Elementary Statistics

Satya Mandal

U. Kansas

Arrowtic K-Theory

Fall 2025

Contents

Chapter 1

The Language and Terminology

1.1 Introduction

Most people understand that statistics is the study of the numerical features of a subject/population. Understanding of the statisticians would not be much different. A statistician would only emphasize on how they do it, in addition. Some statisticians would define statistics as the scientific and mathematical study of the methods of collecting data, summarizing and presenting data, and drawing inferences from data.

American Statistical Association defines statistics follows (https://www.amstat.org/asa/what-is-statistics.aspx): *Statistics is the scientific application of mathematical principles to the collection, analysis, and presentation of numerical data. Statisticians contribute to scientific enquiry by applying their mathematical and statistical knowledge to the design of surveys and experiments; the collection, processing, and analysis of data; and the interpretation of the results.*

The goal of this course is to learn some commonly used methods to use collected sample data to draw inferences about a population and the mathematical basis behind such methods. The following is an example.

Example. Suppose you want to estimate the mean (average) weight of the fish polulation in the nearest lake. The mean weight is a charecteristic of the whole fish population in the lake. To estimate it, you catch a small sample of fish from the lake. Then compute the mean weight of the sample, to be called the sample mean. Then, declair that this sample mean is an estimate of the mean weight of the whole population (also called the population mean).

Another point about the nature of statistics as a science is that it is **not a deterministic science**. It does not have laws like force is equal to mass times acceleration. Statements in statistics come with a probability (i.e., quantified chance) of being correct. When a weatherman says that it will rain today he means that there is, say, a ninety five percent chance that it will rain today. Roughly, this means that if he makes the same

prediction one hundred times he will be correct 95 times, and it will not rain the other 5 days. The problem is that sometimes a weatherman will hide the information that there is a 95 percent chance only. Such information hiding is sometimes done for simplicity.

Skepticism: Skepticism about statistics is widespread and often justifiably so. It may not be an overstatement to say that statistics is misused and abused on regular basis. To put it sarcastically, abuse of statistics to generate opinion may already be a brunch of science or sociology based on scientific theory and models. The part that is based on scientific models may sometimes be ethically wrong, its scientific validity cannot be denied. Unfortunately, such methods include misleading the public with false data and misinformation.

On Sunday talk shows pundits and the political opinion makers try to justify opposing point of views, sometimes based on data from respectable sources. It would be a fair question, how could something be a science when it justifies two opposing point of views? While there is no cure for misleading or incorrect information, sometimes both may be statistitically correct with emphasis on the different aspects of the statistical inferences. Following is an example.

Example: In December 1998, the House of Representatives impeached President Bill Clinton. In February 1999, President was acquitted by the Senate. (In impeachment trial, the house works like the prosecutor and the Senate works like the jury. Search internet for more information).

President Clinton was formally charged with perjury and obstruction of justice. In any case, both stemmed out of allegations of sexual liaison and harassment. During this process of impeachment, there was long political discourse with respect to morality and legalities of the whole episode. Following would be typical discussion on TV.

1. Clinton critics would cite data and point out that, according to statistics, the majority of Americans think that character matters.

2. Clinton sympathizers would cite data and point out, according to statistics, that the majority of Americans think the president is doing a good job.

The implication here is that one of them was "wrong." But the science of statistics says that both were correct. Data was collected and analyzed, and it was found that the majority of Americans think that character matters and that the majority of Americans think the president is doing a good job. It does not matter to the science of statistics which one of the statistically established facts one would have desired. ■

Historically, during the early part of development of statistics, skepticism used to be of different nature. The validity of scientific foundation itself was in question. Statistics was compared with astrology, because both do predictions regarding unknown. An anecdote follows. When the proposal to establish the Indian Statistical Institute in Calcutta was considered by the government of India in the early part of the last century, some critics said, then why not an institute in astrology?

