.0, 16.1, 15.9, 15.8

3.16 The vitamin C concentration (in mg per 100 gm) in a sample of size 17 of canned orange juice is: 16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, and 25. Find a 90% C.I. for the true variation in Vitamin C concentrations.

3.17 In a sample of 31 patients, the amount of an anesthetic required to produce anesthesia suitable for surgery was found to have a standard deviation (from patient to patient) of 10.2 mg. Compute a 98% confidence interval on the population standard deviation.

3.18 Five out of 50 randomly selected time sharing terminals give incorrect character response. A firm has 800 of these terminals.

a) Estimate the proportion of terminals that give incorrect response.
b) eport your estimate as a 95% confidence interval on the true population proportion.

3.19 A manufacturer of flashcubes wants to estimate the probability that a flashcube will work. Since, destructive testing is involved; he wants to keep the sample size as small as possible. Find the number of observations that must be taken to estimate the probability within 0.04 and with 95% confidence of that if

a) He has no idea of the percent defective.
b) He believes that the percent defective is no more than 6%.

3.20 A public Library wants to estimate the percentage of books in its collection that have publication dates of 1970 or earlier. How large a random sample must be taken to be 90% sure of coming within 5% of the actual proportion?

3.21 Five out of 20 randomly selected employees working for a large institution indicate that they are unhappy with their kind of work.

a) Report an estimate of the proportion of unhappy employees as 90% C.I.
b) If there were 50 employees in the institution, convert the interval in part a) to the number of employees who are unhappy.

3.22 A marketing research group found that 25% of the 200 shopper, it recently interviewed at a certain shopping center, resided more than 12 miles from the center. Assume that a random sample was taken, construct a 95% C.I. for the actual; percentage of shoppers who live more than 15 miles from that center.

3.23 Two catalysts are being compared for their effect on the output of a chemical process. A sample of 12 batches is prepared using catalyst 1 and a sample of 10 batches was obtained using catalyst 2. The 12 batches for which catalyst 1 was used gave an average yield of 85 with standard deviation of 4, while the average for the second sample was 81 with a standard deviation of 5. Estimate the difference between the true averages as a 90% confidence interval, assuming equal variances.

3.24 Records for the last 15 years have shown that the average rainfall in a certain region of the country, for the month of March, to be 1.20 inches, with s = 0.45 inches. A second region had an average rainfall of 1.35 inches, with s = 0.54. estimate the difference of the true average rainfalls in those two regions as a 95% C.I. with the assumption of normal populations and unequal variances.

3.25 Two varieties of corn are being compared for yield. Fifty acres of each are planted and grown under the similar conditions. Variety A yielded, on the average, 78.3 bushels per acre with s = 5.6 bushels per acre. Variety B yielded, on the average, 87.2 bushels per acre with s =6.3 bushels per acre. Estimate the difference in the average yields as a 95% confidence interval by assuming equal populations variances.

3.26 The following are the rates of diffusion of carbon dioxide through two soils of different porosity.

Fine Soil: 20, 31, 18, 23, 23, 28, 23, 26, 27, 26, 12, 17, 25
Course Soil: 19, 30, 32, 28, 15, 26, 35, 18, 25, 27, 35, 35.

Estimate the difference, in the two rates, as a 95% confidence interval, a) by assuming equal variability, b) by assuming unequal variability.

3.27 The total nitrogen content of the blood plasma of normal albino rats was measured at 37 and 180 days of age. The results are expressed as gm. per 100 cc of plasma. The results are given below:

At age 37 days, there were 9 rats with: 0.98, 0.83, 0.99, 0.86, 0.90, 0.81, 0.94, 0.92, and 0.87

At age 180 days, there were 8 rats with: 1.20, 1.18, 1.33, 1.21, 1.20, 1.07, 1.13, and 1.12

Set a 95% confidence interval on the difference between the two populations' means, a) Assuming equal variances, b) Assuming unequal variances. C) Comment on the two intervals.

3.28 Two samples of seedlings were grown with different fertilizers. The first sample, with 200 seedlings had an average height of 10.9 inches and a sample standard deviation of 2.0 inches. The second sample, with 100 seedlings had on average height of 10.5 inches with a standard deviation of 3.0 inches. Compute the confidence interval on the difference between the two population means by assuming equal variability.

3.29 An experiment was performed on ten college students to determine the effect of a certain drug on the span of their concentration. The data below represents the time, in minutes, of concentration before fatigue. Estimate the difference as a 99% Confidence interval of the concentration time.

Student	1	2	3	4	5	6	7	8	9	10
Before Drug	45	32	58	59	60	44	47	51	42	38
After Drug	47	34	60	57	63	38	49	53	46	41

3.30 It is claimed that a new diet will reduce a person's weight by 5 kg, on the average, in a period of two weeks. The weights of 7 women who followed this diet and recorded their weights before and after the diet are as shown below. Report your estimate on the difference between the means based on a level of 95%.

Woman	1	2	3	4	5	6	7
Wt. Before	65	67	67	76	70	69	62
Wt. After	66	60	64	68	65	66	60

3.31 Five people selected at random had their breathing capacity measured before and after a certain treatment, obtaining the following data. Based on this data, estimate the difference in the mean capacity of before and after treatment with level of confidence of 95%.

Person	1	2	3	4	5
Before	2750	2360	2950	2830	2250
After	2850	2380	2930	2860	2320

3.32 A certain change in a manufacturing procedure for component parts is being considered. Samples are taken using the existing procedure and the new one. If 75 items out of 1500 items, from the existing procedure, were found to be defective, while 80 items out of 2000 items for the new procedure were found defective. Find a 90% confidence interval for the true difference in the fraction defective between the existing and the new procedures.

3.33 A geneticist is interested in the proportion of males and females in the population that have a certain minor blood disorder. In a random sample of 1000 males, 250 were found to be afflicted, whereas 275 out of 1000 females tested appear to have the disorder. Compute a 95% C.I. on the difference between the proportion of males and females that have the blood disorder.

3.34 In a study of morphological variation in natural populations of fruit fly, it was reported that the mean wing length of 16 females, collected in a certain area, was 4.653 mm with s = 0.012 mm; and the mean wing length of 11 females, collected in a second area, was 4.274 mm with s = 0.02. Let the distribution of the wing length be normal, find a 98% confidence interval on the ratio of the two populations' standard deviations.

3.35 A standardized placement test was given to 10 boys and 11 girls who applied for a certain job. The boys made an average grade of 82 with s = 8, while the girls made an average grade of 78 with s = 7. Estimate the ratio of variability between the boys and the girls as a 90% confidence interval.

3.36 A researcher wishes to estimate the proportion of adults who have high-speed Internet access. What sample size should be obtained if the researcher wishes to estimate within 0.03 with 99% confidence if

a) He uses a 2007 estimate of 0.69 obtained from US Census Bureau?
b) He does not use any prior estimate?
c) Let the confidence level be 90%, and recalculate the needed sample size in both a) and b) above and comment on your conclusion.

RANDOM NUMBERS TABLE

Row	Column Number									
Number	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

TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP Obtaining a Simple Random Sample

TI-83/84 Plus

8. Enter any nonzero number (the seed) on the HOME screen.
9. Press the *sto* > button.
10. Press the *MATH* button.
11. Highlight the PRB menu and select 1: rand.
12. From the **HOME** screen press ENTER.
13. Press the MATH button. Highlight PRB menu and select 5: randInt (.
14. 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

4. 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.**
5. Select **TOOLS** and highlight **Data Analysis….**
6. 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
   Number of Random Numbers: 10
   Distribution: Uniform
   Parameters
   Between 1 grid 10
   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 **Confidence Intervals about μ, σ known**

TI-83/84 Plus

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 7: Z-Interval.
3. If the data is raw, highlight DATA. Make sure List1 is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. Following sigma, enter the population standard deviation.
4. Enter the confidence level following C-Level.
5. Highlight Calculate; press ENTER.

Excel

1. Enter raw data in column A. Highlight the data.
2. Load the data Desk XL. Add-in.
3. Select the **DDXL** menu. Highlight **Confidence Intervals**, and then highlight **1Var z Interval** from the dropdown menu.

4. Select the column of data from the "Names and Column" window. Check" The First Row is Variables names", if appropriate. Use the < arrow to select the data. Click OK.
5. Select the level of confidence and enter the value of the population standard deviation. Click Compute Interval.

TECHNOLOGY STEP-BY-STEP **Confidence Intervals about μ, σ Unknown**

TI-83/84 Plus

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 8: T-Interval.
3. If the data is raw, highlight DATA. Make sure List1 is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. Following sigma: enter the sample standard deviation.
4. Enter the confidence level following C-Level.
5. Highlight Calculate; press ENTER.

Excel

1. Enter raw data in column A. Highlight the data.
2. Load the data Desk XL. Add-in.
3. Select the **DDXL** menu. Highlight **Confidence Intervals**, and then highlight **1Var t-Interval** from the dropdown menu.
4. Select the column of data from the "Names and Column" window. Check "The First Row is Variables names", if appropriate. Use the < arrow to select the data. Click OK.
5. Select the level of confidence and Click Compute Interval.

TECHNOLOGY STEP-BY-STEP **Confidence Intervals about p**

TI-83/84 Plus

1. Press STAT; highlight TESTS, and select A: 1-PropZInt.
2. Enter the values of x and n.
3. Enter the confidence level following C-Level.
4. Highlight Calculate; press ENTER.

Excel

1. Enter raw data in column A. Highlight the data.
2. Load the Data Desk XL Add-in.
3. Select the **DDXL** menu. Highlight **Confidence Intervals**, and then highlight **1 Var Prop Interval** if you have raw data or **Sum 1 Var Prop Interval** if you have summarized data.

4. For raw data, Select the column of data from the "Names and Column" window. Check "The First Row is Variables names", if appropriate. Use the < arrow to select the data. Click OK. For summarized data, click the pencil icon and enter the number of trials and the number of successes. Click OK.
5. Select the level of confidence and Click Compute Interval.

TECHNOLOGY STEP-BY-STEP **Confidence Intervals about σ**

TI-83/84 Plus The TI-83/84 Plus do not construct confidence interval about σ

Excel Excel does not construct confidence intervals about Excel

TECHNOLOGY STEP-BY-STEP

Confidence Intervals about $_2 - _1$, known population Variances

TI-83/84 Plus

1. If necessary, enter raw data in L_1, L_2.
2. Press STAT, highlight TESTS, and select 7: 2- Sample Z-Interval.

3. If the data is raw, highlight DATA. Make sure List1 is set to L_1 and, and List2 is set to L_2, and freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. Following sigma, enter the population's standard deviations, then the samples means, and sizes.
4. Enter the confidence level following C-Level.
5. Highlight Calculate; press ENTER.

TECHNOLOGY STEP-BY-STEP

Confidence Intervals about $_2 - _1$, unknown population Variances, large samples

TI-83/84 Plus

1. If necessary, enter raw data in L_1, L_2.
2. Press STAT, highlight TESTS, and select 7: 2- Sample Z-Interval.
3. If the data is raw, highlight DATA. Make sure List1 is set to L_1 and, and List2 is set to L_2, and freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. The samples standard deviations takeover for the populations standard deviations in this case when the sample sizes are large, following sigma, enter the samples standard deviations in this case.
4. Enter the confidence level following C-Level.
5. Highlight Calculate; press ENTER.

TECHNOLOGY STEP-BY-STEP

Confidence Intervals about $_2 - _1$, unknown population Variances, small samples

TI-83/84 Plus

1. If necessary, enter raw data in L_1, L_2.
2. Press STAT, highlight TESTS, and select 7: 2- Sample T-Interval.
3. If the data is raw, highlight DATA. Make sure List1 is set to L_1 and, and List2 is set to L_2, and freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. The sample standard deviations takeover for the population standard deviations in this case when the sample sizes are large, following sigma, enter the samples standard deviations in this case.
4. Enter the confidence level following C-Level.
5. Select Pooled: No or Yes, based on your choice, there will a difference in the degrees of freedom between the pooled and the non-pooled case, as you will notice.
6. Highlight Calculate; press ENTER.

TECHNOLOGY STEP-BY-STEP **Confidence Intervals about $P_1 - P_2$**

1. Press STAT, highlight TESTS, and select 7: 2- propZInt.
2. Enter the values of x_1, n_1, x_2 and n_2.
3. Select a confidence Level
4. Highlight Calculate and press enter.

TECHNOLOGY STEP-BY-STEP **Comparing Two Population Standard Deviations**

TI-83/84 Plus The TI-83/84 Plus do not construct confidence interval about two the ratio of two population standard deviations

4 Testing of Statistical Hypotheses

Outline

4.1 Introduction

Testing a statistical hypothesis is the second main and major part of inferential statistics. A statistical hypothesis is an assumption or a statement, about one or two parameters and involving one or more than one population. A statistical hypothesis may or may not be true. We need to decide, based on the data in a sample, or samples, whether the stated hypothesis is true or not. If we knew all the members of the population, then it is possible to say with certainty whether or not the hypothesis is true. However, in most cases, it is impossible, and impractical to examine the entire population. Due to scarcity of resources, lack of time, and tedious calculations based on a population, we can only examine a sample that hopefully represents that population very well. So the truth or falsity of a statistical hypothesis is never known with certainty.

Testing a statistical hypothesis is a technique, or a procedure, by which we can gather some evidence, using the data of the sample, to support, or reject, the hypothesis we have in mind.

4.2 Fundamental Concepts

Any field, and statistics is not an exception, has its own definitions, concepts and terminology. These items make the basic building stones in any subject. Knowing these three things, and connecting among them, will make the subject coherent, and more at the will of the reader. Based on this, we like to present some definitions, and terminology for some concepts that will be used in the text.

A) The Null and Alternative Hypotheses

The first step, in testing a statistical hypothesis, is to set up a null hypothesis and an alternative hypothesis. When we conjecture a statement, about one parameter of a population, or two parameters of two populations, we usually keep in mind an alternative conjecture to the first one. Only one of the conjectures can be true. So, in essence we are weighing the truth of one conjecture against the truth of the other. This idea is the first basic principle in testing a statistical hypothesis. For example, an experimenter may think that a newly discovered drug is either as effective as, or better than, a currently used one. The experimenter wants to weigh the truth of the hypothesis that the new drug is as effective as the old drug against the hypothesis that the new drug is actually better than the old one. In statistical terminology, the first hypothesis is called the "Null Hypothesis", i.e. no change, no effect, or no difference, and it is denoted by H_0, H_{Naught} or H_{zero}. The second hypothesis is called the "Alternative hypothesis", i.e., there is a change, and it is denoted by H_a or H_1, and it will be a statement regarding the value of a population parameter. In this text, we will be using H_0 and H_1 as our Null Hypothesis and Alternative Hypothesis respectively.

B) Possible Decisions

The test procedure will lead to either one of the following decisions:

1. Reject the Null Hypothesis, H_0, i.e., conclude that H_0 is a false statement and this will lead to take, or accept, that the alternative hypothesis H_1 as a true statement.
2. Do not reject the Null hypothesis, H_0. This means that there is no evidence from the sample, to disprove the null hypothesis. The non-rejection of H_0 should not imply that it is true. This is because the objective of testing a statistical hypothesis is to disprove, or reject, the null hypothesis with a high certainty, rather than to prove it. Thus if the statistical test rejects H_0 then we are highly certain that it is false. However, if the test does not reject H_0, then we interpret the non-rejection as that the sample does not give enough evidence to disprove the null hypothesis. In other words, the rejection of the null hypothesis is the decisive conclusion that we can depend on.

Based on the decision, whether to "Reject H_0" or "Do Not Reject H_0", we should be careful in stating the null and alternative hypotheses. This is due to the fact that originally we have two statements to be examined against each other and we may call either one of them the null hypothesis. But since we are only highly confident about the conclusion of rejecting the null hypothesis, we take H_0 as the statement that the sample will reject. On the other hand, the alternative hypothesis will be that statement which we hope that the data will support.

In the drug example above, the experimenter wants to prove that the new drug is better than the old one. So, the experimenter wants to disprove the statement that the new drug is as effective as the old one. Based on that, he should set the hypotheses as

H_0: The new drug is as effective as the old one.
H_1: The new drug is more effective than the old drug.

C) Types of Errors

The procedure, in testing a statistical hypothesis, either rejects the null hypothesis or not. Of course the truth is never known, i.e. we do not know whether H_0 is true or not. The "true state of nature" may then be that H_0 is true or H_0 is false. We make the decision of rejecting the null hypothesis or not rejecting it, without knowing the true state of nature. In making a decision about testing a statistical hypothesis, two types of errors may be committed:

Type I Error: A Type I error has been committed if the test rejects the null hypothesis when in fact it is true. The probability of making such an error will be denoted by α, (The Greek letter Alpha). For sure, it is clear that $0 \leq \alpha \leq 1$.

Type II Error: A Type II error has been committed if the test does not reject H_0 when H_0 is false. The probability of making such an error will be denoted by β, (the Greek letter Beta) with $0 \leq \beta \leq 1$. What is more important is that we do not like to make such errors with high probabilities.

In either one of the other two cases, there is no error committed, as shown by Table I.

		Nature State of H_0	
		TRUE	FALSE
Based on Data	Reject H_0	Type I Error	Correct Decision
	Do Not Reject H_0	Correct Decision	Type II Error

Table 1

Each type of error has a certain probability of being committed. These probabilities are given specific names, and values, due to their importance and the severity of the decision.

Level of Significance: The probability of committing a type I error is denoted by (Alpha) α. It is called the theoretical level of significance for the test. The most common used values for α are: 0.01, 0.05, or 0.10. Other values for α are at the discretion of the researcher. More expressions for α are:

α = P (committing a type I error),
= P (rejecting H_0 when H_0 is true),
= P (rejecting H_0 when H_1 is false).

The probability of committing a type II error is denoted by (beta) β. Other labels for β are:

β = P (committing a type II error),
= P (not rejecting H_0 when H_0 is false),
= P (not rejecting H_0 when H_1 is true).

Power of the Test: The value of $1-\beta$, which stands for P (rejecting H_0 when H_0 is false) is the power of the test.

The probabilities of type I and Type II errors tell us how good the test is. Clearly we do not like to make a type I error with high probability, as well as we like to make a correct decision with a very high power. The smaller these probabilities (of type I error and Type II error) are the better is the test. Ideally, we like to use a test procedure for which both type I and type II errors have small probabilities. However, it turns out that the type I error and the type II error are related in such a way that a decrease in the probability of one of them generally results in an increase in the probability of the other. It is not possible to control both probabilities based on a fixed sample size. Traditionally, or by convention, statisticians have adopted to fix the level of significance of the test in advance, and to search for a test procedure that will minimize the probability of making a type II error, and consequently maximize the power for the test. Practically, we can make those probabilities, namely α and β, smaller by taking a larger sample if possible. For calculating the probability of type II error, and the power of the test, we direct the interested reader to Sullivan 2013.

D) The Test Statistic

The test statistic is a quantity that depends on the information, or statistics, that the sample will provide. It is a function of the sample statistics, and the value(s) of the parameter(s) under the null hypothesis. Thus a statistic is a random variable until we get some values from the sample. The numerical value of the test statistic (large or small) leads us to decide whether or not to reject the Null Hypothesis when it is compared to the critical value(s) of the test.

E) The Critical Region or The Rejection Region (CR or RR)

The critical region is an interval, or a union of intervals, which is determined by using special and certain distributions with the appropriate Table values. It depends on the distribution of the test statistic when H_0 is true, on the form of the alternative Hypothesis, and on the level of significance that was set for the test.

F) Conclusion and interpretation

The final conclusion, of the test procedure, is based on whether or not the computed value of the test statistic falls inside the critical region, or not, as follows:

1. Reject H_0 if the computed value of the test statistic falls in the critical region.
2. Do not reject H_0 if the computed value of the test statistic does not fall inside the critical region.

In either of the two cases detailed above, an interpretation and a practical statement are due in order to answer the question that was raised before the test procedure started.

4.3 Methods in Testing a Statistical Hypothesis

There are two methods to test a statistical hypothesis, namely The Classical or Traditional method, and the P-value method. Both of these methods will be introduced and used in this text. Based on the notation and definitions that were set above, we will list the steps in the Classical method, in general, first and then the steps for the p-value method next. More detailed steps will be outlined later based on the parameter, or parameters, involved, or stated in the hypotheses.

I Classical Method Steps

1. Determine, and clearly, state the two hypotheses: H_0 and H_1. Equality to the assigned parameter should be included under the Null Hypothesis.
2. Decide on the significance level α. Find the critical value or values, and locate the rejection region or regions (all based on the parameter and distribution under consideration).
3. Choose the appropriate Test statistic for the hypotheses based on the parameter on hand.
4. Using the information provided by the data in the sample, and the computed statistics, calculate the test statistic that was chosen in Step 3.
5. Make your statistical decision, whether to reject, or not to reject, H_0 based on the comparison between the computed value of the test statistic and the critical value(s) found in Step 2, and as outlined earlier.
6. Give the conclusion, or the answer, in a statement that anyone can understand without any mathematical jargons or statistical ambiguity.

II. P-value Method Steps

1. Determine and clearly state the two hypotheses: H_0 and H_1. Equality to the assigned parameter should be included under the Null Hypothesis.
2. Decide on the significance level α.
3. Choose the appropriate Test statistic for the hypotheses based on the parameter on hand.
4. Using the information provided by the data in the sample, and the computed statistics, calculate the test statistic that was set up in Step 3.
5. Make your statistical decision, whether to reject, or not to reject, H_0 based on the comparison between the theoretical significance level α, (that was set up above) and the calculated p-value. (This p-value is the practical, or attained, significance level, based on the type of the test and the distribution of the parameter involved). A p-value less than α will lead to the rejection of H_0, otherwise do not reject H_0.
6. Give the conclusion, or the answer to the question, in a statement that anyone can understand without any mathematical jargons or statistical ambiguity.

The above steps will be applied to test on one parameter or two parameters whether the test was two tailed test or one tailed, left or right test. In the next section we will introduce the test of a statistical hypothesis on one parameter. The one parameter case will involve; one proportion, one mean and one standard deviation, or one variance.

4.4 Hypothesis Testing About One Parameter

In this section we will discuss, and display, the procedures for testing a statistical hypothesis about one population parameter. The one population parameter, which is of interest, will include: one proportion, one mean, and one variance or a standard deviation. In general, let that parameter be θ, and its assumed value be θ_0. Following the steps that were set earlier, we will give the steps in more details for the case when one proportion is under investigation.

It is to be noted here that the two hypotheses are mutually exclusive sets on the real line for the values of the parameter, with the equal sign always set to go with the null hypothesis. This is so chosen in order to compute the value of the test statistic based on the null hypothesis being true.

Thus the steps will go as follows.

4.4.1 Hypothesis Testing about One Proportion, P

Recall that the best point estimate of p, the proportion of the population with a certain characteristic, is given by

$$\hat{p} = x/n,$$

where x is the number of individuals in the sample with the specified characteristic of interest and n is the sample size. Recall from **Chapter 3** that the sampling distribution of $\hat{p}$ is approximately normal, with mean $_{\hat{p}} = p$ and standard deviation

$$_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

In addition to the above two criteria, the following requirements should be satisfied:

1. The sample is a simple random sample
2. $np(1 - p) \geq 10$.
3. The sample values are independent of each other.

A) Classical Method Steps

For a specified value of the proportion P_0 we have (We are using the z-test on one proportion):

1. State the null and alternative hypotheses:
   There are three ways to set up the null and alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $P = P_0$ versus H_1: $P \neq P_0$, two-tailed test.
   b) At least versus less than: H_0: $P \geq P_0$ versus H_1: $P < P_0$, left-tailed test.
   c) At most versus greater than: H_0: $P \leq P_0$ versus H_1: $P > P_0$, right-tailed test.

2. Let α (the most used values for the level of significance are: 0.01, or 0.05, or 0.10) be the significance level. Based on the three cases in step 1, we have the following three cases that will go along with that:

a) For the two tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$, as shown in the **Figure 1** to include both areas in this case each will be $\frac{\alpha}{2}$.

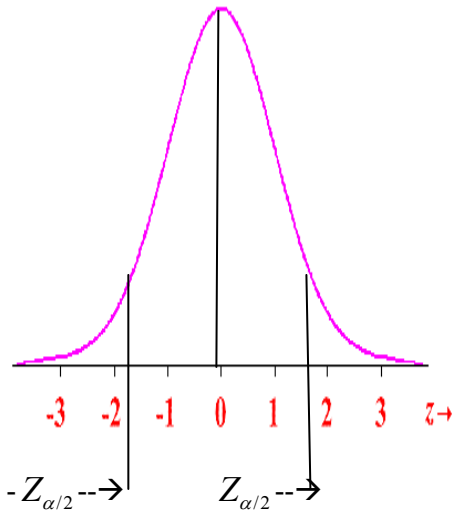


Figure 1

b) For the left-tailed test, the critical value is $-Z_\alpha$, and the rejection region is given by $Z < -Z_\alpha$, as shown in the **Figure 2.** The area to the left of $-Z_\alpha$ is α.

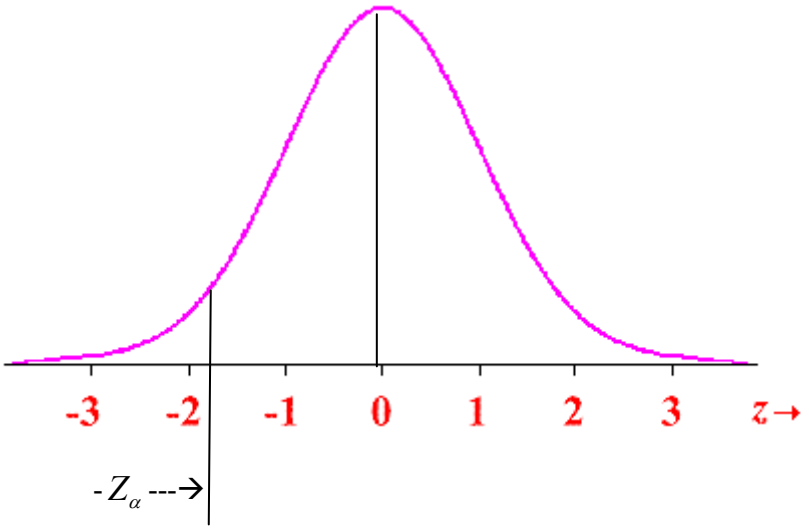


Figure 2

c) For the-right tailed test, again, there is one critical value given by Z_α, and the rejection region is $Z > Z_\alpha$, as shown in **Figure 3.** The area to the right t of Z_α is α.

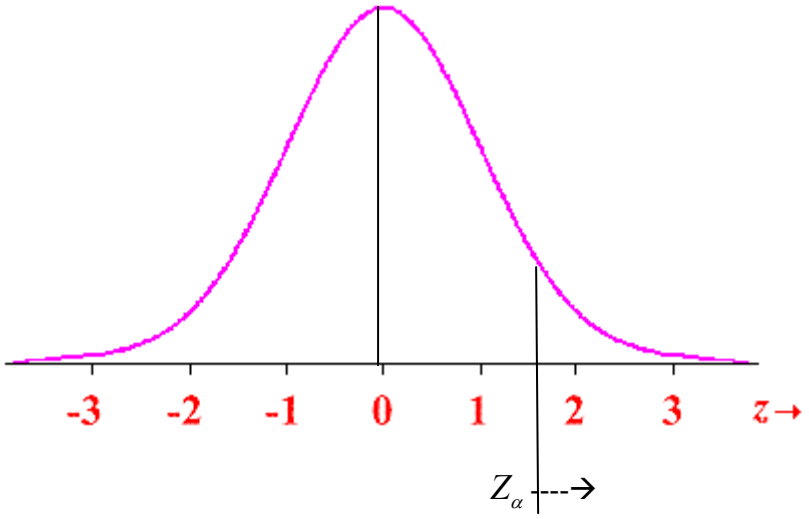


Figure 3

3. For the test statistic we have $Z = \frac{\hat{p} - p_0}{\sqrt{[p_0(1-p_0)/n]}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$, where n is the sample size and $\hat{p}$ = x/n.
4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed, or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

The above steps are the road map for testing a statistical hypothesis on one proportion using the classical or traditional method. Here is an example.

EXAMPLE 4.1

The government of a wealthy country intends to institute a program to discourage investment in foreign countries by its citizens. It is known that in the past 35% of the country's adult citizens held investment in foreign countries. The government wishes to determine if the current percentage of adult citizens, who own foreign investment is greater than this long term figure of 35%. A random sample of 800 adults is selected, and it is found that 320 of these citizens hold foreign assets. Is this percentage greater than 35%? Use a 10% significance level for testing this claim.

Solution:

Using the setup above for the classical method on testing a statistical hypothesis on one proportion we proceed as follows:

1. H_0: P ≤ 0.35 Versus H_1: P > .35. This is a right–tailed test. (It is 1-Prop Z test on TI-84).
2. The level of significance is given to be 10%, or 0.10. Thus $\alpha = 0.10$. Since the test is one-tailed, on the right side, we have one critical value given by: C.V. = $Z_\alpha = Z_{0.10} = 1.282$ and the rejection region is given by: Z > 1.282.

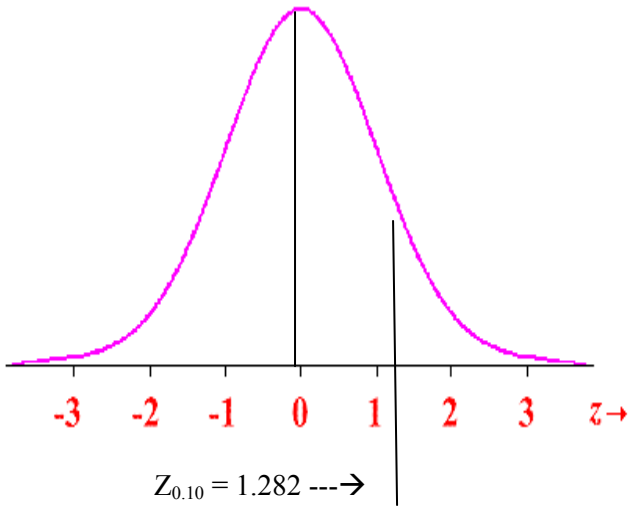


Figure 4

3. The test statistic is $Z = \frac{\hat{p} - p_0}{\sqrt{[p_0(1-p_0)/n]}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$, where n = 800, x = 320, and $P_0 = 0.35$.
4. From the above values, and form for the test statistic, we find $Z_{cal} = 2.965$ where $\hat{P} = 0.40$.
5. Since $Z_{cal} = 2.965 > C.V. = Z_{0.10} = 1.282$, we reject H_0.
6. It is concluded that the percentage of adult citizens, who own investment in a foreign country, is greater than 35%.

❖

B) The P-value Method:

For a specified value of the proportion P_0 and (We are using the z-test on one proportion) for testing on one proportion using the P-value Method, the steps go like this (the steps almost look like what we had on the classical Method except there is a difference between step 2 and step 5):

1. State the null and alternative hypotheses:
   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $P = P_0$ versus H_1: $P \neq P_0$, Two-tailed test
   b) At least versus less than: H_0: $P \geq P_0$ versus H_1: $P < P_0$, Left-tailed test
   c) At most versus greater than: H_0: $P \leq P_0$ versus H_1: $P > P_0$, right-tailed test

2. Let α (= 0.01, or 0.05, or 0.10, or any other value of your choice) be the significance level.
3. For the test statistic we have $Z = \dfrac{\hat{p} - p_0}{\sqrt{[p_0(1-p_0)/n]}} = \dfrac{x - np_0}{\sqrt{np_0(1-p_0)}}$, where n is the sample size and $\hat{p}$ = x/n.
4. The above test statistic is computed based on the information provided to us by the sample data.

5. How to calculate the p-value for your test? Based on the test type stated in Step 1, we have the following options:

   a) For the 2-tailed test, and after calculating the value of the test statistic, Zcal in step 4, the p- value is found by: p-value = 2 *P ($Z < Z_{cal}$), or p-value = 2 *P (Z > Zcal), conditioned on whether the Zcal is negative or positive. The statistical decision will be made based on comparing the computed p-value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.
   b) For the left-tailed test, and after calculating the value of the test statistic, Zcal in step 4, the p-value is found by: p-value = P ($Z < Z_{cal}$). The statistical decision will be made by comparing the computed p- value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.
   c) For the right-tailed test, and after calculating the value of the test statistic, Zcal in step 4, the p-value is found by: p-value = P ($Z > Z_{cal}$). The statistical decision will be made by comparing the computed p- value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.

6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.2

Apply the P-value method to check on the test in EXAMPLE **4.1.**

Solution:

Using the setup above for the p-value method on testing a statistical hypothesis on one proportion we proceed as follows:

1. H_0: P ≤ 0.35 Versus H_1: P > .35. This is a right–tailed test. (It is 1-Prop Z test on TI 84).
2. The level of significance is given to be 10%, or 0.10.
3. The test statistic is $Z = \frac{\hat{p} - p_0}{\sqrt{[p_0(1-p_0)/n]}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$, where n = 800, x = 320, P_0 = 0.35.
4. From the above values and form for the test statistic we find Z_{cal} =2.965 where $\hat{P} = 0.40$.
5. Since Z_{cal} = 2.965. Since we have a right-tailed test, then the p-value for the test will be calculated by finding P (Z > 2.965). This is done by applying step No. 5. Part c) in the p-value method steps. Using the table for standard normal distribution we find ourselves trapped in rounding to two decimal places.

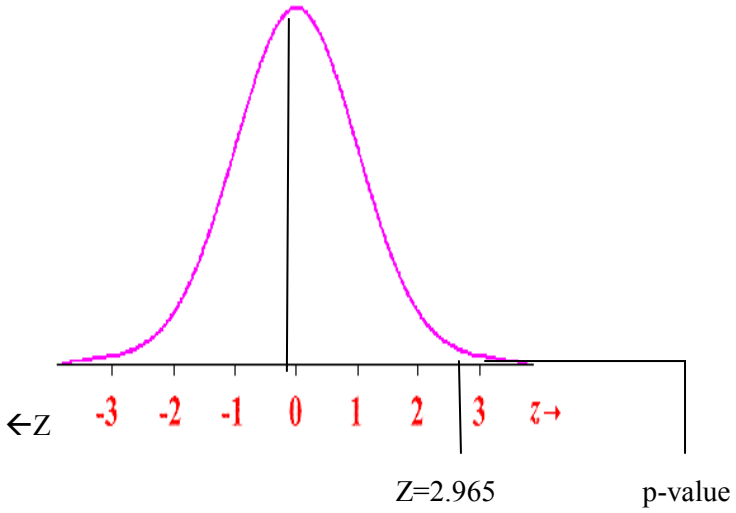


Figure 5

First, let us take Z_{cal} = 2.97. Based on that we see then P (Z > 2.97) = 1–P (Z < 2.97), and from the Standard Normal Table, we have P (Z > 2.97) = 1–0.9985 = 0.0015 < 0.10, hence The Null hypothesis is rejected, See Figure 2.

Second, let us take Z_{cal} = 2.96. Based on that we see then P (Z > 2.96) = 1–P (Z <2.96), and from Standard Normal Table, we have P (Z > 2.96) = 1–0.9985 = 0.0015 < 0.10, hence The Null hypothesis is rejected. In this case it did not make a difference whether you rounded up or down the calculated value for the test statistic.

Using a graphing calculator, and testing the same hypothesis, we find that the p-value, to 4 decimal places is, again, 0.0015. Thus the null hypothesis is rejected.

6. It is concluded that the percentage of adult citizens, who own investment in a foreign country, is greater than 35%.

❖

4.4.2 Hypothesis Testing about One Mean, μ

This section will display the procedure, by using the two methods outlined above for Testing a statistical hypothesis about one population mean. There are three cases to be considered in this section. Again, as it was stated above for one proportion, it is to be noted here that the two hypotheses are mutually exclusive sets on the real line for the values of the parameter, with the equal sign always set to go with the null hypothesis. This is so chosen in order to compute the value of the test statistics based on the null hypothesis being true.

Case I: Testing on One Mean, μ when Population variance, σ^2 known

Classical Method Steps: For a specified value of the population mean μ_0 we have (We are using the z-test):

1. State the null and alternative hypotheses.

   There are three ways to set up the null and alternative Hypotheses:

   a) Equal hypothesis versus not equal hypothesis: H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, two-tailed test
   b) At least versus less than: H_0: $\mu \geq \mu_0$ versus H_1: $\mu < \mu_0$, left-tailed test
   c) At most versus greater than: H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$, right-tailed test

2. Let α be the significance level, and based on the three cases in step 1, we have the following three cases that will go along for finding the critical values and rejection regions:

a) For the two-tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$, as shown in the **Figure 6**, with the area on each is $\alpha/2$.

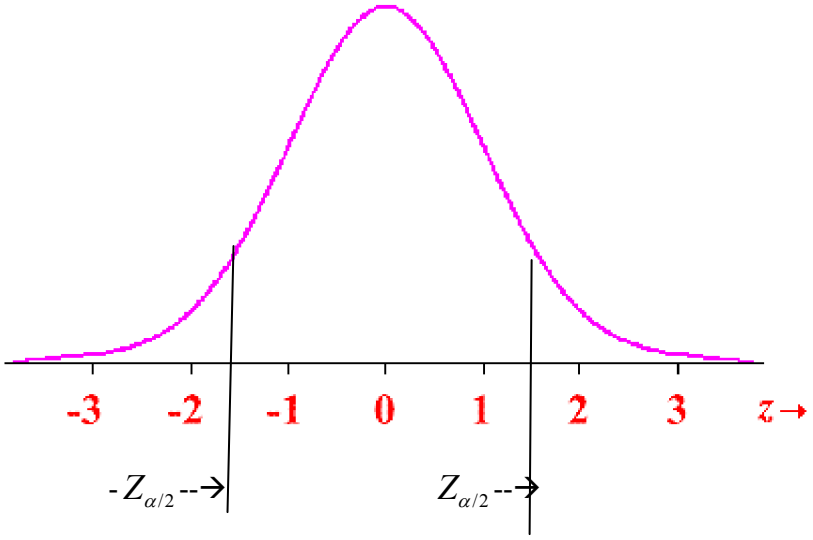


Figure 6

b) For the left tailed test, there is the critical value of $-Z_{\alpha}$, and the rejection region is given by $Z < -Z_{\alpha}$, as shown in the **Figure 7**, with the area to the left of $-Z_{\alpha}$ is equal to α.

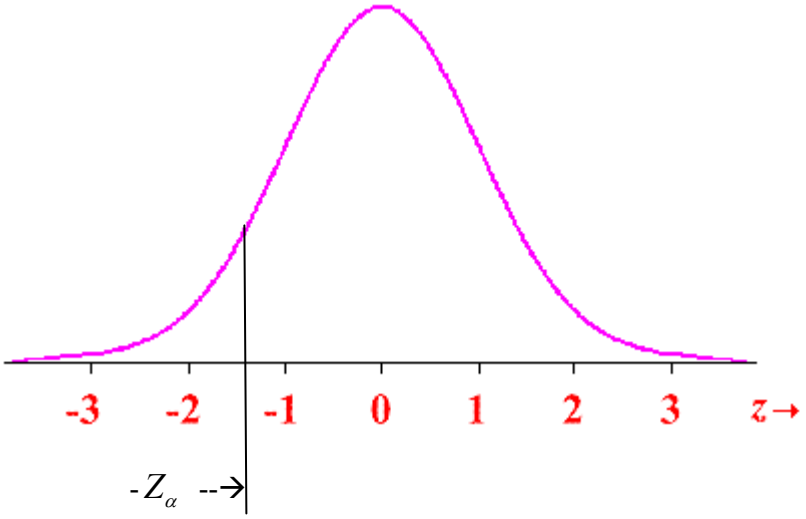


Figure 7

c) For the right tailed test, again, there is on critical value given by Z_{α}, and the rejection region is $Z > Z_{\alpha}$, as shown in the **Figure 8**, with the area to the right ofZ_{α} is equal to α.

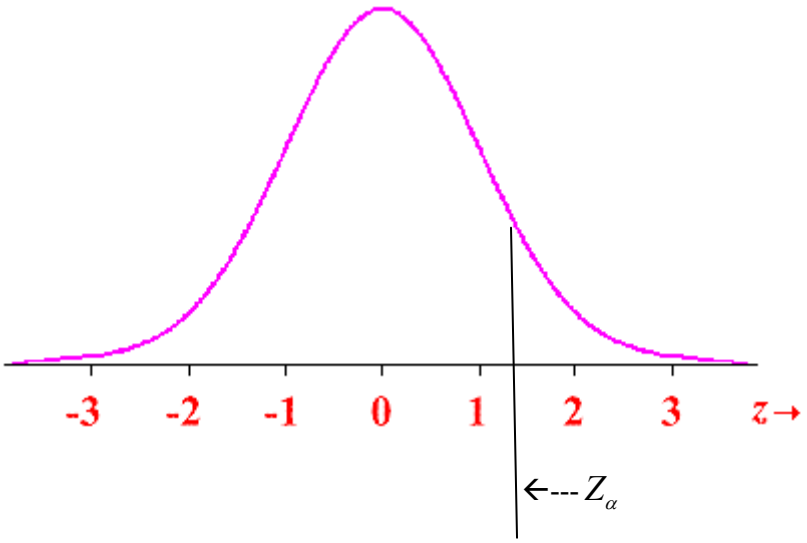


Figure 8

3. The test statistic is given by $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, where n is the sample size, $\bar{x}$ is the mean, and Z has the standard normal distribution, N (0, 1).
4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.3

To test H_0: $\mu \geq 50$ versus H_1: $\mu < 50$, a random sample of n = 24 is obtained from a population that is known to be normally distributed with $\sigma = 12$, and we got a sample mean of 47.1. Will the null hypothesis be rejected?

Solution:

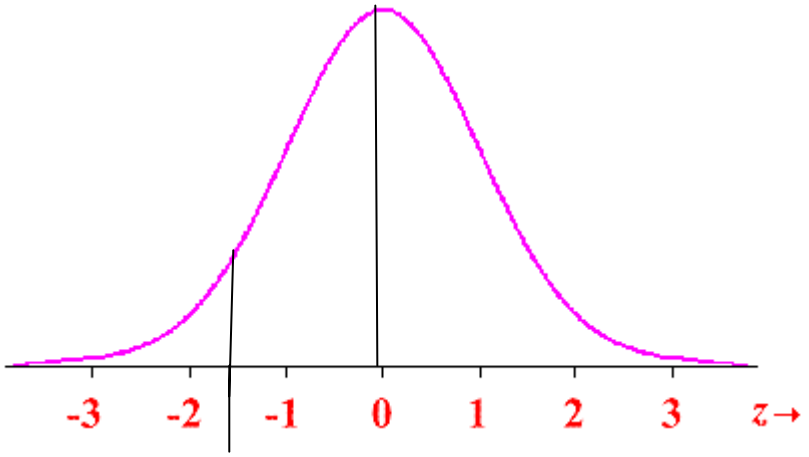
Applying the classical Method steps for test on one mean we have

1. State the null and alternative hypotheses.

H_0: $\mu \geq 50$ versus H_1: $\mu < 50$, left-tailed test

2. Let $\alpha = 0.05$ be the significance level.

   a) For the left tailed test, there is the critical value of $-Z_{0.05} = -1.645$, and the rejection region is given by $Z < -1.645$, as shown in the Figure 1, the green area. The green area now is equal α.



CV $-Z_{0.05} = -1.645$ and R.R. $Z < -1.645$

Figure 9

3. The test statistic is given by $Z = \frac{\bar{x} - \quad _0}{/\sqrt{n}}$,

4. The above test statistic is computed as $Z = \frac{47.1 - 50}{12/\sqrt{24}}$, based on the information provided to us by the sample data. Thus Z = -1.1839

5. Since the calculated value of the test statistic, namely -1.1839, does not fall in the rejection region, then H_0 is not rejected.
6. The conclusion is that μ is not less than 50. Therefore $\mu \geq 50$.

❖

P-value Method Steps: For a specified value of the population mean μ_0 we have (We are using the z-test on one mean) for testing on one proportion the steps go like this:

1. State the null and alternative hypotheses :
   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, Two-tailed test
   b) At least versus less than: H_0: $\mu \geq \mu_0$ versus H_1: $\mu < \mu_0$, Left-tailed test

   c) At most versus greater than: H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$, right-tailed test

2. Let α be the significance level for the test.
3. For the test statistic we have $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, where n is the sample size.
4. The above test statistic is computed based on the information provided to us by the sample data.
5. Apply how the p-value is calculated based on the type of your test, as shown above. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed p-value, for the test, to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.
6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.4

Using the information in Example 4.3, test the above hypothesis there by the P-value method.

Solution:

As in the steps we do not need to find a critical value for this method. The significance level was given to be 0.05.

Following the steps as if it were the classical method, we find that the test statistic has the value of -1.1839. Let us find the p-value for this left-tailed test, see **Figure 10**

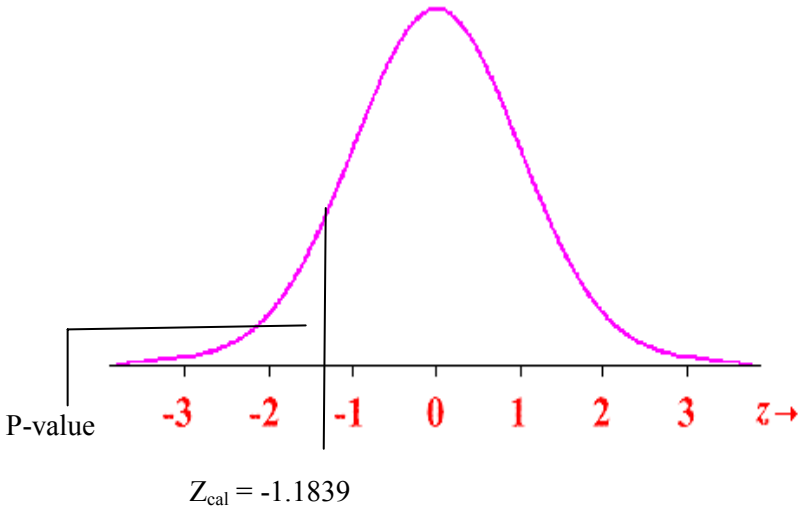


Figure 10

The p-value = P (Z < -1.18) for using the Standard Normal Table, we get P-value = 0.1190. It is greater than Alpha. We do not reject the null hypothesis. Hence we conclude that μ is not less than 50, therefore $\mu \geq 50$.

It is to be recalled that the two methods used above lead to the same conclusion. In case there is a contradiction between them, i.e. if you reject the Null hypothesis by using one of them while you did not reject the Null Hypothesis by using the other method. It is for sure you have made a mistake in one of them. Check it again.

❖

Case II: Testing on One Mean, μ When Population Variance, σ^2 Unknown

Since the population variance, σ^2 is not known, it is traditionally reasonable to ask about the sample size. This is based on the earlier presentations done in Chapter 3, when we compare the standard normal distribution with the Student's t-distribution. We found out there that when n, the sample size is large, usually n ≥ 30, is suitable to use the standard normal distribution for the test statistic involving one mean. Based on this discussion we have two cases to consider.

A) large Sample Size

In this part, since the sample size is large, n ≥ 30, we will use the same procedure for testing a statistical hypothesis about one population mean with one change. That change will take place in calculating the test statistic Z, when the standard deviation of the population is replaced by that of the sample. Based on that, the test statistic will be given by

$$Z = \frac{\bar{x} - {}_0}{s/\sqrt{n}}.$$

The test that will be used is the Z-Test. The steps, in the classical method and the P-value method, are the same as above for this case of testing about one mean when the population variance is unknown.

B) Small Sample Size

In this part, since the sample size is small, n < 30, we will use the same procedure for testing a statistical hypothesis about one population mean with one change. That change will take place in replacing Z as the test statistic with T, where T will have a student t-distribution with degrees of freedom ν = n-1. Based on that, the test statistic will be given by

$$T = \frac{\bar{x} - {}_0}{s/\sqrt{n}}.$$

The t-distribution is another continuous distribution that is widely used in statistics. Because of that let us describe that distribution before getting to use it here.

In case I, above, we discussed testing a statistical hypothesis when the population variance was known. We now have another case on hand when the population variance, or the standard deviation, is unknown, and we have a small sample. The Z-Test discussed in Case I and Case II A) does not apply any more. We have to appeal for another distribution. This distribution is the t-distribution. In this case we do not replace σ by s anymore, and say that

$$T = \frac{\bar{x} - {}_0}{s/\sqrt{n}}$$

is normally distributed, with mean 0 and variance 1. Instead $T = \frac{\bar{x} - {}_0}{s/\sqrt{n}}$, and this random variable follows **Student's t-distribution** with n-1 degrees of freedom. So, let us have the properties of the t-distribution as listed below.

1. The t-distribution is controlled by its degrees of freedom. It is different for different degrees of freedom.
2. The mean of the distribution is 0, and it is symmetric about its mean.
3. As it was the case with the standard normal distribution, the total area under the curve is 1.
4. The horizontal axis acts like a horizontal asymptote, i.e., as t increases (or decreases) without any bound; the graph approaches the horizontal axis but never intersects it.
5. Compared with the standard normal distribution, and if drawn on the same scale, we find that the peak for standard normal distribution is higher than that of the t-distribution. This makes the tails for the t-distribution thicker than those for the standard distribution.
6. The variance for the t-distribution is > 1.
7. As the number of degrees of freedom increases (i.e. as the sample size n increases) the t-distribution gets closer to Z, the standard normal distribution. In this case the two curves for the two distributions will look almost alike. That is, and based on the law of large numbers, the estimator S, of σ, gets closer and closer.

When testing regarding one mean with the population variance unknown, and a small sample, the steps in the classical and the p-value methods look almost the same as in the cases discussed above, but for more clarity and consistency we will list the steps here again.

Classical Method Steps: For a specified value of the population mean μ_0 we have (We are using the T-test on one mean):

1. State the null and alternative hypotheses:

   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, two-tailed test
   b) At least versus less than: H_0: $\mu \geq \mu_0$ versus H_1: $\mu < \mu_0$, left-tailed test

   c) At most versus greater than: H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$, right-tailed test

It is to be noted here that the two hypotheses are mutually exclusive sets on the real line for the values of the parameter, with the equal sign always set to go with the null hypothesis. This is so chosen in order to compute the value of the test statistics based on the null hypothesis being true.

2. Let α be the significance level for the test. Based on the three cases in step 1, we have the following three cases that will go along, for finding the critical value(s)and the rejection region(s)

   a) For the two tailed test there are two critical values: $-t_{\alpha/2}$ & $t_{\alpha/2}$, and the critical region is given by $|T| > t_{\alpha}$, as shown in the figure.
   b) For the left tailed test, there is the critical value of $-t$, and the rejection region is given by $T < -t_{\alpha}$, as shown in the figure.
   c) For the right tailed test, again, there is on critical value given by t_{α}, and the rejection region is $T > t_{\alpha}$, as shown in the figure.

3. For the test statistic we have $T = \frac{\overline{x} - \quad_0}{s/\sqrt{n}}$, where n is the sample size, s is the standard deviation of the sample, and T has the student t-distribution with n-1 degrees of freedom.

4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

In order to complete the picture, as before, let us list the steps in testing a statistical hypothesis regarding one mean when the population variance is unknown by using the p-value method.

P-value Method Steps For a specified value of the population mean μ_0, we have (We are using the T-test on one mean) the following steps need to be followed:

1. State the null and alternative hypotheses:

   There are three ways to set up the null and alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, Two-tailed test
   b) At least versus less than: H_0: $\mu \geq \mu_0$ versus H_1: $\mu < \mu_0$, Left-tailed test
   c) At most versus greater than: H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$, right-tailed test

2. Let α be the significance level for the test.
3. The test statistic is $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, where n is the sample size, s is the standard deviation of the sample, and T has the student t-distribution with n –1 degrees of freedom.
4. The above test statistic is computed based on the information provided to us by the sample data.

5. Apply how the p-value is calculated based on the type of your test, as shown above. In this case we need to use the t-value that was found in step 4, instead of the z-value, since we have a T-test now. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed p-value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.
6. The interpretation and conclusion are due to answer the question that was raised.

It is to be noted here that the p-value cannot be precisely found from the Tables as it as the case with Z-Test. By using the T-Table, we can put a range on the p-value only, while with technology the value will be given by the program used.

EXAMPLE 4.5

A colony of laboratory mice consisted of several hundred animals. Their average weight was believed to 30 gm. An experiment was conducted to check on this belief. A simple random sample of 25 animals was taken. The average weight for this sample turned up to be 33 grams with a sample standard deviation of 5 gm. What conclusion can be made using if the level of significance will be 5%?

Solution:

Using the classical method steps for testing on one mean, using the T-test, we have:

1. H_0: $\mu = 30$ versus H_1: $\mu \neq 30$, Two-tailed test.
2. It is assumed that the significance level is 0.05. Since we have a two-tailed test, we have the following critical values and the corresponding Rejection regions.

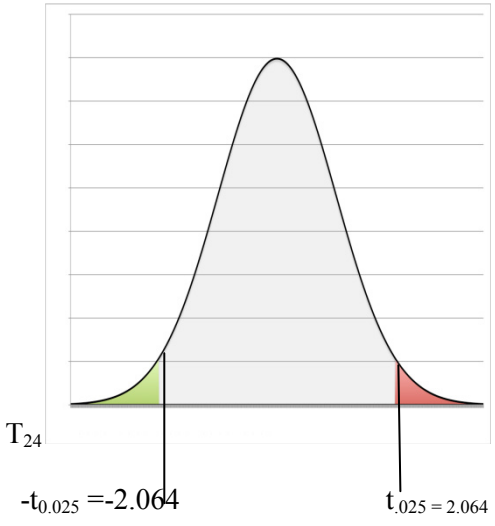


Figure 11

We see, from **Figure 11**, that the critical values are ± 2.064, and the rejection regions are given by |T| > 2.064.

3. The test statistic is $T = \dfrac{\overline{x} - {}_0}{s/\sqrt{n}}$.
4. Using the information given on hand, by calculating the above expression for T we have T = 3.
5. Since the value of the test statistic falls in the rejection region, we reject H_0.
6. Based on the data provided the average weight is > 30.

❖

4.4.3 Hypothesis Testing about One variance

Testing a statistical hypothesis about a population variance, or standard deviation, is surely different from testing about one population mean or proportion. The difference lies in the distribution of the statistic involved in the test. The statistic we are talking about here is the sample variance, or standard deviation, and its distribution. The test on one population variance is carried out based on the interest to check on the variability in the population. As it was the case in finding the confidence interval for one variance, we appealed to the random variable for the statistic given by

$$^2 = \frac{(n-1)S^2}{^2}.$$

This Random variable has a chi-squared distribution with n-1 degrees of freedom, and S^2 is the sample variance, σ^2 is the population variance, and n is the sample size. (See the properties of the Chi-squared distribution set up in **Chapter 3**).

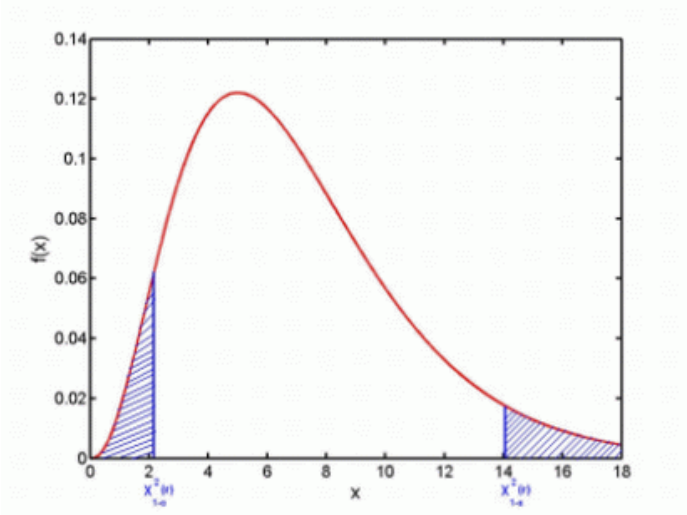


Figure 12

EXAMPLE 4.6

Let $\alpha = 0.05$, and n =15. Find the following values $\chi^2_{\alpha/2}$, $\chi^2_{1-\alpha/2}$, χ^2_{α} and $\chi^2_{1-\alpha}$.

Solution:

Since n = 15, then the degrees of freedom for the ^{2} distribution is given by v=14. And based on α=.05, then 1-α=0.95, α/2= 0.025, 1-α/2 = 0.975, from the ^{2}-Table, with 14 degrees of freedom as find that $^2_{0.025} = 26.119$, $^2_{0.975} = 5.629$, $^2_{0.05} = 23.685$, and $^2_{0.95} = 6.571$.

❖

When we tested on one population mean, with small sample size and unknown variance, using the t-test, we had to assume that the population has a normal distribution. For testing on one population variance we will need the same condition, i.e., the sampled population needs to have a normal distribution. As it was the case on one population mean or proportion, we can use either the classical or the p-value method, with the one restriction that we will consider only the right-tailed test. This is because we like to check on the large value of the variance, in other words, we like to check on a high variation in the population.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n from a normal population with mean μ and variance σ^2. The steps for the statistical test on σ^2, for a specified value of the population variance σ^2 are as follows:

Classical Method Steps:

1. State the null and alternative hypotheses:

   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\sigma^2 = \sigma_0^2$ versus H_1: $\sigma^2 \neq \sigma_0^2$, two-tailed test
   b) At least versus less than: H_0: $\sigma^2 \geq \sigma_0^2$ versus H_1: $\sigma^2 < \sigma_0^2$, left-tailed test
   c) at most versus greater than: H_0: $\sigma^2 \leq \sigma_0^2$ versus H_1: $\sigma^2 > \sigma_0^2$, right-tailed test

2. Let α be the significance level. Based on the three cases in step 1, we have the following three cases that will go along to calculate the critical values and the rejection regions.

   a) For the two tailed test there are two critical values: $^2_{1-\ /2}$ & $^2_{\ /2}$, and the critical regions are given by $^2 > Z_{\alpha/2}$, or $^2 < {}^2_{1-\ /2}$, as shown in Figure 13

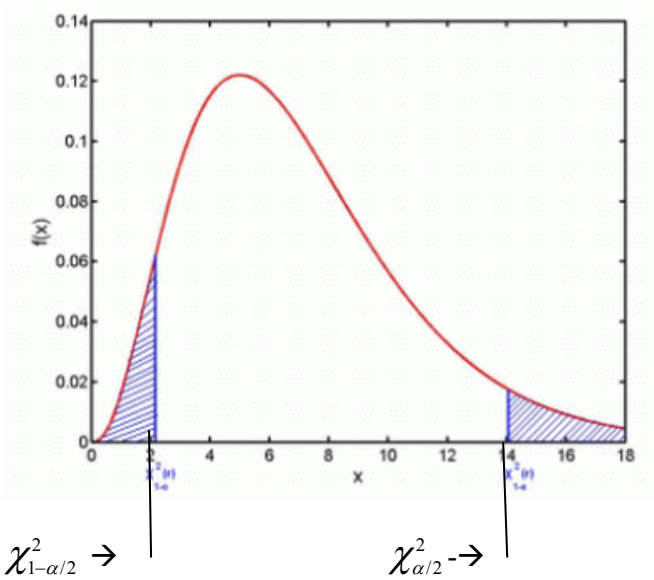


Figure 13

   b) For the left tailed test, there is the critical value of $^2_{1-\ }$, and the rejection region is given by $^2 < {}^2_{1-\ }$ as shown in the figure.
   c) For the right tailed test, again, there is on critical value given by ^{2}, and the rejection region is $^2 > {}^2$, as shown in the figure.

3. The test statistic is given by $^2 = \frac{(n-1)S^2}{^2_0}$, where n is the sample size, S^2 is the sample variance, and ^{2} has the Chi-Square distribution, with n-1 degrees of freedom.
4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

It is needless to say that the P-value method steps can be applied to test on one variance. The difficulty lies in finding the exact value of the p-value. As it was the case with t-test, there would be range of values on the p-value. However, if a software program is used, in either the T-Test on one mean, or the Chi-square test on one variance, the v p-value will be calculated exactly.

EXAMPLE 4.7

Consider a random sample of size 15 is taken from a normal population, that yielded $S^2 = 3$. Test whether $\sigma > 1$, by using a) the classical method, b) the p-value method, by taking the significance level of 0.05.

Solution:

a) Using the classical method steps for testing the hypothesis on one variance we have:

1. H_0: $\sigma^2 \leq 1$ versus H_1: $\sigma^2 > 1$, right-tailed test.
2. $\alpha = 0.05$, n = 15, ν =14, we see that the critical value is $^2_{0.05} = 23.685$, Thus the rejection region is given by $^2 > 23.685$.
3. The test statistic is $^2 = \frac{(n-1)S^2}{^2_0}$.
4. The calculated value of the test statistic is 126.
5. Since the calculated value of the test statistic is > 23.685, H_0 is rejected.
7. We conclude that $\sigma^2 > 1$ and thus $\sigma > 1$.

b) By using the P-value Method, we have

1. H_0: $\sigma^2 \leq 1$ versus H_1: $\sigma^2 > 1$, right-tailed test.
2. $\alpha = 0.05$
3. The test statistic is $^2 = \frac{(n-1)S^2}{^2_0}$.
4. The calculated value of the test statistic is 126.

5. We need to calculate the p-value. From the Chi-square table it is hard to find exactly how much the p-value is. Never the less we can put a range on the p value. How? Since the degrees of freedom = 14, and our test statistic value is 126, we go look along 14 for a number close to, or greater than, 126. Doing so we find the largest value in the table along 14 degrees of freedom is 31.319, which fall under 0.005, in the Table. Hence the P-value is $< 0.005 < 0.05$, and H_0 is rejected.
   On the other hand, using a graphing calculator, we find that the p-value will be given precisely as

6. We conclude that $\sigma^2 > 1$ and thus $\sigma > 1$.

❖

4.5 Hypothesis Testing Concerning Two Parameters

In this section we will discuss, and display, the procedures for testing a statistical hypothesis about two populations' parameters. In general, let those parameters be denoted by θ_1, and θ_2. Thus the steps go as follows:

4.5.1 Tests about two proportions, $P_1 - P_2$.

The populations' parameters which are of interest in this section are the populations' proportions. In section 4.4.1, we discussed inference regarding one population proportion. We will now tackle the question how to run a statistical hypothesis testing on two populations' proportions. To conduct inference about two population proportions, we must first determine the sampling distribution of the difference of two proportions. Recall that the best point estimate of p, the proportion of the population with a certain characteristic, is given by

$$\hat{p} = x/n,$$

where x is the number of individuals in the sample with the specified characteristic of interest and n is the sample size. Recall from **Chapter 3** that the sampling distribution of $\hat{p}$ is approximately normal, with mean $_{\hat{p}} = p$ and standard deviation

$$_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Provided that $np(1-P) \geq 10$. So

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is normally distributed, with mean 0 and standard deviation 1. Using this information along with the independence of the two samples taken from two different populations, we obtain the sampling distribution of the difference between two proportions. Let us set the sampling distribution for $P_1 - P_2$.

Suppose that two simple random samples were taken from two different populations with proportions of P_1 and P_2, for a certain property that we are interested in. The First sample is of size n1 that produced x individuals having that interesting characteristic, while sample 2 is of size n_2 that produced y individuals having the specified characteristic. Thus the sampling distribution of $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1 = \frac{x}{n_1}$, and $\hat{p}_2 = \frac{y}{n_2}$, is approximately normal with mean

$$_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

and standard deviation

$$_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

provided that the condition on each sample is being satisfied. Thus the standardized version of $\hat{p}_1 - \hat{p}_2$ is given by

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

This has an approximate standard normal distribution.

When comparing two population proportions, the null hypothesis will always have the equal sign with it i.e., $p_1 = p_2 = p$ where p is the common value for the population proportion. Based on this setting we can write the above z statistic in this form

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

$\widehat{p} = \frac{x+y}{n_1+n_2}$. The above test statistic will be used in the two methods for testing a statistical hypothesis about the difference between two population proportions. We start with the Classical Method.

Classical Method Steps: We are using the z-test on two proportions (2-prop Z-Test):

1. State the null and alternative hypotheses:
   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $P_1 = P_2$ versus H_1: $P_1 \neq P_2$, two-tailed test.
   b) At least versus less than: H_0: $P_1 \geq P_2$ versus H_1: $P_1 < P_2$, left-tailed test.
   c) At most versus greater than: H_0: $P_1 \leq P_2$ versus H_1: $P_1 > P_2$, right-tailed test.

2. Let α be the significance level. Based on the three cases in step 1, we have the following three cases that will go along for finding the critical values and rejection regions.

   a) For the two tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z|>Z_{\alpha/2}$, as shown in the **Figure 1.**
   b) For the left tailed test, there is the critical value of $-Z_{\alpha}$, and the rejection region is given by $Z < -Z_{\alpha}$, as shown in the **Figure 2.**
   c) For the right tailed test, again, there is on critical value given by Z_{α}, and the rejection region is $Z > Z_{\alpha}$, as shown in the **Figure 3.**

3. For the test statistic we have $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, as defined above.

4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

P-value Method Steps: We are using the z-test on two proportions (2-prop Z-Test):

1. State the null and alternative hypotheses:
   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $P_1 = P_2$ versus H_1: $P_1 \neq P_2$, two-tailed test.
   b) At least versus less than: H_0: $P_1 \geq P_2$ versus H_1: $P_1 < P_2$, left-tailed test.
   c) At most versus greater than: H_0: $P_1 \leq P_2$ versus H_1: $P_1 > P_2$, right-tailed test.

2. Let α(= 0.01, 0.05, 0.10) be the significance level.

3. The test statistic for this case, as defined above, is given by

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

4. The above test statistic is computed based on the information provided to us by the sample data.
5. How to calculate the p-value for your test? Based on the test type stated in Step 1. above, we have the following options:

a) For the 2-tailed test, and after calculating the value of the test statistic, Zcal in step 4, the p-value is found by: p-value = 2*P($Z < Z_{cal}$) or p-value = 2*P($Z > Z_{cal}$), conditioned on whether the Zcal that is negative or positive. The statistical decision will be made based by comparing the computed p- value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.

b) For the left-tailed test, and after calculating the value of the test statistic, Zcal in step 4, the p-value is found by: p-value = P ($Z < Z_{cal}$). The statistical decision will be made by comparing the computed p- value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.

c) For the right-tailed test, and after calculating the value of the test statistic, Zcal in step 4, the p-value is found by: p-value = P ($Z > Z_{cal}$). The statistical decision will be made by comparing the computed p- value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.

6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.8

In clinical trials of testing a certain drug, before it is released for the public, 3800 adults were randomly divided into two groups. The patients in Group 1 (Experimental group) received 200 mg of the drug, while the patients in group 2 (control group) received a placebo. Out of the 2100 patients in the experimental group, 550 reported headache as a side effect. Of the 1700 patients in the control group 370 reported headaches as a side effect. Is there significant evidence to support the claim that the proportion of the drug users that experienced headaches as a side effect is greater than the proportion in the control group at the $\alpha = 0.05$ level of significance.

Solution:

Using the conditions, and all requirements, to carry the test of a statistical hypothesis on the difference between two proportions, we have

1. The samples are independently obtained using simple random sampling
2. $\hat{p}_1 = \frac{x}{n_1} = \frac{550}{2100} = 0.2619$, $\hat{p}_2 = \frac{y}{n_2} = \frac{370}{1700} = 0.2176$.
3. Therefore $n_1\hat{p}_1(1-\hat{p}_1) = 2100.(0.2619).(1\text{-}0.26190) = 405.9476 \geq 10$, and
4. $n_2\hat{p}_2(1-\hat{p}_2) = 1700(0.2176)(1\text{-}0.2176) = 289.4254 \geq 10$.

Thus we proceed with the classical method using the 6 steps, and then we apply the p-value method second. So we have:

1. H_0: $P_1 \leq P_2$ versus H_1: $P_1 > P_2$, right-tailed test.
2. $\alpha = 0.05$ is the level of significance. The Critical value is $Z_{0.05} = 1.645$ and the rejection region is given by $Z > 1.645$.
3. The test statistic is $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
4. $Z_{cal} = 3.1668$, based on the data provided.
5. Since the test statistic falls in the rejection region, the null hypothesis is rejected, i.e., H_0: $P_1 \leq P_2$ is rejected, and H_1: $P_1 > P_2$ is being supported.
6. There is sufficient evidence at the $\alpha = 0.05$ level of significance to support the claim that the proportion of adults taking 200 mg of the drug who experienced headaches is greater than the proportion of adults taking a placebo who experienced headaches.

❖

Now we will apply the p-value method steps on testing the claim that was set up in EXAMPLE 4.8. The steps go like this.

1. State the null and alternative hypotheses:
   H_0: $P_1 \leq P_2$ versus H_1: $P_1 > P_2$, right-tailed test.

2. Let $\alpha = 0.05$, be the significance level.
3. The test statistic for this case, as defined above, is given by

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

4. $Z_{cal} = 3.1668$, based on the data provided.
5. How to calculate the p-value for your test? Based on the test type stated in Step 1, we have a right-tailed test, then the p-value is found by: p-value = P $(Z > Z_{cal})$ = P(Z > 3.1668) = 0.0007706. Since the computed p-value is less than α, H_0 is rejected, and H_1: $P_1 > P_2$ is being supported.
6. There is sufficient evidence at the $\alpha = 0.05$ level of significance to support the claim that the proportion of adults taking 200 mg of the drug who experienced headaches is greater than the proportion of adults taking a placebo who experienced headaches.

❖ ----------

4.5.2 Tests about Two Means, μ_1-μ_2.

A) populations variances known

In this section we will discuss, and display, the procedure for testing a statistical hypothesis about two populations' parameters. The populations parameters which are of interest in this section are the populations means, μ_1 and μ_2. To perform inference about the difference between two means, we must first check if the two samples on hand had been taken as independent or dependent samples. A sampling method is independent when the individuals in one sample do not dictate which individuals are to be taken for the second sample. A sampling method is dependent when the individuals selected to represent the first population are used to determine the individuals to be included in the second sample that is representing the second population. The interest, in this section, lies in the difference between those two means for two populations when the samples have been taken independently. The case when the sampling had been dependent will be discussed under another section later.

We will give below the systematic steps in the two methods for testing a hypothesis, namely: the classical method and the p-value method. (For a calculator, this is the 2-sample z-test)

Classical Method Steps:

1. State the null and alternative hypotheses:
   There are three ways to set up the null and alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, two-tailed test.
   b) At least versus less than: H_0: $\mu_1 \geq \mu_2$ versus H_1: $\mu_1 < \mu_2$, left-tailed test.
   c) at most versus greater than: H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$, right-tailed test.

2. Let α be the significance level. Based on the three cases in step 1, we have the following three cases that will go along to find the critical values and the rejection regions for the test

   a) For the two tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$, as shown in the figure.
   b) For the left tailed test, there is the critical value of $-Z_{\alpha}$, and the rejection region is given by $Z < -Z_{\alpha}$, as shown in the figure.
   c) For the right tailed test, again, there is on critical value given by Z_{α}, and the rejection region is $Z > Z_{\alpha}$, as shown in the figure.

3. The test statistic we have $Z = \dfrac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\dfrac{^2_1}{n_1} + \dfrac{^2_2}{n_2}}}$, where n is the sample size, $\bar{x}$ is the sample mean, and Z has the standard normal distribution.

4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

P-value Method Steps: (We are using the 2-sample z-test on the difference between two means).

The steps go like this:

1. State the null and alternative hypotheses:
   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, Two-tailed test
   b) At least versus less than: H_0: $\mu_1 \geq \mu_2$ versus H_1: $\mu_1 < \mu_2$, Left-tailed test
   c) at most versus greater than: H_0: $\mu_1 \leq \mu_2$ versus H_1: $\mu_1 > \mu_2$, right-tailed test

2. Let $\alpha = 0.05$, and based on the three cases in step 1, we have the following three cases that will go along

   a) For the two tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$,
   b) For the left tailed test, there is the critical value of $-Z_{\alpha}$, and the rejection region is given by $Z < -Z_{\alpha}$,
   c) For the right tailed test, again, there is on critical value given by Z_{α}, and the rejection region is $Z > Z_{\alpha}$,

3. The test statistic we have $Z = \dfrac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\dfrac{^2_1}{n_1} + \dfrac{^2_2}{n_2}}}$, where n is the sample size.

4. The above test statistic is computed based on the information provided to us by the sample data.
5. Apply how the p-value is calculated based on the type of your test, as shown above. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed p- value for the test statistic to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0.
6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.9

Test the claim that $\mu_1 \neq \mu_2$ at the 0.05 level of significance for the given data

	Population 1	Population 2
n	15	15
$\bar{x}$	15.3	14.2
σ	3.2	3.5

Solution:

We will do the 2-sampleZTest, assuming the populations are normally distributed with known variances. Using the classical method we have

1. State the null and alternative hypotheses:
   H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$ two-tailed test.

2. Let $\alpha = 0.05$ be the significance level. For the two-tailed test there are two critical values: - $Z_{0.025} = -1.96$ & $Z_{0.025} = 1.96$ and the critical region is given by $|Z| > Z_{0.025} = 1.96$, as shown in the **Figure 1**

3. The test statistic is $Z = \dfrac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\dfrac{^2_1}{n_1} + \dfrac{^2_2}{n_2}}}$,
4. The above test statistic is computed and we have Z = 0.898, based on the information provided.
5. The null hypothesis is not rejected.
6. The two populations have the same mean.

❖

Note: The above test could have been carried using the p-value method, and getting the p-value=0.3690. Thus the null hypothesis is not rejected since the p-value is greater than the significance level.

B) Populations Variances unknown, large samples

In this situation a sample size is considered large if n ≥ 30. Populations can be assumed to be normally distributed with the populations means, μ_1 and μ_2, and populations' variances $^2_1$ and $^2_2$. We are still interested in the difference between the two populations' means.

The steps will go as it was when the population variances are known with the exception that the test statistic will take a different form. It is still a 2-sample z-test with the sample variances replacing the populations' variances in the form for the test statistic. Hence the test statistic will have the following form:

$$Z = \frac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}.$$

All the other set up step for the classical method, and the p-value method, will be applicable here also.

Example 4.10

Test the claim that $\mu_1 > \mu_2$ at the 0.05 level of significance for the given data

	Population 1	Population 2
n	35	35
$\bar{x}$	15.3	14.2
S	3.2	3.5

Solution:

We have two large samples each n > 30. We will do the p-value method on testing the difference between two means, with population variances unknown.

1. State the null and alternative hypotheses:
   H_0: $\mu_1 \le \mu_2$ versus H_1: $\mu_1 > \mu_2$, right-tailed test

2. Let $\alpha = 0.05$
3. The test statistic we have $Z = \dfrac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$.
4. The above test statistic, based on the information provided is Z = 1.3722
5. Apply the p-value for the right-tailed test we see that p-value = 0.08499 >α. Hence the null hypothesis is rejected.
6. The two population means are the equal.

❖ ----------

C) Populations Variances Unknown, Small Samples

Once more the steps in the classical method, and the p-value method, for testing on the difference between two means will apply here as well. However there will be two sub-cases to be considered as the samples' sizes are small. Small sample are those samples with n <30. In addition to that, as it was the case with one mean, population variance unknown, and small sample size, the t-distribution will the underlying distribution for the test statistic in this case. In this section we will discuss, the procedure for testing a statistical hypothesis about two populations' means, when the variances are not known. This case, by itself, raises the following question: (the fact that the variances are unknown), Are the variances equal, or they are different? Without answering this question now, but the situation is worth looking at from these two different sub-cases. The following discussion will address this phenomenon.

1. Let us consider the case when the two populations' variances are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$, their common value. In this case we estimate this common value between the two populations' variances by pooling the two variances of the two samples, and we have

$$\hat{\ }^2 = S_{pooled}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The underlying distribution, for the test statistics will be given by

$$T = \frac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{S_{pooled}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}.$$

The above test statistic will have the student's t-distribution with $= n_1 + n_2 - 2$. The steps in the classical method and the p-value method will apply on the difference between the two means; μ_1-μ_2, by using the 2-sample t-test, with the difference in the test statistic as given above.

EXAMPLE 4.11

Test the claim that $\mu_1 \neq \mu_2$ at the 0.05 level of significance for the given data

	Population 1	Population 2
n	15	15
$\bar{x}$	15.3	14.2
S	3.2	3.5

Solution:

For this example we will consider the case that was outlined above; i.e. we will assume the population variances are equal and carry the 2-sample TTest, with the estimated pooled variance for the common value for the population variances, as given by

$$\hat{}^2 = S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

We will carry the 2-sapmleTTest, using the classical method, as follows:

1. State the null and alternative hypotheses:

   H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$ two-tailed test.

2. Let $\alpha = 0.05$ be the significance level. In this case the degrees of freedom $v = 28$. For the two-tailed test there are two critical values: $t_{.025,28}$ =2.048, -$t_{.025,28}$ = 2.048 and the critical region is given by $|T| > t_{.025,28} = 2.048$.
3. The test statistic we have $T = \dfrac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{S_{pooled}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$,
4. The above test statistic is computed based on the information provided T = 0.898 to us by the sample data.
5. The statistical decision is that the null hypothesis is not rejected.
6. The two population means are equal.

❖

The second sub-case is when $\sigma_1^2 \neq \sigma_2^2$**.** In this case, without testing on the equality of the two variances (This case is coming up later), we will use the Student's t-distribution, but with different degrees of freedom. The degrees of freedom will be calculated from

$$= \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

The degrees of freedom, given above, will not be a whole number, in most cases. The rounding will be done downwards to the nearest positive integer. More over the test statistic will have the following form:

$$T = \frac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

However the steps in the test, whether in the classical method, or the p-value method, will follow as it was in case 1.above for the 2 sample t-test.

In addition to the above two sub-cases, when the populations' variances are unknown, there what we call the Welch's test. This test will consider the test statistic to be given by

$$T = \frac{(\bar{x} - \bar{y}) - (\ _1 - \ _2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

For the above test statistic, Welch's test will take the smaller of $n_1 - 1$ or $n_2 - 1$ as the degrees of freedom for the test. This has to be done, or calculated, manually since no software will calculate the test statistic value based on this type of degrees of freedom. Awareness, by the reader, or the investigator, should be taken into consideration for what case he/she will consider.

Remark:

The interested student can check on the above example using the Welch's Test, or the case when the sample variances are not pooled for the 2-sampleTTest.

4.5.3 Tests about Two Variances

Testing a statistical hypothesis about a population variance, or standard deviation, is surely different from testing about one population's mean or proportion. The difference lies in the distribution of the test statistic involved in the process. The statistic we are talking about here is the sample variance, or standard deviation, and its distribution. The test on one population variance is carried out based on the interest to check on the variability in the population. As it was the case in finding the confidence interval for one variance, we appealed to the random variable for the test statistic given by

$$\chi^2 = \frac{(n-1)S^2}{\ _2}.$$

This random variable has a (chi-square) χ^2-distribution with n-1 degrees of freedom, and S^2 is the sample's variance, σ^2 is the population's variance, and n is the sample's size. (See the properties of the χ^2-distribution set up in Chapter 3).

When we tested on one population mean with small sample size and unknown variance, using the t-test, we had to assume that the population has a normal distribution. For testing on one population variance we will need the same condition, i.e. the sampled population needs to have a normal distribution. As it was the case on one population mean or proportion, we can use either the classical or the p-value method, with the one restriction that we will consider only the right-tailed test. This is because we like to check on the large value of the variance, or on a high variation in the population.

Sometimes it is of interest to compare the variability (precision) of two procedures for taking measurements or analysis (e.g. two machines, two operators). This comparison can be carried out through comparing the variances, or the standard deviations, in the two procedures.

Recall the case when we tested on μ_1-μ_2, when the population variances were unknown, we had to consider the two cases whether the variances are equal, in order to pool, or not. The 2-sample T-test was carried out without really checking on the equality of the two populations' variances. May be it is time now. We tested on one variance or one standard deviation using the Chi-square test give by the random variable defined as above in 4.4.5.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n from a normal population with mean μ_1 and variance σ_1^2. Also let $Y_1, Y_2, \ldots, Y_m$ be another simple random sample from another normal population with mean μ_2 and variance σ_2^2. Here it is assumed that the populations' variances are unknown. In other words we let $X \sim N(\ _1,\ _1^2)$, $Y \sim N(\ _2,\ _2^2)$, respectively be the two independent variables with the assigned distribution that the samples were taken from. The steps for testing the statistical hypothesis on the ratio between the two variances need another random variable to be introduced.

Classical Method Steps: One more time the steps will go as follows:

1. State the null and alternative hypotheses:
   There are three ways to set up the null and alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ two-tailed test.
   b) At least versus less than: H_0: $\sigma_1^2 \geq \sigma_2^2$ versus H_1: $\sigma_1^2 < \sigma_2^2$, left-tailed test.
   c) at most versus greater than: H_0: $\sigma_1^2 \leq \sigma_2^2$ versus H_1: $\sigma_1^2 > \sigma_2^2$, right-tailed test

2. Let α be the significance level, and based on the three cases in step 1, we have the following three cases that will go along in order to find the critical values and the rejection region.

   a) For the two tailed test there are two critical values: $F_{1-\alpha/2}$ & $F_{\alpha/2}$, and the critical regions are given by $F > F_{\alpha/2}$, or $F < F_{1-\alpha/2}$, as shown in Figure 14.

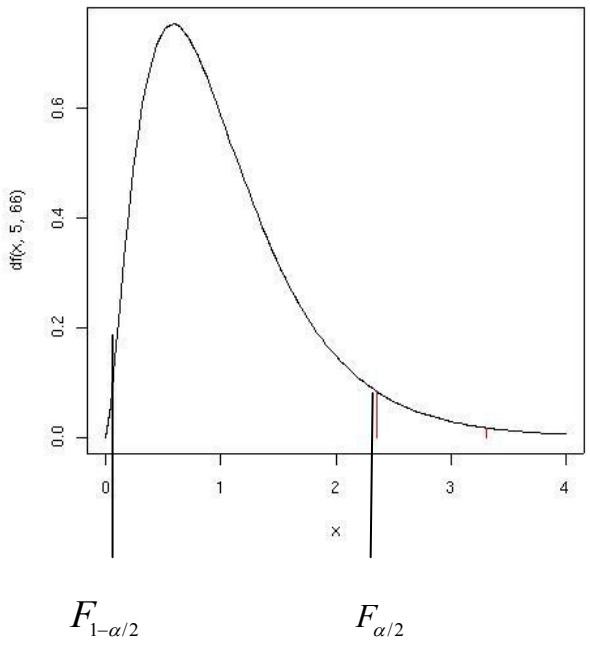


Figure 14

   b) For the left tailed test, there is the critical value $F_{1-\alpha}$, and the rejection region is given by $F < F_{1-\alpha}$ as shown in Figure 15.

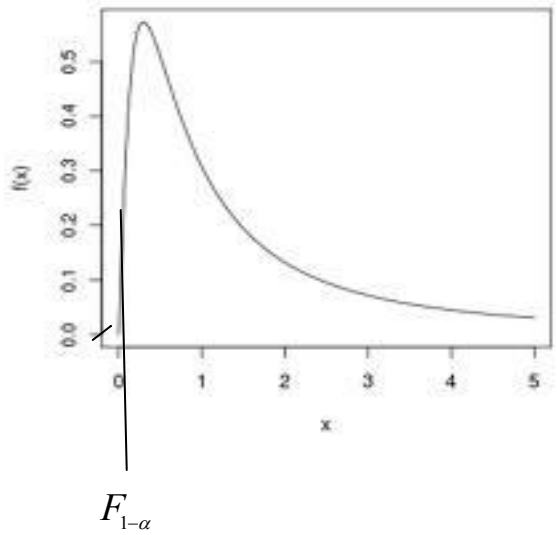


Figure 15

c) For the right tailed test, again, there is on critical value given by F_α, and the rejection region is $F > F_\alpha$, as shown in Figure 16.

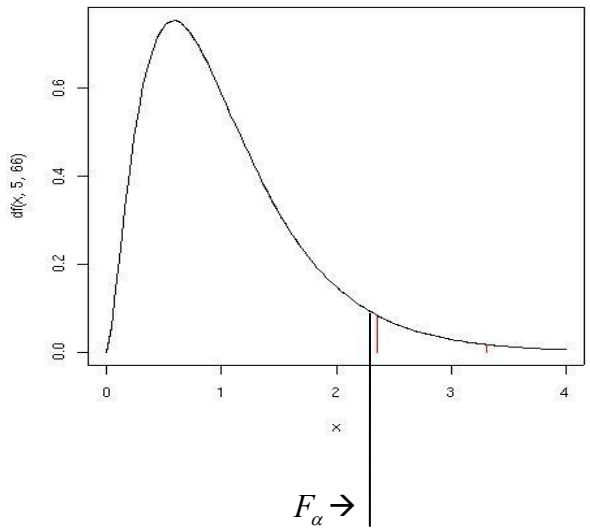


Figure 16

3. The test statistic is $F = \frac{S_1^2}{S_2^2}$ where the larger of the two samples' variances is S_1^2. Then F has the F-distribution, with n_1^2–1 degrees of freedom for the numerator and n_2–1 degrees of freedom for the denominator, with n_1 and n_2 are the samples' sizes.
4. The above test statistic is computed based on the information provided to us by the sample data.

5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.12

A company produces machined engine parts that are supposed to have a diameter variance no larger than 0.0002 (diameters are measured in inches). The company wishes to compare the variation in diameters produced by the company with the variation in diameters parts produced by another competitor. Our company had a sample of size n=10 that produced a sample variance $S_1^2 = 0.0003$. In contrast, the sample variance of the diameter measurements for 20 of the competitor's parts was $S_2^2 = 0.0001$. Do the data provide sufficient information to indicate a smaller variation in diameters for the competitor? Test with $\alpha = 0.05$, by using the classical method.

Solution:

1. H_0: $\sigma_1^2 \leq \sigma_2^2$ versus H_1: $\sigma_1^2 > \sigma_2^2$, right-tailed test
2. $\alpha = 0.05$, and $n_1 = 10$, $n_1 - 1 = 9$, $n_2 = 20$, $n_2 - 1 = 19$. Thus we have an F-distribution with 9 and 19 degrees of freedom for the numerator and denominator respectively, the CV is given by F (9, 19; .05) = 2.42. Moreover, the rejection region is given by F > 2.42.
3. The Test statistic is given by (since under H_0 $\sigma_1^2 = \sigma_2^2$) $F = \frac{S_1^2}{S_2^2}$
4. The observed value of the test statistic, based on what we have on hand, is F = 3.
5. We see that F > F (9, 19; .05), therefore at the $\alpha = 0.05$ level of significance, we reject the null hypothesis H_0: $\sigma_1^2 \leq \sigma_2^2$ in favor of the alternative hypothesis, namely H_1: $\sigma_1^2 > \sigma_2^2$.
6. We conclude that the competitor produces parts with smaller variation in their diameters.

❖

Remark

Recall that it is not easily, by using the table for the F-distribution, to pin point the p-value for the test. All we can do is that putting a range on the p-value.

EXAMPLE 4.13

Give the bounds on the p-value associated with the data in the above Example ###.

Solution

As it was shown, the calculated F-value for this right-tailed test is F = 3. Since this value is to be compared with F (9, 19; .05) = 2.42, by using the F-distribution Table for F (9, 19; .025) we find that F (9, 19; .025) = 2.82, whereas F (9, 19; .01) = 3.52. Hence based on the observed value of the F-Statistic, F = 3, would lead to the rejection of the null hypothesis for $\alpha = 0.025$ but not for $\alpha = 0.01$. Therefore, we have the following of values on the p-value, which is 0.01 < p-value < 0.025.

On the other hand, by using technology, and more specifically, TI Calculator 84-Plus, we find the p-value for our test is given exactly as 0.020960, a value clearly in the above cited range.

❖

CHAPTER 4 EXERCISES

In Exercises 1–6, a null and alternative hypothesis is given. Determine whether the hypothesis test is left-tailed, right-tailed, or two-tailed. What is the parameter that being tested?

4.1 H_0: $\mu = 5$ versus H_1: $\mu > 5$.

4.2 H_0: $p \geq 0.2$ versus H_1: $p < 0.2$.

4.3 H_0: $\sigma = 4.5$ 5 versus H_1: $\sigma \neq 4.55$.

4.4 H_0: $p \leq 0.75$ versus H_1: $p > 0.75$.

4.5 H_0: $\mu \geq 115$ versus H_1: $\mu < 115$.

4.6 H_0: $\sigma \leq 5$ versus H_1: $\sigma > 5$.

In exercises 7–9,

a) Determine the null and alternative hypotheses.
b) Explain what it would mean to a type I error, and
c) Explain what it would mean to make a type II error.

4.7 According to the Federal Housing Finance Board, the mean price of a single-family home, in 2003, was $245,950. A real estate broker believes that because of the credit standing and the interest rate, the mean price has increased since then.

4.8 According to the Centers for Disease Control and Prevention, 16% of children aged 6 to 11 years are overweight. A school nurse thinks that the percentage, of 6- to 11-year-olds that is overweight, is higher in her school district.

4.9 In 2006, the standard deviation of SAT math score for all students taking the exam was 115. A teacher believes that, due to changes to the SAT Reasoning test in 2007, the standard deviation of SAT math scores will increase.

4.10 State the conclusion based on the results of the test for the claim in Exercises 1–6, when the null hypothesis is rejected.

4.11 State the conclusion based on the results of the test for the claim in Exercises 7–9, when the null hypothesis is not rejected.

In Exercises 4.12 – 4.14, test the hypothesis using (a) the classical method and then (b) the P-value method. Be sure to verify the requirements of the test, and the conclusions in both methods are the same.

4.12 H_0: $p \geq 0.9$ versus H_1: $p < 0.9$, n 500, x = 440, $\sigma = 0.01$.

4.13 H_0: $p \leq 0.75$ versus H_1: $p > 0.75$, n = 200, x = 75, $\alpha = 0.05$

4.14 H_0: $p = 0.55$ versus H_1: $p \neq 0.55$ n = 150, x = 785, $\alpha = 0.10$

4.15 A tomato juice cannery attempts to put 46 ounces in the can. The measuring device puts in X ounces, a random variable that is normally distributed. If the average content is below 46 ounces, the company may get in trouble with the government inspectors for false labeling. On the other hand if the average content is above 46 ounces, the company will make less profit. In order to determine whether or not the weighing process is operating satisfactorily the plant statistician inspects a simple random sample of 25 cans and finds that $\bar{x} = 46.18$ ounces with S = 0.5. What conclusion should the company make, by using $\alpha = 1\%$?

4.16 To test H_0: $\mu \leq 15$ versus H_1: $\mu > 15$, a random sample of size 25 is obtained from a population that is known to be normally distributed with $\sigma = 5$. a) If the sample mean is determined to be 14.7, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.05, determine the critical value. c) Draw a normal curve that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why not?

4.17 To test H_0: $\mu \geq 40$ versus H_1: $\mu < 40$, a random sample of size 25 is obtained from a population that is known to be normally distributed with $\sigma = 6$. a) If the sample mean is determined to be 42.3, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.10, determine the critical value. c) Draw a normal curve that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why not?

4.18 To test H_0: $\mu = 100$ versus H_1: $\mu \neq 100$, a random sample of size 30 is obtained from a population that is known to be normally distributed with $\sigma = 7$. a) If the sample mean is determined to be 42.3, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.01, determine the critical value. c) Draw a normal curve that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why not?

4.19 A standard variety of wheat produces, on the average, 30 bushels per acre. A new imported variety is planted on nine randomly selected acre plots. The observed sample average for the new variety is 33.4 bushels per acre, with a standard deviation of 5.1 bushels. Should the new variety be used instead of the standard one, by using 5% level of significance?

4.20 To test H_0: $\mu = 45$ versus H_1: $\mu \neq 45$, a random sample of size 40 is obtained from a population that has a standard deviation of $\sigma = 8$. a) Does the population need to be normally distributed to compute the P value? b) If the sample mean is determined to be 48.3, compute the p-value, and interpret it. c) If the researcher decides to test this hypothesis at the level of significance of 0.05, determine the critical value. d) Draw a normal curve that will show the critical (or rejection) region. e) Will the researcher reject the null hypothesis? Why or why not?

4.21 To test H_0: $\mu \geq 40$ versus H_1: $\mu < 40$, a random sample of size 25 is obtained from a population that is known to be normally distributed. a) If the sample mean is determined to be 42.3, and $s = 4.3$, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.10, determine the critical value. c) Draw a t-distribution curve that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why not?

4.22 To test H_0: $\mu = 100$ versus H_1: $\mu \neq 100$, a random sample of size 23 is obtained from a population that is known to be normally distributed. a) If the sample mean is determined to be 104.8, with $s = 9.2$, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.01, determine the critical value. c) Draw a t-distribution curve that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why not?

4.23 To test H_0: $\mu \geq 20$ versus H_1: $\mu < 20$, a simple random sample of size 18 is obtained from a population that is known to be normally distributed. a) If the sample mean is determined to be 18.3, and s = 4.3, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.10, determine the critical value. c) Draw a t-distribution curve that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis, use the P-value method? Why or why not?

4.24 State the requirements to test a claim regarding a population standard deviation, or a population variance.

4.25 Determine the critical value for a right-tailed test of a population standard deviation with 18 degrees of freedom at the 5% level of significance.

4.26 Determine the critical values for a two-tailed test of a population variance with 18 degrees of freedom at the 5% level of significance.

4.27 To test H_0: $\sigma \leq 35$ versus H_1: $\sigma > 35$, a random sample of n = 15 is obtained from a population that is normally distributed. a) If the sample standard deviation is determined to be s = 37.4, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.01, determine the critical value. c) Draw a Chi-square distribution that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why not?

4.28 To test H_0: $\sigma \geq 0.35$ versus H_1: $\sigma < 0.35$, a random sample of n = 41 is obtained from a population that is normally distributed. a) If the sample standard deviation is determined to be s = 0.23, compute the test statistic. b) If the researcher decides to test this hypothesis at the level of significance of 0.01, determine the critical value. c) Draw a Chi-square distribution that will show the critical (or rejection) region. d) Will the researcher reject the null hypothesis? Why or why no.

TECHNOLOGY STEP-BY-STEP

TECHNOLOGY STEP-BY-STEP **Hypothesis Tests Regarding μ, σ known**

TI-83/84 Plus

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 1: Z-Test
3. If the data is raw, highlight DATA. Make sure List1 is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. Following sigma: enter the population standard deviation. For the value of U0 enter the value sated in the null hypothesis.

4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw; press ENTER.

Excel

1. Enter raw data in column A. Highlight the data.
2. Load the data Desk XL. Add-in, if necessary.
3. Select the DDXL menu. Highlight Hypothesis Tests. From the drop-down menu under Function Type:, select 1 VarZ test.
4. For raw data, Select the column of data from the "Names and Column" window. Use the < arrow to select the data. Click OK.
5. Fill in the Data Desk window. In step 1, click "set μ_0 and s.d." Enter the values of the hypothesized mean (this is the value that is stated in the null hypothesis) and the population standard deviation. In step 2, select the level of significance. In step 3, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP Hypothesis Tests Regarding μ, σ Unknown

TI-83/84 Plus

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 2: T-Test
3. If the data is raw, highlight DATA. Make sure List1 is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. For the value of U0 enter the value stated in the null hypothesis.
4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw; press ENTER. The TI-83/84 gives the P-value.

Excel

1. Enter raw data in column A. Highlight the data.
2. Load the data Desk XL. Add-in, if necessary.
3. Select the **DDXL** menu. Highlight **Hypothesis Tests**. From the drop-down menu under Function Type: select **1 Var t test**.
4. For raw data, Select the column of data from the "Names and Column" window. Use the < arrow to select the data. Click OK.
5. Fill in the Data Desk window. In step 1, click "set μ_0". Enter the value of the hypothesized mean (this is the value that is stated in the null hypothesis). In step 2, select the level of significance. In step 3, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP Hypothesis Tests Regarding a Population Proportion

TI-83/84 Plus

1. Press STAT, highlight TESTS, and select 5: 1-PropZTest
2. For the value of p0 enter the value stated in the null hypothesis.
3. Enter the number of successes, x, and the sample size, n.
4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw; press ENTER. The TI-83/84 gives the P-value.

Excel

1. If you have, enter the raw data in column A. If you have summarized data, enter the number of successes in column 1 and the number of trials in column B. Highlight the data.
2. Load the data Desk XL. Add-in, if necessary.
3. Select the **DDXL** menu. Highlight **Hypothesis Tests**. From the drop-down menu under Function Type: select **1 Var Prop Test**, if you have raw data; select Sum 1-Var-Prop-Test if you have summarized data.
4. If you have raw data, Select the column that contains the raw data from the "Names and Column" window. Use the < arrow to select the data. Click OK. If you have summarized data, highlight the observations in "Names and Columns" and use the < arrow to select the number of successes; repeat this for the number of trials. Click OK.
5. Fill in the Data Desk window. In step 1, click "set hypothesized value of proportion" and enter them value of the hypothesized. Enter the value of the hypothesized proportion (this is the value that is stated in the null hypothesis). In step 2, select the level of significance. In step 3, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP Two-Samples t-Tests, Dependent Sampling

TI-83/84 Plus

1. If necessary, enter raw data in L1 and L2. Let L3 = L1 – L2 (or L2 – L1), depending on how the alternative hypothesis is defined.
2. Press STAT, highlight TESTS, and select 2: T-Test
3. If the data is raw, highlight DATA. Make sure List1 is set to L3 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw; press ENTER. Calculate gives the test statistic and P-value. Draw will draw the t-distribution with the P-value shaded.

Excel

1. Enter raw data in Column A and B.
2. Select **TOOLS** menu and highlight **Data Analysis**….
3. Select "t-test: Paired two-sample for means." With the curser in the "Variable 1 Range" cell, highlight the data in Column A. With the cursor in the "Variable 2 Range" cell, highlight the data in column B. Enter the hypothesized difference in the means (usually 0) and a value for alpha. Click OK.

TECHNOLOGY STEP-BY-STEP Two -Samples t-Tests, Independent Sampling

TI-83/84 Plus

Hypothesis Tests

1. If necessary, enter raw data in L1 and L2.
2. Press STAT, highlight TESTS, and select 4: 2-SampleTTest
3. If the data is raw, highlight DATA. Make sure List1 is set to L1, List2 is set to L2, and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
4. Highlight the appropriate relation between mu1 and mu1 in the alternative hypothesis. Set Pooled to NO.
5. Highlight Calculate or Draw; press ENTER. Calculate gives the test statistic and P-value. Draw will draw the t-distribution with the P-value shaded.

Confidence Interval

Follow the steps for hypothesis tests, except select 0: 2-SampleTInt. Also, select a confidence level (such as 95% = 0.95).

Excel

1. Enter raw data in Column A and B.
2. Select **TOOLS** menu and highlight **Data Analysis**….
3. Select "t-test: Two-sample Assuming Unequal Variances." With the curser in the "Variable 1 Range" cell, highlight the data in Column A. With the cursor in the "Variable 2 Range" cell, highlight the data in column B. Enter the hypothesized difference in the means (usually 0) and a value for alpha. Click OK.

TECHNOLOGY STEP-BY-STEP Inference for Two Population Proportions

TI-83/84 Plus

Hypothesis Tests

1. Press STAT, highlight TESTS, and select 6: 2-PropZTest
2. Enter the values of x_1, n_1, x_2, n_2.
3. Highlight the appropriate relation between p_1 and p_2 in the alternative hypothesis.
4. Highlight Calculate or Draw; press ENTER. Calculate gives the test statistic and P-value. Draw will draw the Z-distribution with the P-value shaded.

Confidence Interval

Follow the steps for hypothesis tests, except select 0: 2-PropZInt. Also, select a confidence level (such as 95% = 0.95).

Excel

1. Enter raw data in Column A and B, using 0 for failures and 1 for successes. Highlight the data.
2. Load the DDXL Add-in, if necessary
3. Select **DDXL** menu and highlight **Hypothesis Tests**. From the drop-down window under Function

   Type: select 2Var Prop Test

4. Select the column that contains the data for sample 1 from the "Names and Columns" window. Use the < arrow to select data for 1 prop variable. Select the column that contains the data for sample 2 from the "Names and Columns" window. Use the < arrow to select data for 2 Prop Variable. Click OK. Note: If the first row contains the names of the variable, check the "first row is variable names" box.
5. Fill in the Data Desk window. In step 1, click "set Success 1" so that 1 is a success. Click "set Success 2" so that 1 is also a success. In step2, click "set p" and enter the, value of the difference in the proportions stated in the null hypothesis (usually 0). In step 3 select the level of significance. In step 4, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP **Hypothesis Tests regarding μ, σ known**

TI-83/84 Plus

6. If necessary, enter raw data in L1.
7. Press STAT, highlight TESTS, and select 1: Z-Test
8. If the data is raw, highlight DATA. Make sure List1 is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. Following sigma: enter the population standard deviation. For the value of U_0 enter the value sated in the null hypothesis.
9. Select the direction of the alternative hypothesis.
10. Highlight Calculate or Draw; press ENTER.

Excel

6. Enter raw data in column A. Highlight the data.
7. Load the data Desk XL. Add-in, if necessary.
8. Select the **DDXL** menu. Highlight **Hypothesis Tests**. From the drop-down menu under Function Type:, select **1 Var Z test**.
9. For raw data, Select the column of data from the "Names and Column" window. Use the < arrow to select the data. Click OK.
10. Fill in the Data Desk window. In step 1, click "set μ_0 and sd". Enter the values of the hypothesized mean (this is the value that is stated in the null hypothesis) and the population standard deviation. In step 2, select the level of significance. In step 3, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP **Hypothesis Tests Regarding μ, σ Unknown**

TI-83/84 Plus

6. If necessary, enter raw data in L1.
7. Press STAT, highlight TESTS, and select 2: T-Test
8. If the data is raw, highlight DATA. Make sure List1 is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics. For the value of U_0 enter the value stated in the null hypothesis.
9. Select the direction of the alternative hypothesis.
10. Highlight Calculate or Draw; press ENTER. The TI-83/84 gives the P-value.

Excel

6. Enter raw data in column A. Highlight the data.
7. Load the data Desk XL. Add-in, if necessary.
8. Select the **DDXL** menu. Highlight **Hypothesis Tests.** From the drop-down menu under Function Type: select **1- Var t-test.**
9. For raw data, Select the column of data from the "Names and Column" window. Use the < arrow to select the data. Click OK.
10. Fill in the Data Desk window. In step 1, click "set μ_0". Enter the value of the hypothesized mean (this is the value that is stated in the null hypothesis). In step 2, select the level of significance. In step 3, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP **Hypothesis Tests Regarding a Population Proportion, P**

TI-83/84 Plus

6. Press STAT, highlight TESTS, and select 5: 1-PropZTest
7. For the value of p_0 enter the value stated in the null hypothesis.
8. Enter the number of successes, x, and the sample size, n.
9. Select the direction of the alternative hypothesis.
10. Highlight Calculate or Draw; press ENTER. The TI-83/84 gives the P-value.

Excel

6. If you have, enter the raw data in column A. If you have summarized data, enter the number of successes in column 1 and the number of trials in column B. Highlight the data.
7. Load the data Desk XL. Add-in, if necessary.
8. Select the **DDXL** menu. Highlight **Hypothesis Tests.** From the drop-down menu under Function Type:, select **1 Var Prop Test,** if you have raw data; select Sum 1Var Prop Test if you have summarized data.

9. If you have raw data, Select the column that contains the raw data from the "Names and Column" window. Use the < arrow to select the data. Click OK. If you have summarized data, highlight the observations in "Names and Columns" and use the < arrow to select the number of successes; repeat this for the number of trials. Click OK.
10. Fill in the Data Desk window. In step 1, click "set hypothesized value of proportion" and enter them value of the hypothesized. Enter the value of the hypothesized proportion (this is the value that is stated in the null hypothesis). In step2, select the level of significance. In step 3, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP **Two-Samples t-Tests, Dependent Sampling**

TI-83/84 Plus

6. If necessary, enter raw data in L1 and L2. Let L3 = L1 – L2 (or L2 – L1), depending on how the alternative hypothesis is defined.
7. Press STAT, highlight TESTS, and select 2: T-Test
8. If the data is raw, highlight DATA. Make sure List1 is set to L3 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
9. Select the direction of the alternative hypothesis.
10. Highlight Calculate or Draw; press ENTER. Calculate gives the test statistic and P-value. Draw will draw the t-distribution with the P-value shaded.

Excel

4. Enter raw data in Column A and B.
5. Select **TOOLS** menu and highlight **Data Analysis**….
6. Select "t-test: Paired two-sample for means." With the curser in the "Variable 1 Range" cell, highlight the data in Column A. With the cursor in the "Variable 2 Range" cell, highlight the data in column B. Enter the hypothesized difference in the means (usually 0) and a value for alpha. Click OK.

TECHNOLOGY STEP-BY-STEP **Two -Samples t-Tests, Independent Sampling**

TI-83/84 Plus

Hypothesis Tests

6. If necessary, enter raw data in L1 and L2.
7. Press STAT, highlight TESTS, and select 4: 2-SampleTTest
8. If the data is raw, highlight DATA. Make sure List1 is set to L1, List2 is set to L2, and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
9. Highlight the appropriate relation between mu1 and mu1 in the alternative hypothesis. Set Pooled to NO.
10. Highlight Calculate or Draw; press ENTER. Calculate gives the test statistic and P-value. Draw will draw the t-distribution with the P-value shaded.

Confidence Interval

Follow the steps for hypothesis tests, except select 0: 2-SampleTInt. Also, select a confidence level (such as 95% = 0.95).

Excel

4. Enter raw data in Column A and B.
5. Select **TOOLS** menu and highlight **Data Analysis**….
6. Select "t-test: Two-sample Assuming Unequal Variances." With the curser in the "Variable 1 Range" cell, highlight the data in Column A. With the cursor in the "Variable 2 Range" cell, highlight the data in column B. Enter the hypothesized difference in the means (usually 0) and a value for alpha. Click OK.

TECHNOLOGY STEP-BY-STEP Inference for Two Population Proportions, $P_1 - P_2$

TI-83/84 Plus

Hypothesis Tests

5. Press STAT, highlight TESTS, and select 6: 2-PropZTest
6. Enter the values of x_1, n_1, x_2, n_2.
7. Highlight the appropriate relation between p_1 and p_2 in the alternative hypothesis.
8. Highlight Calculate or Draw; press ENTER. Calculate gives the test statistic and P-value. Draw will draw the Z-distribution with the P-value shaded.

Confidence Interval

Follow the steps for hypothesis tests, except select 0: 2-PropZInt. Also, select a confidence level (such as 95% = 0.95).

Excel

6. Enter raw data in Column A and B, using 0 for failures and 1 for successes. Highlight the data.
7. Load the DDXL Add-in, if necessary
8. Select **DDXL** menu and highlight **Hypothesis Tests**. From the drop-down window under Function
   Type:, select 2Var Prop Test
9. Select the column that contains the data for sample 1 from the "Names and Columns" window. Use the < arrow to select data for 1 prop variable. Select the column that contains the data for sample 2 from the "Names and Columns" window. Use the < arrow to select data for 2 Prop Variable. Click OK. Note: If the first row contains the names of the variable, check the "first row is variable names" box.
10. Fill in the Data Desk window. In step 1, click "set Success 1" so that 1 is a success. Click "set Success 2" so that 1 is also a success. In step 2, click "set p" and enter the, value of the difference in the proportions stated in the null hypothesis (usually 0). In step 3 select the level of significance. In step 4, select the direction in the alternative hypothesis. Click Compute.

TECHNOLOGY STEP-BY-STEP **Inference for Two Population variances**

TI-83/84 Plus

1. If necessary, enter raw data in L_1, L_2.
2. Press STAT, highlight TESTS, and select 7: 2- Sample F-Test
3. If the data is raw, highlight DATA. Make sure List1 is set to L_1 and, and List2 is set to L_2, and freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics..
4. Highlight the appropriate relationship between the two standard deviations under the alternative Hypothesis
5. Select Pooled: No or Yes, based on your choice, there will a difference in the degrees of freedom between the pooled and the non-pooled case, as you will notice.
6. Highlight Calculate; OR draw, and press ENTER. Calculate gives the test statistics and the p-value. Draw draws the F-distribution with the p-value shaded.

5 Simple Linear Regression and Correlation

Outline

5.1 Introduction

It is quite often that the investigator likes to check how variables are related, and to "find if he can" a relationship 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, 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. For 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 variable** 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 linear regression. In case we have one response and many predictor variables, this is the case of multiple regressions. In this chapter, we are interested in the case of two variables, and when the relationship between them is assumed to be a linear one. In addition to linear regression between two variables there might be logarithmic, exponential, quadratic, cubic, and so on regression.

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.

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, see exercise 5.15.

In much of the experimental work, it is desired to investigate 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. (See Draper and Smith, pp. 4–6).

5.1.1 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 "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 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 for I for a particular value for V (with R as a fixed constant) we should use the plotted straight line.

5.1.2 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 for some population. If we plot the ordered pair (weight, height) = (x, y), a diagram, like Figure 1 below, will result. Such a representation is conventionally called a scatter diagram.

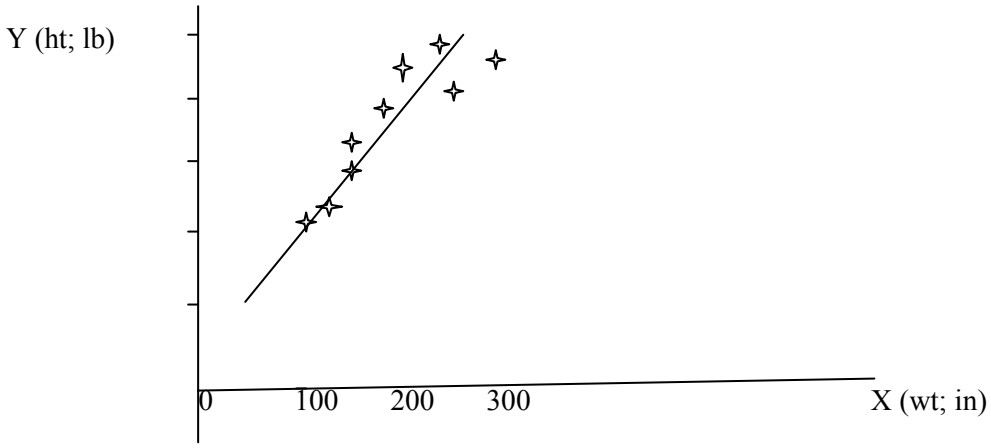


Figure 1

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? The investigator 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 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 relation implied.

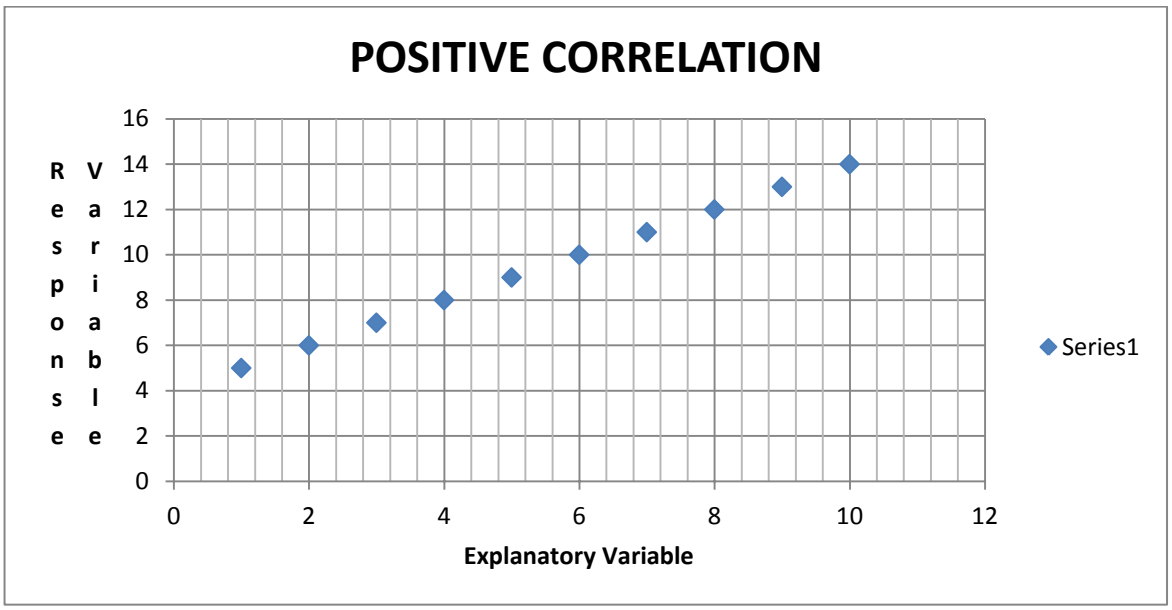


Figure 2 Perfect Positive Correlations

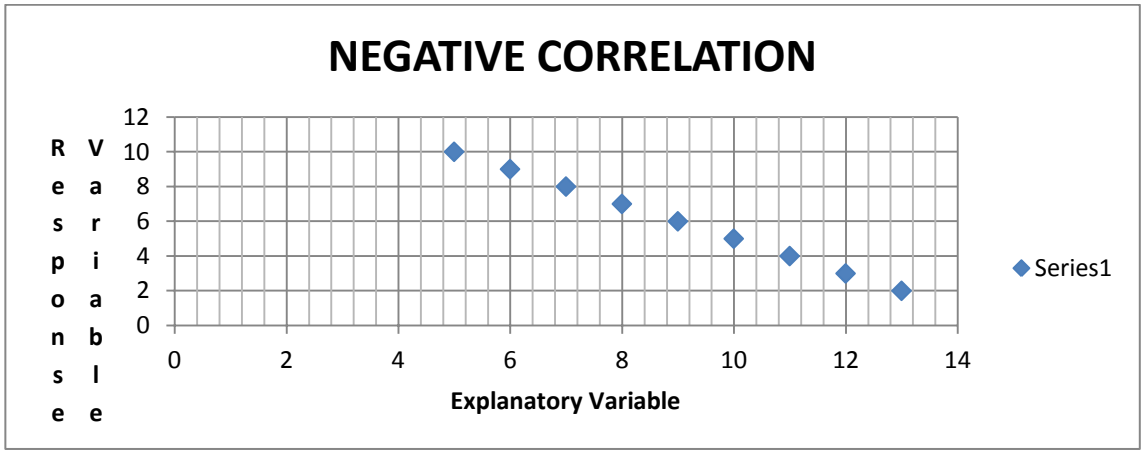


Figure 3 Perfect Negative Correlation.

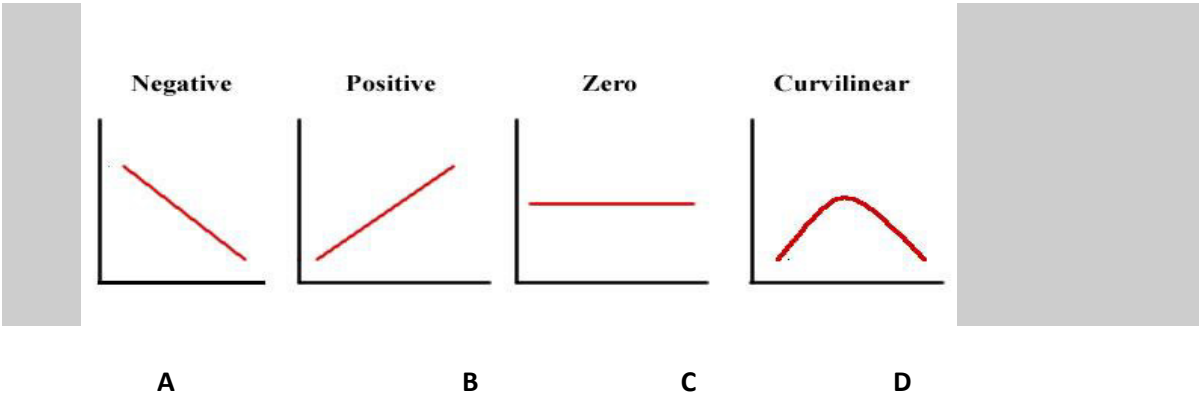


Figure 4

Figure 4, above, shows the type 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, is not affected by the change in the explanatory variable, x. Thus Y = a constant.

Figure 4D shows a non-linear 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 5.4 we will check on the strength of that linear relationship between the explanatory variable x and the response variable Y.

5.2 Regression Models

The simplest linear regression model is of the form

$$\mathbf{y} = \beta_0 + \beta_1 \mathrm{x} + \mathrm{E},$$

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 and β_1 are unknown constants, 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 E 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 relationship 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 the 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 could not be used for predictive purposes outside that range.

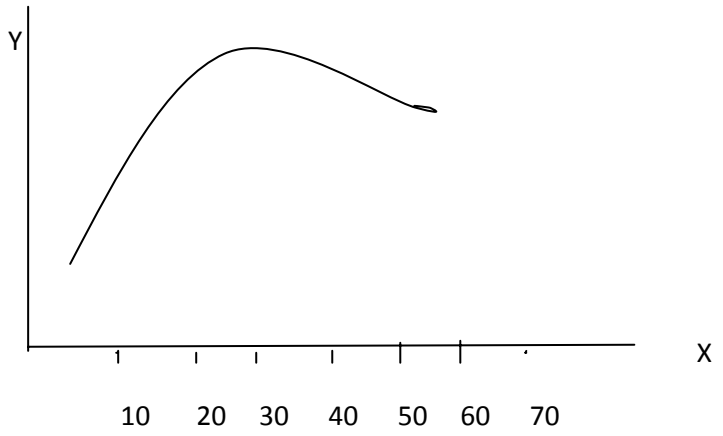


Figure 5

5 .3 Fitting a straight line (First order Model)

The first order model given by

$$y = \beta_0 + \beta_1 x + e,$$

has the unknown parameters β_0 and β_1, that need to be estimated in order to use the relationship for prediction. Thus a sample of size n of ordered observations (x_1, y_1), (x_2, y_2), … , (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

$$\hat{y} = b_0 + b_1 x.$$

In the notation, above, $\hat{y}$ is used in the least squares regression line to serve as a 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

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

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Each observation (x_i, y_i), i = 1, 2,…, n, in the sample, satisfies the equation

$$\mathbf{Y}_i = \beta_0 + \beta_1 x_i + \quad _i ,$$

Where $_i$ is the value assumed by E_i when Y_i takes on the value y_i. The above equation can be viewed as the model for a single observation y_i. Similarly, using the estimated, or fitted regression line

$$\hat{y} = b_0 + b_1 x.$$

Each pair of observations satisfies the relation

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

Where $e_i = y_i - \hat{y}_I$ is called a residual and describes the error in the fit of the model at the ith 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 denoted by SSE. This minimization procedure for estimating the parameters is called the Method of Least-Squares. Thus we shall find b0 and b1 so as to minimize

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 .$$

(The interested reader, who is aware of differential calculus, can check the procedure in R.E. Walpole, and R.H. Myers; Probability and Statistics for Engineers and Scientists, 4th Edition, p. 361.) The least –squares regression line is the line that minimizes the square of the vertical distance 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

$$b_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \right)}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \right)}, \text{ and } \qquad b_0 = \bar{y} - b_1 \bar{x} .$$

EXAMPLE 5.1:

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 ^{0}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^{0}C.

Solution:

From the data we have: 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 $\hat{y} = 0.514 + 0.0028x$. When the temperature x = 25^{0}C, the predicted specific heat is 0.514+0.0028*25 = 0.584.

EXAMPLE 5.2 (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 about zero to make $\sum_{i=1}^{n} x_i = 0$, and thus the estimators of coefficients in the regression line become

$$b_1 = \sum_{i=1}^{n} x_i y_i \,/\, \sum_{i=1}^{n} x_i^2 \text{ , and } b_0 = \bar{y}\,.$$

If there are an odd number of evenly spaced data values for X, denote them by: …, -3, -2, -1, 0, 1, 2, 3, ….

If there is an even number of evenly spaced data values for X, denote them by: …, -5, -3, -1, 1, 3, 5, ….

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 to look like the following table.

Output (1000 tons)	11.1	12.3	13.7	14.6	15.6
Year	-2	-1	0	1	2

For manual calculations, we have the following Table.

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

Hence the estimated values 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*3 = 16.85 or 16850 tons.

EXAMPLE 5.3: (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 linearized easily. For example consider the experiment growth (or Decay, when the rate is negative) model:

$$P = Ae^{rt}.u,$$

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.

Consider the following data for the population of the USA (in millions) during the years 1860–1900:

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 to a linear relation by taking the logarithm to base e (i.e. ln) of both sides to obtain

$$\text{Ln } P = \text{Ln } A + rt + \text{Ln } u,$$

Which can be written as: $\mathbf{y} = \beta_0 + \beta_1 x + E$, where y = Ln P, $\beta_0 = r$, x = t and E = Ln u = 0. Now we have a linear model, we can fit using the data given above. Moreover, by appealing to EXAMPLE **5.2,** and coding the data, we have the following generated data in Table 2 for manual calculations.

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

Thus we have

$$b_1 = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 = 0.222, \text{ and}$$

$$b_0 = \bar{y} = 3.902,$$

and the linear model takes the form

$$\hat{y} = 3.902 + 0.222x.$$

By encoding we reach at $\widehat{P} = \mathbf{49.5e^{0.222t}}$.

❖

5.4 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 two variables see Figure 1. For the sake of argument, what if someone else used a different scale for the same data and he/she 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 of 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 to 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?

In order to keep the math background to a minimum for this course, we will give a 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. Recall the formula for the z-scores, in words, with S being the standard deviation of the data set

$$\text{Z-score} = \frac{data\ point - mean\ of\ the\ data\ set}{S}.$$

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

$$Z_x = \frac{x_i - \overline{x}}{S_x}, \text{ and } Z_y = \frac{y_i - \overline{y}}{S_y}.$$

Based on the above z's expressions, we can write the sample correlation coefficient r as

$$r = \frac{\sum_i Z_x Z_y}{n-1}.$$

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 is

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\left(\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}\right)\cdot\left(\sqrt{\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}\right)}$$

(See Wackerly, Mendenhall and Scheaffer 2008)

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).

Let us give some properties of the Linear Correlation Coefficient.

1. The Linear Correlation coefficient is always between -1 and 1, inclusive. In other words, $-1 \leq r \leq 1$.
2. If r = 1, there is a perfect positive linear relation between the two variables. See Figure 4B.
3. If r = -1, there is a perfect negative linear relation between the two variables. See Figure 4A.
4. The closer r to 1, the stronger is the evidence of positive association between the two variables.
5. The closer r to -1, the stronger is the evidence of negative association between the two variables.
6. The closer r to 0, there is 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.

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 5.4

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.

5.5 Hypothesis Testing in Regression Analysis

In this section we discuss the precision of the regression coefficients, the construction of confidence limits, and testing the statistical hypotheses about the regression coefficients. Before proceeding we need to make some basic assumptions on the first order model as follows. While the model is given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, 3, \ldots, n.$$

Provided that the ε_i are independent random variables which have a normal distribution with mean zero and variance σ^2, i.e. $E(\varepsilon_i) = 0$, and $Var(\varepsilon_i) = \sigma^2$. This implies that

$$E(Y_i) = \beta_0 + \beta_1 x_i,$$

and

$$Var(Y_i) = \sigma^2,$$

with the Y_i's taken as independent and normally distributed Random variables.

5.5.1 Partitioning the Total Variation

The total variation in the response variable (the dependent variable), y, will be decomposed into meaningful components for later use in testing the hypothesis about the coefficients of the regression line. Consider the identity

$$Y_i = \overline{y} + (\hat{y}_i - \overline{y}) + (Y_i - \hat{y}_i), \text{ for i =1, 2, 3, …, n.}$$

Equivalently we have

$$Y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (Y_i - \hat{y}_i), \text{ for i = 1, 2, 3, …, n.}$$

In other words, we can write

Total deviation = Deviation due to regression + Deviation about regression

Squaring both sides and summing over i = 1, 2, ….n, we have (interested reader might need to check this identity by recalling the identity $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$)

$$\sum_{i=1}^{n}(Y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$$

$$SS_{total} = SS_{regr} + SS_{residuals}$$

This decomposition shows that, of the total variation in the y's about their mean, some of the variation can be ascribed to the regression line and some to the fact that the actual observations do not all lie exactly on a regression line. It is seen that the value of SS_{regr} can be used as a measure of the importance of the fitted regression line. If SS_{regr} is large in relative to $SS_{residuals}$, this indicates the fitted line is significant. One measure for the importance of regression is the "Coefficient of Determination" r^2, which is given by $r^2 = SS_{regr} / SS_{total}$. The above sums of squares can be computed easily using the following formulas:

$$SS_{total} = \sum_{i=1}^{n}(Y_i - \overline{y})^2 = \sum_{i=1}^{n}Y_i^2 - \frac{\sum_{i=1}^{n}Y_i^2}{n},$$

$$SS_{regr} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = b_0\sum_{i=1}^{n}Y_i + b_1\sum_{i=1}^{n}X_iY_i - \frac{\sum_{i=1}^{n}Y_i^2}{n}, \text{ and}$$

$$SS_{residuals} = SS_{total} - SS_{regr} = \sum_{i=1}^{n}Y_i^2 - b_0\sum_{i=1}^{n}Y_i - b_1\sum_{i=1}^{n}X_iY_i.$$

What is that r^2? It is not surprising to see that it is the square of the linear correlation coefficient found above, when it needs to be calculated. If r^2 is found first, clearly we can find r by taking the square root. Another question: is $r > 0$ or $r < 0$? How do you decide? No doubt that finding r first is more enlightening than finding r^2 first. Here is another definition.

The coefficient of determination, r^2, measures the percentage of total variation in the response variable that is explained by the least square regression line based on the one explanatory variable x. it is a number between 0 and 1 inclusive, that is, $0 \le r^2 \le 1$. Recall that r^2 is the estimated value of the random variable R^2. Practically r^{2-} is expressed as a percentage, and it is < 100%. What about that difference percentage between the 100% and r^2? This difference is the unexplained variation in Y due to not including all the variables that the response variable can get affected by.

The results are summarized in the ANOVA Table below.

ANOVA Table

Source	df	SS	MS	E(MS)
Regression	1	SS_{regr}	$SS_{regr}/1$	$\sigma^2 + b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$
Residual	$n-2$	$SS_{residuals} = SS_{total} - SS_{regr}$	$\frac{SS_{residuals}}{n-2}$	σ^2
Total	$n-1$	SS_{total}		

(See Bakir and Shayib 1990)

5.5.2 Testing for Significant Regression

In the model $y = \beta_0 + \beta_1 x + e$, if $\beta_1 = 0$, then Y is not related to x by the first linear model, i.e., this first order model is incorrect. However if $\beta_1 \neq 0$, then there is a linear relationship between x and y as described in the model. At this time we cannot tell whether the relationship is positively related, i.e. y increases as x increases, or negatively related, i.e., x increases then Y decreases. Thus to test for significant linear (first order) relationship, between x and Y is equivalent to testing the following hypotheses:

H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.

The test statistic for the above hypothesis is given by

$$F = MS_{regr} / MS_{residuals}, \text{ with}$$

$$MSregr = \frac{SS_{regr}}{1}, \text{ and}$$

$$MSresidual = \frac{SS_{residuals}}{n-2}.$$

With the understanding of SS_{regr}, $SS_{residuals}$, as described above, with 1 is the degree of freedom for the numerator in the F expression, and n – 2 degrees of freedom for the denominator. Thus the test statistic F, as defined above, will have an F-distribution with (1, n – 2) degrees of freedom, when H_0 is true. Hence the critical region of size α, to reject H_0, using the F-distribution, is given by $F > f_{1-}$, where f_{1-} is that value of the F-distribution such that $P(F < f_{1-}) = 1-\alpha$.

EXAMPLE 5.5

The following data were collected on the amount of fertilizer and the corresponding yield of wheat. Test, at the 5% level of significance, for a significant first order (linear) regression of wheat on fertilizer.

Fertilizer (Lb/Acre) x	Yield (bushels/acre) y	x^2	y^2	xy
10	18	100	324	180
20	24	400	576	480
30	25	900	625	750
40	30	1600	900	1200
Total 100	97	3000	2425	2610

Solution:

The first order model (or the linear model) is given by

$$y = \beta_0 + \beta_1 x + e.$$

The above model is estimated by

$\hat{y} = b_0 + b_1 x$, where

$$b_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \right)}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \right)}$$

$$b_1 = \frac{\left(2610 - \frac{(100)(97)}{4} \right)}{\left(3000 - \frac{(100)^2}{4} \right)} = 0.37$$

And

$b_0 = \bar{y} - b_1 \bar{x} = 24.25 - (0.37)(25) = 15.0$

Thus $\hat{y} = 15.0 + 0.37x$.

The sums of the squares are given by

$$SS_{total} = 2425 - \frac{97^2}{4} = 72.75,\ SS_{regr} = (15.0)\ (97) + (0.37)(2610) - \frac{97^2}{4} = 68.45, \text{ and}$$

$$SS_{residuals} = SS_{total} - SS_{regr} = 72.75 - 68.45 = 4.30.$$

Using the classical method steps regarding the statistical hypothesis about the significance of the first order regression, we have.

1. H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$, this is a two-tailed test.
2. Based on the level of significance α =0.05, and the F-test with 1 and 2 degrees of freedom for the numerator and denominator respectively, the critical value is given as $f_{0.95} = 18.51$.
3. The test statistics is $F = MS_{regr} / MS_{residuals}$
4. F = 68.45/2.15 = 31.84
5. Since the calculated value of 31.84 > the critical value of 18.51, the null hypothesis is rejected.
6. It is concluded that there exists a significant linear relationship between the yield and the amount of fertilizer used. There is a positive, direct proportional relationship between the amount of fertilizer x and the yield Y.

Using the ANOVA setup above we have the following ANOVA Table

ANOVA Table

Source	df	SS	MS	F-ratio	
Regression	1	68.45	68.4	31.84	sig at 0.05
Residual	2	4.30	2.15		
Total (corrected)	3	72.75			(See Bakir and Shayib 1990)

The above test procedure for the significance of the linear regression has used the ANOVA technique. Example 5.5 was an example on that procedure. Recall we had two methods to test on a parameter of the population. The steps are identical in both cases, and the difference is in the test statistic. Just for comparison we will list the steps for the classical and p-value methods for testing the hypothesis regarding the slope coefficient β_1 in the least –squares regression linear model.

It is to be noted that to test the claim that two quantitative variables are linearly related, we make sure that the following two conditions are satisfied:

1. The sample is obtained using random sampling.
2. The residuals are normally distributed with a constant error variance.

Classical Method Steps:

1. State the null and alternative hypotheses.
   There are three ways to set up the null and alternative Hypotheses:

   d) Equal hypothesis versus not equal hypothesis: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$, two-tailed test
   e) At least versus less than: H_0: $\beta_1 \geq 0$ versus H_1: $\beta_1 < 0$, left-tailed test
   f) At most versus greater than: H_0: $\beta_1 \leq 0$ versus H_1: $\beta_1 > 0$, right-tailed test

2. Let α be the significance level, and based on the three cases in step 1, we have the following three cases that will go along for finding the critical values and rejection regions:

   a) For the two-tailed test there are two critical values: $-t_{\alpha/2}$ & $t_{\alpha/2}$, and the critical region is given by $|T| > t_{\alpha/2}$, as shown in the Figure 1, with each of the shaded areas equals $\alpha/2$.

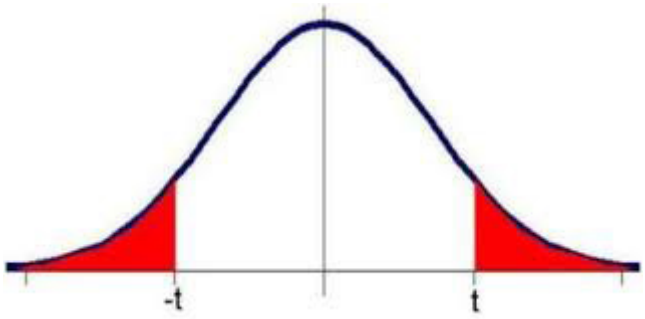


Figure 1

b) For the left-tailed test, there is the critical value of $-t_\alpha$, and the rejection region is given by $T < -t_\alpha$, and should be the one on the left side in Figure 1,
c) or the right-tailed test, again, there is on critical value given by t_α, and the rejection region is $T > t_\alpha$, as shown in the Figure 1, on the right side with area equals α.

3. The test statistic is given by $t = \frac{b_1 - \quad_1}{S_{b_1}}$, which follows Student's t-distribution with n-2 degrees of freedom ere n is the sample size, $\bar{x}$ is the mean, and Z has the standard normal distribution, N (0, 1).
4. The above test statistic is computed based on the information provided to us by the sample data, and on the assumption that the null hypothesis is true, i.e. $\beta_1 = 0$.
5. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

P-value Method Steps: For testing regarding the slope β_1, we have:

1. State the null and alternative hypotheses :
   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$, Two-tailed test
   b) At least versus less than: H_0: $\beta_1 \geq 0$ versus H_1: $\beta_1 < 0$, Left-tailed test
   c) At most versus greater than: H_0: $\beta_1 \leq 0$ versus H_1: $\beta_1 > 0$, right-tailed test

2. Let α be the significance level for the test.
3. For the test statistic we have $t = \frac{b_1 - \quad_1}{S_{b_1}}$.
4. The above test statistic is computed based on the information provided to us by the sample data, and on the assumption that the null hypothesis is true, i.e. $\beta_1 = 0$.

5. Apply how the p-value is calculated based on the type of your test, as shown above. The statistical decision will be made based on the case on hand whether we have a two-tailed, a left-tailed or a right-tailed test, by comparing the computed p-value, for the test, to the significance level stated in step 2. If the computed p-value is less than α, H_0 will be rejected; otherwise do not reject H_0. Using the T-Test, and technology we can have an exact value for the p-value, but only a range of values on it if things are done manually and based on the tables.
6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 5.6

Based on data in EXAMPLE **5.5,** test the significance of the least-squares regression linear model. Apply the above techniques with the LinRegTTest.

Solution:

Following Technology – Step-by-Step for testing on the significance of the linear model in Example 5.5, we have the following output.

LinRegTTest

y = a + bx

$\beta_1 \neq 0$ and $\rho \neq 0$

t = 5.64245

p = 0.03000

df = 2

a = 15

b = 0.37

s = 1.4663

$r^2 = 0.94089$

r = 0.969997

On applying the classical method with level of significance of 0.05, df = 4-2 = 2, and two-tailed test, the critical values are ± 4.303, and the rejection region is given by |T| > 4.303. From the output above we see that the calculated value of the test statistic is > the critical value, thus it is in the rejection region. So, the null hypothesis is rejected, and the two variable are positively associated since r > 0.

Applying the p-value method, based on the practical significance level of 0.03000 < 0.05, and therefore the null hypothesis is rejected.

The conclusions just reached are in line with the conclusion that was based on the ANOVA and the F-Test.

❖

5.6 Confidence Interval on β_0 and β_1

As it was the case when we constructed a confidence interval on one parameter of the population (i.e. on one proportion, or one mean, or one variance) or on two parameters of the population, we can construct a confidence interval on the coefficients in the least squares regression line, namely, the parameters β_0 and β_1. Clearly, those parameters needed to be estimated by b_0 and b_1 respectively. We refer the interested reader, for the proofs and realization of the following formulas, to Walpole and Myers. Thus we have

$$E(b_0) = \beta_0, \text{ Var}(b_0) = \frac{^2\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

$$E(b_1) = \beta_1, \text{ Var}(b_1) = \frac{^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

It is to be noticed that the estimators b_0 and b_1 are unbiased, with σ^2 to be estimated by $MS_{residuals}$.

Hence the 100(1- α) % C.I. on β_0 is given by:

$$b_0 \pm t_{1-\ /2}\left(MS_{residuals}.\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)^{1/2}$$

And The 100(1- α) % C.I. on β_1 is:

$$b_1 \pm t_{1-\ /2}\left(\frac{MS_{residuals}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)^{1/2}$$

When $t_{1-\ /2}$ is read from the t-table with n–2 degrees of freedom = residual degrees of freedom.

Remember there are equivalent forms for the above intervals in terms of the standard error.

For more C.I. on the mean of the response variable at a given value for the explanatory variable, or for a C.I. for the regression line at a given x value or for a C.I. on a predicted single y value at a specified x value, we refer the interested reader to Bakir and Shayib 1990, pp. 172–176. In addition to the above the interested reader can check that the random variable T^2 has an F-distribution.

From the model : y = β_0 + β_1x + e, again for the sake of completion, with the usual assumptions for the E's that are independent random variables with mean zero and variance σ^2, let us give, without derivation, the 100(1 –α) % C.I. for a mean response, i.e. on E(Y) = $_{y|x}$ = β_0 + β_1x, which depends on the x-value when it is given. The following formula can be used to construct a confidence interval on $_{y|x}$

$$\hat{y}-t_{1-\ /2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}\right)^{1/2} \quad < \quad _{y|x} < \hat{y}+t_{1-\ /2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}\right)^{1/2},$$

When the specified value for the explanatory variable is x_0, n is the sample size, and $t_{1-\ /2}$ is the critical value, with n–2 degrees of freedom.

The procedure for obtaining a prediction interval for an individual response is identical to that for finding a C.I. on the mean of that response. The difference is in the standard error. More variability is associated with one value than with the mean of some values. The following formula can be used to construct a prediction interval for a predicted y value at a specified x-value, x = x_0.

$$\hat{y} - t_{1-\ /2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}+1\right)^{1/2} < Y < \hat{y} + t_{1-\ /2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}+1\right)^{1/2},$$

EXAMPLE 5.7

A personnel officer, at a certain institution, believes that there is a relationship between a worker's age and absenteeism. In order to discover this relationship, she has compiled the following information for 10 randomly selected employees.

X (age)	19	22	25	27	30	33	36	39	42	57
Y(days absent)	8	10	9	7	5	6	5	4	2	4

a) Estimate the least-squares regression line of absenteeism on age.
b) Construct a 95% C.I. on the regression coefficients.
c) Construct a 95% C.I. on the regression line when x = 35.
d) Construct a 95% prediction interval for the number of days of absenteeism when the age is 35.

Solution: (a)

X (age)	19	22	25	27	30	33	36	39	42	57
Y(days absent)	8	10	9	7	5	6	5	4	2	4
X^2	361	484	625	729	900	1089	1296	1521	1764	3249
Y^2	64	100	81	49	25	36	25	16	4	16
XY	152	220	225	189	150	198	180	156	84	228

Using the formulas for the coefficients, we have

$$b_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}\right)}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} = -0.1755, \text{ and } b_0 = \bar{y} - b_1\bar{x} = 11.7915.$$

Based on the above calculations the predicted equation is given by $\hat{y} = 11.7915 - 0.1755x$.

To construct C.I. on the slope, we have $MS_{residual} = \frac{SS_{residuals}}{n-2} = 2.656$, The 100(1- α) % C.I. on β_1 is:

b. $$b_1 \pm t_{1-\ /2}\left(\frac{MS_{residuals}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^{1/2} = -0.1755 \pm 2.306 = -0.1755 \pm 0.1118 \text{ or}$$

$$-0.2873 < \beta_1 < -0.0638.$$

c. When x = 35, $\hat{y}$ then = 11.7915 – 0.1755x= 11.7915 – 0.1755(35) = 5.649. Also $(x_0 - \bar{x})^2$ = 4 .Hence

$$\hat{y} - t_{1-\alpha/2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}\right)^{1/2} < \mu_{y|x} < \hat{y} + t_{1-\alpha/2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}\right)^{1/2} \text{ or}$$

5.267 < $_{y|x}$ < 6.031.

d. For the prediction interval for the value of y when x = 35, we use

$$\hat{y} - t_{1-\ /2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}+1\right)^{1/2} < \text{Y} < \hat{y} + t_{1-\ /2}.S_e\left(\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+\frac{1}{n}+1\right)^{1/2}.$$

When x =35, $\hat{y}$ = 5.649, t.$_{975}$ = 2.306, thus we reach the following interval: 1.701 ≤ y ≤ 9.597.

❖ ----------

CHAPTER 5 EXERCISES

5.1 What does it mean to say that two variables are positively associated

5.2 What does it mean to say that the linear correlation coefficient between two variables equals 1? What would the scatter diagram look like?

5.3 What does it mean if r = 0

5.4 For the data set

X:	0	2	3	5	6	6
Y:	5.8	5.7	5.2	2.8	1.9	2.2

Draw a scatter diagram. Comment on the type of relation that appears to exist between x and y.

a) Determine the least squares regression line.
b) Graph the least squares regression line on the scatter diagram drawn in part a).

5.5 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 $\mathbf{y} = \beta_0 + \beta_1 x + E$, 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?

5.6 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	1.96
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 $\alpha = 0.05$.

5.7 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

a) Fit the model $\mathbf{y} = \beta_0 + \beta_1 x + E$
b) Test for significance regression, with $\alpha = 0.05$.
c) Obtain a 95% confidence interval for β_0 and β_1.

5.8 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:	26.825.4	28.9	23.6	27.7	23.9	24.7	28.1	26.9	27.4	22.6	25.6
Y:	26.527.3	24.2	27.1	23.6	25.9	26.3	22.5	21.7	21.4	25.8	24.9

a) Estimate the regression line $\mu_{y|n} = \beta_0 + \beta_1 x$.
b) Estimate the shear resistance for a normal stress of 24.5 pounds per square inch.

5.9 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

b. Test for significant correlation at 5% level.

5.10 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

5.11 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.

a) Following the suggested procedure in Example 3, find the least square estimates of γ and C
b) Estimate P when V= 80 cubic centimeters.

5.12 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							

In the following Exercises: 5.13–5.16, you are given the regression equation:

a) Calculate the predicted values
b) Calculate the residuals, and the sum of their squares
c) Construct a scatterplot of the residuals versus the predictyed values
d) Construct a normal probability plot of the residuals using technology
e) Verify that the regression assumptions are valid

5.13 $\hat{y} = 2.5x + 13.5$

x	1	2	3	4	5
Y	15	20	20	25	25

5.14 $\hat{y} = 3.2x + 8$

x	-5	-4	-3	-2	-1
y	0	8	8	16	16

5.15 $\hat{y} = -2x + 8$

x	1	2	2	2	3
y	6	5	4	3	2

5.16 $\hat{y} = -0.5x + 104$

x	10	20	30	40	50
y	100	95	85	85	80

Practical Problem and Project on Chapter 5

A car dealership believes that the relationship between the number of cars sold per week, Y, and the average number of sales consultants, x, on the floor each day during the week, is linear. The following data was provided for analysis:

X	6	3	4	6	2
Y	20	11	10	18	6

a) Draw a scatter diagram for the data
b) Calculate the least-squares regression linear equation
c) Plot the least-squares line on the scatter plot
d) Give an explanation, and what each of the following stands for:
   1. The least –squares estimate of s,
   2. The y-intercept, and
   3. The slope of the line.

e) Based on the linear relationship that has been calculated, approximate how many cars should the company expect to sell in a week, if, on the average, there are 5 sales consultants on the floor each day.
f) Assuming the distribution, of the cars sold per week, follows a normal distribution with mean as found in part e) and standard deviation s, find the minimum and maximum number of cars sold per week using the empirical rule.

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 curser on "**ON**" and press ENTER, to turn the plots on
4. Highlight the scatter diagram icon, and press ENTER. Make sure that Xlist s 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

6. 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.
7. Enter the explanatory variable in column A, and the response variable in column B.
8. Select **TOOLS** and highlight **Data Analysis**…
9. Select **Regression** option.
10. 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**.

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 R2. 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 L1 and the response variable in L2.
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**.

6 Other Tests and Analysis Of Variance

Outline

6.1 Introduction

For example; are children born equally on the days of the week? When rolling a die, how do you check if that die is fair?

In the previous chapters we have seen how to

1. Collect, organize, summarize and analyze data
2. Test statistical hypotheses concerning population means, populations' proportions, and populations' variances.

The idea of testing statistical hypotheses can be extended to other situations that involve different parameters and use different test statistics. Whereas the standardized test statistics that appeared in earlier chapters followed either a normal, Student t-distribution Chi-square tests, in this chapter the tests will involve two other very common and useful distributions, the chi-square and the F-distributions. The chi-square distribution a particular probability distribution specified by a number of degrees of freedom, df, arises in tests of hypotheses concerning the independence of two random variables and concerning whether a discrete random variable follows a specified distribution, that is, testing on goodness of fit to some assumptions. The F-distribution, a particular probability distribution specified by two degrees of freedom, df_1 *and* df_2 arises in tests of hypotheses concerning the equality if two populations' variances are equal. It is also used in testing statistical hypotheses concerning whether, or not, three or more populations' means are equal.

In Chapter 4 we discussed procedures for testing a statistical hypothesis about one population parameter, or two populations' parameters. In chapter 6, we seek an extension of the previously described procedures to test statistical hypotheses about three or more means. The extended procedure, to do this testing, is called Analysis of Variance (ANOVA). In general, we call the variability among the sample means **the between-sample variability** and the variability of each sample **the within-sample variability.** ANOVA works by comparing the variability **between** the samples to the variability **within** the samples. To carry out such a task, we need to introduce, and learn about, a new distribution, the F-distribution.

The F distribution was named in honor of the "grandfather of Statistics," **Sir Ronal A. Fisher.** The F distribution is like the Chi-square distribution, it is right-skewed, a non-negative, and has an infinite number of F curves. It is characterized by the degrees of freedom of the numerator and the denominator that can take infinite values, theoretically speaking. The resemblance between the F and the Chi-Square distributions is not surprising since, as we will find out later, that the F distribution is a ratio of two chi-square distributions with degrees of freedom df_1 for the numerator and df_2 for the denominator.

Properties of the F Distribution

1. The total area under the F-distribution curve equals 1.
2. The value of the F random variable is never negative, so the F curve starts at 0. However, it extends indefinitely to the right,
3. The curve never intersects the horizontal axis. The horizontal axis acts as a horizontal asymptote for the F-distribution curve.
4. Because of the property 2, the distribution is positively skewed, or right-skewed.
5. There is a different F-distribution curve for each pair of the degrees of freedom, df_1 and df_2, the degrees of freedom of the numerator and the degrees of freedom of the denominator, respectively.
6. The F-distribution is continuous, and thus we find probabilities associated with the values of F, as areas under the curve and between any two specified points on the horizontal axis within the range of values for the random variable, just as we did with the normal, t-, and chi-square distributions.
7. Let U and V be independent Chi-square random variables with r_1 and r_2 degrees of freedom, respectively, then the random variable is defined by

$$F = \frac{U / r_1}{V / r_2}$$

This random variable will have an F-distribution with r_1 degrees of freedom for the numerator, and r_2 degrees of freedom for the denominator. Based on that, we write $F(r_1, r_2)$. Clearly the reciprocal of F,

$$\frac{1}{F} = \frac{V / r_2}{U / r_1},$$

will be another F-distribution but with reversed degrees of freedom, with regard to the original F-distribution. It is given by $F(r_2, r_1)$.

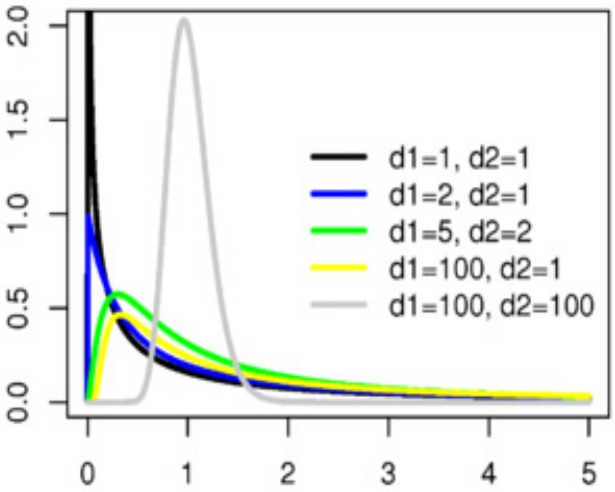


Figure 1

Figure 1 displays an F-distribution with different degrees of freedom for the numerator and denominator. As it is clear from the graph, the F-distribution is positively skewed. For large degrees of freedom, the graph will have a high peak, and a very thin right tail.

In section 6.2 the Goodness-Of-Fit will be introduced. The test for independence and Homogeneity of proportions will be discussed in 6.3. In section 6.4, below, we will introduce the one-way ANOVA. We will restrict ourselves to the simplest case of ANOVA when comparing a set of "J" treatments with no restriction on randomization. This presentation will use the linear model in its two forms, as it will become clear later.

6.2 Goodness-of-Fit Tests

Data, or measurements from experiments, can be qualitative or categorical rather than quantitative like many of the measurements discussed in the earlier chapters. Earlier we have dealt mostly with quantitative data, for estimating the perameters of the populations, finding confidence intervals about them and testing statistical hypotheses concerning those parameters as well. When data comes in, and it is summarized in categories, or classes, the measurements will be given as counts that fall in each of the distinct categories associated with the varable. For Example

- Employees may be classified into one of five income brackets.
- Faculty members at any institution can be classified in more than 3 categories.
- Cars might fall into one of 4 or 5 types of vehicles.

In this section, we like to answer the question just raised in the first sentence in this chapter: How do we check if a die is fair? In chapter 2 we considered discrete random variables and especially the binomial random variable where there are only two possible outcomes for the experiment. As example for the binomial random variable is this question again: Asking the students individually: Did you have eggs this morning for breakfast? We are not interested in how did he/she have their eggs, as much we are concerned about whether the answer is Yes or No. On the other hand, if the question was put in this way: What did have for breakfast this morning? Clearly there are more than two choices for the answer. Now let us consider the following type of random variable that will go with the question just been asked, i.e. a question which can have more than two possible values. Hence we are defining what is called a **Multinomial Random Variable.**

A Multinomial Random Variable is that variable which satisfies the following conditions:

a) Each independent trial of the experiment has k possible outcomes, k = 3, 4, …
b) The i^{th} outcome occurs with probability $p_i > 0$, i = 1, 2, …, k, (that is is the population proportion for the ith outcome, or category.
c) $\sum p_i = 1$ (by the law of total probability).

Data from a multinomial random variable will be labeled as following a multinomial distribution.

EXAMPLE 6.1

In a class of 40 students taking stat 1350, there was the following question: What is your political affiliation?

Solution:

The answer is a multinomial random variable since the answere could be; Republican, Democrat, Independent, TEA party, or others (No affiliation). Clearly this is a multinomial data. If in the class of 40 students we have 10 Rerublicans, 12 democrats, 8 independent, 5 TEA Party, and 5 with no affiliation. Hence we have, based on k = 5, we see that
$p_1 = .25$, $p_2 = 0.3$, $p_3 = 0.20$, $p_4 = 0.125$, $p_5 = 0.125$. Clearly we see that $\sum_1^5 p_i = 1$.

❖

Is the class above a good representative of the population in the USA? If those proportions are correct throughout the nation then a class of 60 students will be composed of how many of each political category? Recall how did we find the mean of a binomial random variable? Is it binomial here? Yes, it is as far as each category is concerned. If you ask are a republican? That is a binomial Random variable with two outcomes; yes or no. Thus the expected value in this case will be: The Number of trials * the proportion of being a republican = np. To check the validity of this claim, we will introduce a new test called Chi-Square, χ^2 Goodness Of-Fit-Test.

A χ^2 Goodness of fit test is a statistical hypothesis test used to determine whether a random variable follows a particular distribution or not. In this case the two Hypotheses will be stated as follows:

H_0: The random variable follows a particular distribution, i.e. with specified and assigned values for the probabilities of the categories,

H_1: The random variable does not follow the particular distribution stated in.

In statistics, a goodness of fit test checks on a collection of data, in categories, to see if that data falls according to a specified probability distribution. The data collected is a sample, and we use the goodness of fit test to see if the sample is consistent with the null hypothesis. The sample will be summarized in a set of "observed" frequencies under the categories of interest. The χ^2 Goodness of fit test then compares the observed frequencies with the expected frequencies calculated based on the assumed distribution. We can see $\sum(O - E) = 0$, and this is not agood value to uase in a test. Based on that we see that The χ^2 Goodness of fit test is based on the following formula: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. The test will reject the null hypothesis for large value of $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, where we have k categories for the multinomial random variable, O_i is the observed frequency in category i, and E_i is the expected frequency based on the assumed probability distribution. Having said that so far, let us check on our fair die!

EXAMPLE 6.2

We rolled a normal die 60 times, and got the following outcomes

Category	1	2	3	4	5	6
Observed	15	9	12	10	8	6

Solution:

As it was the case with a binomial random variable, we can ask the question: did you get one or not? So, each outcome can be looked at as binomial random variable, and thus the expected value is given by np, where n is the total number of trials and p is the probability to get that outcome. Therefore we have the following table

Category	1	2	3	4	5	6
Observed	15	9	12	10	8	6
Expected	10	10	10	10	10	10
O – E	5	-1	2	0	-2	-4
$(O - E)^2$	25	1	4	0	4	16

The comparison of the observed frequencies with the expected ones is the basis for the Chi-Square goodness-of-fit test. Clearly, we have $\sum E_i = n = 60$, in this case where we have k = 6 categories, and each probability $p_i = 1/6$, i = 1, 2, 3, 4, 5, 6.

From ther table above, we see: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.1(25+1+4+0+4+16) = 5.$ Here, k = 6, and thus $v = 5$, degrees of freedom. Based on the significance level of 0.05, we find, from the Chi-square tabkle, that the critical value for the test is 11.071. Thus H_0 is not rejected, and the die is fair.

❖

As it was the case with any statistical test, there are some conditions to be satisfied before carrying the test. The Chi-squre test is not an exception. Here are the conditions in order to carry the test. For a multinomial random variable, with k categories and n independent trials, let O_i be the observed frequency in category i, and Ei is the expected frequency for the same category, then the test statistic for a goodness-of-fit test is given by $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. This Random variable approximately follows a χ^2 distribution with k-1 degrees of freedom, if the following conditions are satisfied:

1. None of the expected frequencies is less than 1.
2. At most 20% of the expected frequencies are less than 5.

As it was discussed in Chapter 4, we can carry the process of testing a statistical hypothesis by the classical method or the p-value method. We are not listing the steps here again, (check chapter 4). It is to be noticed here that, we will consider a one-tailed test, i.e. a right-tail test, and reject the null hypothesis for a large value of the test statistic. The goodness-of-fit test is used when we have ONE Multinomial Random Variable (one factor at different levels) which is characterized in k categories. In case we have two multinomial random variables the above goodness-of-fit does not apply. This is the the topic for the next section.

6.3 Contingency Tables

In section 6.2 we learned that The Chi-Square Goodness-of-fit-Test, using the Chi-Square distribution could help to determine that if a model falls into some specified distribution. A problem frequently encountered in the analysis of count data concerns assessment of the independence of two methods for classification. For example we may classify a sample of people by gender and by their political affiliation in order to test the hypothesis that political affiliation and gender are independent. This kind of classifying our sample will generate a crosstabulation or a two-way table, and this will form what we call a contingency table. The categories of one variable label the rows, say r rows; and the categories of the other variable will label the columns, say c columns. Thus there are rXc cells in the table.Each cell will contain the number of observations that fit the categories of that row and that column.

Suppose that we are interested in classifying the type of defects found in a manufactured product according to (1) the type of defect and (2) the production shift. A total of n = 309 defects were recorded and the defects were classified as one of four types, A, B, C, or D. At the same time each item was identified according to the shift during which it was manufactured. Based on that classification we come up with the following contingency table of 3 rows and 4 columns,(it could be vice-versa), or 3×4 contingency table. Our objective is to test the Null hypothesis

$H_{0:}$ The type of defect is independent of the shift, against the alternative hypothesis
H_1: The categorization schemes are dependent.

Let p_A equal the unconditional probability that an item has a defect of type A. Similarly, define, ,and as the probabilities of observing the three other types of defects. These are called the column probabilities and their sum is one as you have expected. In a like manner let p_1, p_2, p_3 be the row probabilities that a defective item was produced by the shift 1, 2 or 3, where $p_1+p_2+ p_3 = 1$.

If the two classifications are independent of each other, then each cell probability will be equal to the product of the row probability and the column probability. Hence for cell C_{11} the probability for an item to be there is $p1 \cdot pA$, and so on for the other cells. But it is to be noticed here that neither the row probabilities neither the column probabilities are give. We appeal to the relative frequency definition of probabilities to find all those probabilities for the specified cells. Thus

P_i = (total of row i) / grand total of the observation $= \frac{r_i}{n}$ for i = 1, 2, 3; and r_i is the sum of the ith row. Similarly, the probabilities for the types of defects, i.e, the columns, can be found by the same way. What about the probability of each cell. These are defined in the same way as relative probabilities. Hence we have

For i = 1, 2, 3; and j = A, B, C, D. $p_{ij} = \frac{n_{ij}}{n}$

Likewise, viewing row i as a single cell, the probability for row i is give by $\hat{P}_i = r_i/n$.

If we look at esch cell individually and ask ourselves, is that item here or not? Thus we have a Bernoulli experiment, and we recall how to find the mean of a binomial random variable. The mean = number of trials x the probability of success, i.e. the probability falling into that cell. In otherwords we have, based on the independence of the rows and the columns we find that, for i = 1, 2, 3; and j = A, B, C, D, as above

$$E(n_{ij}) = n(p_{ij}) = n(p_i X\ p_j).$$

Because of independence we find, rellying on the relative frequency for the definition of probability, that

$$E(\hat{n}_{ij}) = \frac{r_i . c_j}{n}$$

Where r_i and c_j are the totals of the ith row and jth column respectively. Based on what had been said and done we have the following table:

Type of defect

Shift	A	B	C	D	Total
1	15 (22.51)	21 (20.99)	45 (38.94)	13 (11.56)	94
2	26 (22.99)	31 (21.44)	34 (39.77)	5 (11.81)	96
3	33 (28.50)	17 (26.57)	49 (49.29)	20 (14.63)	119
Total	74	69	128	38	309

The numbers in the parantheses in each cell is the expected number of observations to be there based on the independence of the rows and columns, and it is found by the formula $E(\hat{n}_{ij}) = \frac{r_i . c_j}{n}$. For example, 22.51 = (74)(94)/ 309, and son for the other cells.

The test of independence is a Chi-square test using the Chi-square statistic defined by

$$^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For i = 1, 2, 3; and j = A, B, C, D; with O representing the observed itmes while E stands for the expected items in the $(i,j)^{th}$ cell.

As we are aware that the Chi-square distribution is characterized by its degrees of freedom, For the contingency table the number of degrees of freedom is given by (r-1)(c-1), where r is the number of rows and c is the number of columns, (for more detailed calculations for the degrees of freedom, see Wackerly et al 2008). Using the data provided in the table we see that

$$^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ii}} = \frac{(15 - 22.51)^2}{22.51} + \ldots + \frac{(20}{}$$

= 19.17.

The critical value for the test is, with 0.05 level of significance, 12.5916, for 6 degrees of freedom. Since 19.17 > 12.5916, the null hypothesis is rejected. Therefore the conclusion is that the type of defect is dependent on the shift. The interested student can work exercise 6.6 and consult Wackerly et al 2008 for more exercises.

6.4 The one-Way Analysis of Variance

The one-way analysis of variance (ANOVA) is used to determine whether there are any significant differences among the means of three or more independent (unrelated) populations. This guide will provide a brief introduction to the one-way ANOVA, including the assumptions of the test and when you should use this test.

The one-way ANOVA compares the means among the groups you are interested in and determines whether any of those means are significantly different from each other. Specifically, it tests the null hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$

where μ_i, i =1,2,…, k, is the mean of the i^{th} population and k is the number of populations. If, however, the one-way ANOVA returns a significant result, we accept the alternative hypothesis: H_1, which is that there are at least 2 populations' means that are significantly different from each other.

At this point, it is important to realize that the one-way ANOVA is an *omnibus* test statistic and cannot tell you which specific groups were significantly different from each other only that at least two groups were. To determine which specific groups differed from each other, you need to use a *post-hoc* test. Post-hoc tests are described in Montgomery 2001, and Bakir and Shayib 1990.

The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values.

Typically, however, the one-way ANOVA is used to test about differences among at least three groups, since the two-group case can be covered by a t-test (Gosset 1908). When there are only two means to compare, the t-test and the F-test are equivalent; and the relation between ANOVA and T is given by $F = T^2$.

6.4.1 Assumptions

The results of a one-way ANOVA can be considered reliable as long as the following assumptions are met:

- Response variable are normally distributed (or approximately normally distributed).
- Samples are independent.
- Variances of populations are equal.
- Responses for a given group are independent and identically distributed normal random variables (not a simple random sample (SRS)).

ANOVA is a relatively robust procedure with respect to violations of the normality assumption; Kirk, R.E. 1995. If data are ordinal, a non-parametric alternative to this test should be used such as Kruskal–Wallis one-way analysis of variance, (interested reader may consult the references on the subject).

6.4.2 The case of fixed effects, fully randomized experiment, unbalanced data

The model will be the normal linear model that describes treatment groups with probability distributions which are identically bell-shaped (normal) curves with different means. Thus fitting the models requires only the means of each treatment group and a variance calculation (an average variance within the treatment groups is used). Calculations of the means and the variance are performed as part of the hypothesis test. The commonly used normal linear models for a completely randomized experiment are; (see Montgomery, D.C. 2001)

$$Y_{ij} = \mu_j + E_{ij} \text{ (The means model)} \qquad i = 1, 2, \ldots, n_j \text{ and } j = 1, 2, \ldots, k$$

Or

$$Yij = \mu + T_j + E_{ij} \text{ (The effects model)} \quad i = 1, 2, \ldots, nj \text{ and } j = 1, 2, \ldots, k.$$

Where: i = 1, 2, …,n_j is an index over experimental units,

j = 1, 2, …, k is an index over treatment groups
n_j is the number of experimental units in the jth treatment group
$N = \sum n_j$ is the total number of experimental units
Y_{ij} are the observations
μ_j is the mean of the observations for the jth treatment group
μ is the grand mean of the observations
$T_j = \mu_j - \mu$ is the jth treatment effect, a deviation from the grand mean
$\sum T_j = 0$, and $E \sim N(0, \sigma^2)$, E_{ij} are normally distributed zero-mean random errors.

The E_{ij} are called "random errors", "random residuals", or "experimental errors". They represent effects of all extraneous fluctuations that are beyond the control of the researcher. Large uncontrolled variations are common in most sciences. There are two main sources of experimental errors:

1. The first source is the inherited variability in the experimental material to which the treatments are applied. An experimental unit is a group of material to a single treatment is applied in a single trial. It could be a lot of land, a patient in a hospital, a chemical batch, etc. Such units usually produce different results even when subjected to the same treatment. These differences, whether large or small, contribute to the experimental errors.

2. The second source of random error is lack of uniformity in the physical conduct of the experiment, in other words, failure to standardize the experimental technique.

6.4.3 The data and statistical summaries of the data

One form of organizing experimental observations Y_{ij} is with groups in columns:

Treatment	1	2	….	k
	Y_{11}	Y_{12}	….	Y_{1j}
	Y_{21}	Y_{22}	….	Y_{2j}
	.	.	…	.
	.	.	….	.
	$Y_{n1,1}$	$Y_{n2,2}$	….	$Y_{nj,j}$
Totals	$Y_{.1}$	$Y_{.2}$	...	$Y_{.k}$
Means	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	...	$\overline{Y}_{.k}$
Variances	$S_{.1}^2$	$S_{.2}^2$	...	$S_{.k}^2$

The grand mean and grand variance are computed from the grand sums, not from group means and variances.

6.4.4 The hypothesis test

The purpose is to test that all k treatment have equal effects. This hypothesis is

$H_0 =$ $_1 =$ $_2 = \ldots =$ $_k$ Versus H_1: Not all means are equal.

The Model $Y_{ij} = \mu + T_j + E_{ij}$ (The effects model) i = 1, 2, …,n_j and j = 1, 2, …, k, can be associated with the simple identity

$$Y_{ij} = \overline{Y}_{..} + (\overline{Y}_{.j} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{.j}) .$$

Given the summary statistics, the calculations of the hypothesis test are shown in tabular form.

Source of Variation	df	SS	MS	E(MS)	F-ratio
$Treatments(betweenGroups)$	k-1	$\sum_{j}^{k}\frac{y_{.j}^2}{n_j}-\frac{y_{..}^2}{N}$	$\frac{SS_{treat}}{k-1}$	$\sigma^2+\frac{\sum_{j=1}^{k}n_j t_j^2}{k-1}$	$\frac{MS_{treat}}{MS_{error}}$
Error	N – k	Difference	$\frac{SSerror}{N-k}$	σ^2	
Total	N -1	$\sum\sum y^2{}_{ij}-\frac{y_{..}^2}{N}$			

The hypotheses to be tested are

$$H_0 : t_1 = t_2 = ... = t_k = 0 \quad agianst$$
$$H_1 : some\, t_j \neq 0, for\ some\ j.$$

The test statistic is $F = \frac{MS_{treat}}{MS_{error}}$,, which has an f-distribution with (k-1, N-k) degrees of freedom, when H_0 is true.

Critical region of size: Reject H_0 if $F > f_{1-\alpha}$, where $f_{1-\alpha}$ is the (1- α) percentile point of the F-distribution with the above degrees of freedom.

The rationale of the ANOVA test is as follows: If H_0 is true, then each $t_j = 0$, then E (MS_{treat}) = σ^2 = E (MS_{error}) in the ANOVA table. This will imply that F =1, otherwise F > 1 when H0 is false because in that case $MS_{treat} > MS_{error}$. Hence we reject H_0 if F is large.

While two columns of SS are shown for their explanatory value, only one column is required to display results.

MS_{Error} is the estimate of the variance corresponding to σ^2 of the model.

6.4.5 Analysis summary

The core ANOVA analysis consists of a series of calculations. The data is collected in tabular form. Then

- Each treatment group is summarized by the number of experimental units, two sums, a mean and a variance. The treatment group summaries are combined to provide totals for the number of units and the sums. The grand mean and grand variance are computed from the grand sums. The treatment and grand means are used in the model.
- The three DFs and SSs are calculated from the summaries. Then the MSs are calculated and a ratio determines F.
- A computer typically determines a p-value from F which determines whether treatments produce significantly different results. If the result is significant, then the model provisionally has validity.

If the experiment is balanced, all of the nj terms are equal, so the SS equations simplify. In a more complex experiment, where the experimental units (or environmental effects) are not homogeneous, row statistics are also used in the analysis. The model includes terms dependent on *i*. Determining the extra terms reduces the number of degrees of freedom available.

EXAMPLE 6.3

Four sections of the same elementary course in statistics were taught by the same instructor. The final grades out of 20 were recorded as follows:

Section	Grades						Totals
1	12	10	7	8	9	14	60
2	12	16	15	9			52
3	9	7	6	11	7		40
4	12	8	8	10			38
Total							190

Is there a significant difference in the average grades of the four sections? Use a 0.05 level of significance.

Solution:

We will be using the Model

$Y_{ij} = \mu + T_j + E_{ij}$ (The effects model) i = 1, 2, …,n_j and j = 1, 2, …, 4; with $n_1 = 6$, $n_2 = 4$, $n_3 = 5$, and $n_4 = 4$.

The hypotheses to be tested are

$$H_0 : t_1 = t_2 = ... = t_k = 0 \quad agianst$$
$$H_1 : some\, t_j \neq 0, for\ some\ j.$$

Using the data provided, we have the following computations:

$S_{Stotal} = 148$, $SS_{treat} = 57$, thus $SS_{error} = 91$, which will make the following table:

ANOVA Table

Source	df	SS	MS	F-ratio
Treatment	3	57	19	3.13
Error	15	91	6.07	
Total	18	148		

With the critical value of $f_{.95}(3, 15) = 3.29$, H_0 is not rejected, since 3.13 < 3.29. The conclusion is that there is no difference in the means of the four sections on that final exam.

On another way, if the level of significance has changed to be 0.10,we find that $f_{.90}$ (3,15) = 2.49, and since 3.13 > 2.49, in this case the null hypothesis will be rejected indicating that there is a difference between the means of the four sections on this final exam.

The interested reader, who likes to make the comparison after the ANOVA, or carrying the test of hypotheses involving a linear combination of the treatment effects, can consult Bakir and Shayib 1990, or other references of his choice.

CHAPTER 6 EXERCISES

For Exercises 1–3, determine whether the distribution is multinomial.

6.1 A random sample of 25 students is drawn from the class of 2013, and the major in their degrees is observed.

6.2 We select 5 students from a group of 25 students at random and without replacement, and we observed student's class: Freshman, sophomore, junior, or senior.

6.3 In a class of 40 students, we ask: your state of residence is.

6.4 If the alternative hypothesis takes the form
$H_{0:}$ $p_1 = .5$, $p_2 = 0.3$, $p_3 = 0.20$, n = 100
H_1: The random variable does not follow the distribution specified in H_0
a) Find the expected frequencies
b) Determine whether the conditions for performing the χ^2 goodness of fit test.

6.5 If the following values are given O_i: 10 12 14, and E_i: 12 12 12. Calculate the value of the test Statistic.

6.6 Astudy of the amount of violence viewed on telvesion as it relates to the age of the viewer yielded the following resuls as tabulated below

	Age		
Viewing	16-34	35-54	55 and over
Low Violence	8	12	21
High violence	18	15	7

Do the data indicate that viewing of violence is not independent of age of viewer, at the 5% significance level?

For exercises 7–10, calculate the test statistic for the goodness-of-fit test.

a) df_1 and df_2 b) $\overline{\overline{x}}$ c) SSTR d) SSE e) SST

6.7

	Sample A	Sample B	Sample C
Mean	10	12	8
Standard Deviation	1	1	1
Sample Size	5	5	5

6.8

	Sample A	Sample B	Sample C	Sample D
Mean	10	12	8	14
Standard Deviation	1	1	1	1
Sample Size	5	5	5	5

6.9

	Sample A	Sample B	Sample C	Sample D
Mean	50	75	100	125
Standard Deviation	5	4	6	5
Sample Size	100	150	200	250

6.10

	Sample A	Sample B	Sample C	Sample D
Mean	0	10	20	10
Standard Deviation	1.5	2.25	1.75	2.0
Sample Size	50	100	50	100

In exercises 11–14, refer to the exercises cited and calculate the following measures

a) MSTR
b) MSE
c) Fdata
d) Construct the ANOVA table.

6.11 Exercise 7.

6.12 Exercise 8.

6.13 Exercise 9.

6.14 Exercise 10.

6.15 It is suspected that four filling machines, in a plant, are turning out products of no-uniform weight. An experiment is run and the data, in ounces, are as follows:

A	12.25	12.27	12.24	12.25	12.20
B	12.18	12.25	12.26	12.22	12.19
C	12.24	12.23	12.23	12.20	12.16
D	12.20	12.17	12.19	12.18	12.16

a) Write out the model and the ANOVA Table.
b) Test for a machine difference at the 5% level of significance.
c) If the test in b) is significant, Perform Duncan's range test
d) Test that the average of A and B = the average of C and D.

6.16 The following data show the effects of four operators, chosen at random from all operators at a certain factory, on the output of a particular machine:

I.	175.4	171.7	173.0	170.5
II.	168.5	162.7	165.0	164.1
III.	170.1	173.4	175.7	170.7
IV.	175.2	175.7	180.1	183.7

a) Perform the analysis of variance(random effects),
b) Estimate the operator variance component
c) Estimate the experimental error variance component.

6.17 Perform Duncan's multiple Range Tests on the following data. α = 5%, 7 treatments each of which having 5 observations, MSerror = 4.0, $\bar{y}_1$ = 19.84, $\bar{y}_2$ = 16.75, $\bar{y}_3$ = 20.00, $\bar{y}_4$ = 20.50, $\bar{y}_5$ = 15.00 $\bar{y}_6$ = 25.10, $\bar{y}_7$ = 25.70. (Check some references to do this exercise.)

7 Appendix A Tables

	Column									
Row										
	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
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 ≤ 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	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
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

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

Table III Binomial Distribution

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

Table IV t-Distribution

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
1.306	1.688	2.028	2.434	2.719	3.333	
1.305	1.687	2.026	2.431	2.715	3.326	
1.304	1.686	2.024	2.429	2.712	3.319	
1.304	1.685	2.023	2.426	2.708	3.313	
1.303	1.684	2.021	2.423	2.704	3.307	
1.303	1.683	2.020	2.421	2.701	3.301	
1.302	1.682	2.018	2.418	2.698	3.296	
1.302	1.681	2.017	2.416	2.695	3.291	
1.301	1.680	2.015	2.414	2.692	3.286	
1.301	1.679	2.014	2.412	2.690	3.281	
1.300	1.679	2.013	2.410	2.687	3.277	
1.300	1.678	2.012	2.408	2.685	3.273	
1.299	1.677	2.011	2.407	2.682	3.269	
1.299	1.677	2.010	2.405	2.680	3.265	
1.299	1.676	2.009	2.403	2.678	3.261	

Table IV t-Distribution

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

df	0.90	0.95	0.975	0.99	0.995	0.999
76.	1.293	1.665	1.992	2.376	2.642	3.201
77.	1.293	1.665	1.991	2.376	2.641	3.199
78.	1.292	1.665	1.991	2.375	2.640	3.198
79.	1.292	1.664	1.990	2.374	2.640	3.197
80.	1.292	1.664	1.990	2.374	2.639	3.195
81.	1.292	1.664	1.990	2.373	2.638	3.194
82.	1.292	1.664	1.989	2.373	2.637	3.193
83.	1.292	1.663	1.989	2.372	2.636	3.191
84.	1.292	1.663	1.989	2.372	2.636	3.190
85.	1.292	1.663	1.988	2.371	2.635	3.189
86.	1.291	1.663	1.988	2.370	2.634	3.188
87.	1.291	1.663	1.988	2.370	2.634	3.187
88.	1.291	1.662	1.987	2.369	2.633	3.185
89.	1.291	1.662	1.987	2.369	2.632	3.184
90.	1.291	1.662	1.987	2.368	2.632	3.183
91.	1.291	1.662	1.986	2.368	2.631	3.182
92.	1.291	1.662	1.986	2.368	2.630	3.181
93.	1.291	1.661	1.986	2.367	2.630	3.180
94.	1.291	1.661	1.986	2.367	2.629	3.179
95.	1.291	1.661	1.985	2.366	2.629	3.178
96.	1.290	1.661	1.985	2.366	2.628	3.177
97.	1.290	1.661	1.985	2.365	2.627	3.176
98.	1.290	1.661	1.984	2.365	2.627	3.175
99.	1.290	1.660	1.984	2.365	2.626	3.175
100.	1.290	1.660	1.984	2.364	2.626	3.174
	1.282	1.645	1.960	2.326	2.576	3.090

Table IV t-Distribution

$\lambda = E(x)$									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.905	0.819	0.741	0.67	0.607	0.549	0.497	0.449	0.407
1	0.995	0.982	0.963	0.938	0.607	0.878	844	8099	0.772
2	1	0.999	0.996	0.992	0.91	0.977	0.966	0.953	0.937
3		1	1	0.999	0.986	0.977	0.994	0.991	0.987
4				1	1	1	0.999	0.999	0.998
5							1	1	1
6									
$\lambda = E(x)$									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
0	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.15
1	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434
2	0.9	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704
3	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875
4	0.995	0.992	0.989	0.986	0.981	0.976	0.97	0.964	0.956
5	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.99	0.987
6	1	1	1	0.999	0.999	0.999	0.998	0.997	0.997
7				1	1	1	1	0.999	0.999
8								1	1
$\lambda = E(x)$									
x	2.2	2	2.6	2.8	3	3.2	3.4	3.6	3.8
0	0.111	0.091	0.074	0.061	0.05	0.041	0.033	0.027	0.022
1	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126	0.107
2	0.623	0.57	0.518	0.469	0.423	0.38	0.34	0.303	0.269
3	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515	0.473
4	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706	0.668
5	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844	0.816
6	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927	0.909
7	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969	0.96
8	1	0.999	0.999	0.998	0.996	0.994	0.992	0.988	0.984
9		1	1	0.999	0.999	0.998	0.997	0.996	0.994
10				1	1	1	0.999	0.999	0.998
11							1	1	0.999
12		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
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
16										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	0.446
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	0.9653	0.587	0.522	0.458	0.397	0.341
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

	$\lambda = E(x)$									
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

	Chi-	Square		P(X ≤ x)				
	0.010	**0.025**	**0.050**	**0.100**	**0.900**	**0.950**	**0.975**	**0.990**
r								
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
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	df of Num									
P(F<f)	Den	1	2	3	4	5	6	7	8	9	10
0.95		161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9
0.975	1	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		18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4
0.975	2	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		10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.975	3	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		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96
0.975	4	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		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.975	5	10.01	8.43	7.76	70.39	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		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06
0.975	6	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		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
0.975	7	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		5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35
0.975	8	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		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
0.975	9	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		4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
0.975	10	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	0.932	24	0.404
11	0.602	25	0.396
12	0.576	26	0.388
13	0.553	27	0.381
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

8 References

1. Bakir, S.T. and M.A. Shayib, 1990, Applied Statistical Methods, Dar Al-Qalam, For Publishing and Distribution, Kuwait.
2. Draper, N. and H. Smith, 1981, Applied Regression Analysis, 2nd Edition, J. Wiley and Sons, Inc., USA.
3. Haghighi, A.M., Lian, Jian-ao, and D.P. Mishev, 2011, Advanced Mathematics for Engineers, with Applications in Stochastic Processes, Revised Edition, Nova Science Publishing, Inc. New York.
4. Kirk, R.E., 1995, Experimental Design: Procedures for the behavioral Sciences, 3rd, Brooks/Cole, Pacific Grove, CA, USA.
5. Larose, D.T., 2011, Discovering the Fundamentals of Statistics, 2nd Edition, Freeman, New York.
6. Montgomery, D.C., 2001, Design and Analysis of Experiments, 5th Edition, Wiley, New York
7. Sullivan, Michael, III, 2011, Fundamentals of Statistics, 3rd Edition, Prentice Hall, New York.
8. Sullivan, Michael, III, 2013, Statistics, Informed Decisions Using data, 4th Edition, Pearson, New York.
9. Wackerly, D.D., Mendenhall, W.III, and R.L. Scheaffer, 2008, Mathematical Statistics, with applications, 7th edition, Brooks/Cole Cengage Learning, USA.
10. Walpole, R.E., and R.H. Myers, 1989, Probability and Statistics for Engineers and Scientists, 4th Edition, MacMillan Publishing Company, New York.