At the inception of statistics as a science there was a lot of skepticism about its scientific validity. Those days are gone, and statistics is not likened to astrology any more! Statistics is a well-founded and a **precise science**. It is a nondeterministic science in nature; it is made precise by making probabilistic statements only.

Descriptive and Inferential Statistics: In this course we will be talking about two branches of statistics. The first one is called descriptive statistics which deals with methods of processing, summarizing, and presenting data. The other part deals with the scientific methods of drawing inferences and forecasting from the data, and is called inferential or inductive statistics.

Course Organization: The course has nine lessons that can be divided into three parts:

1. Chapter 1 and 2: Descriptive Statistics. TI-84 (Silver Edition) would be used to solve problems.
2. Chapter 3, 4, 5, 6: Probability and Mathematical Basis. There is no direct TI-84 method for these lessons. However, after explaining the mathematics involved, the DISTR key (menu) of TI-84 (Silver Edition) will be used to compute probability.
3. Chapter 7, 8, 9: Inferential Statistics or Estimation. The goal of this course is to develop methods to do estimation, which would be accomplished in these lessons. Again, DISTR key (menu) of TI-84 (Silver Edition) will be used heavily.

In the rest of this lesson and the next we deal with descriptive statistics, which include the presentation of data in the form of tables, graphs, and computations of various averages of data.

1.2 Basic Definitions and Concepts

In statistics, we use a small "sample" to make inferences about a "big population". Statistics serves a purpose only when we do not have a way to find full or accurate information about the whole population. Sometimes, the population is such that it is intrinsically impossible to find full and accurate information. The same may be the situation because of the cost associated with full enumeration. The following are some examples:

1. The mean weight of the fish population in the nearest lake. Realistically, it would be impossible to catch all the fish in the lake, measure them and find the mean weight.
2. You are a quality control inspector in a lamp factory. To give an idea to the consumers, you want to know the mean lifetime (in hours) of the lamps produced. There is no way you can measure the mean before you sell.
3. The mean annual expenditure (in year 2011) of the KU student population.

4. **Remark.** In some cases, in spite of existence of full information regarding the whole population, for cost effectiveness, samples are used to estimate the population. Suppose you want to know the mean GPA of the KU population. Although, KU has the full information, you may not be able to access the full data and then do the computations due to the associated cost. So, you may like to be content with a sample and the sample mean GPA as an estimate.

5. **Remark.** Interestingly, advent of computers in abundance caused certain usages of statistics obsolete. Thirty years ago, KU used to keep record in papers. Those days, they would have used sample data to avoid dealing with the huge amount of data in papers.

1.2.1 Population and Sample

Definition 1.2.1. A complete collection of data on the group under study is called the **population** or the **universe**. A member of the population is called a **sampling unit**. Therefore, the population consists of all its sampling units. A **Sample** is a collection of sampling units selected from the population.

Most often, we will work with numerical characteristics (like height, weight, and salary) of a group. So usually the population is a large collection of numbers and the sample is a small subset of the population.

Example. Suppose we are studying the daily rainfall in Lawrence. Since daily rainfall could be from 0 inches to anything above 0, the population here is all nonnegative numbers (i.e., the interval $[0, \infty)$). A sample from this population would be the observed amount of daily rainfall in Lawrence on some number of days. A sample of size 11 would be the observed daily rainfall in Lawrence on 11 days.

Definition 1.2.2. A **variable** is something that varies or changes value. Most often, we consider numerical variables. Numerical variables are also called **quantitative variables**. Examples of quantitative variables include height, length, weight, number of typos in books, number of credit hours completed by students, number of accidents (or number of anything) and time. Non-numerical variables are also considered. They are called **qualitative variables**. Examples of qualitative variables include blood group and gender. In fact, any genetic property (genotype or phenotype) is a qualitative variable, because they vary from human to human (or trees to trees).

In Chapter 4, we will have an elaborate discussion on a specific type of variables called **random variables**, which would be more relevant for our purpose.

1.2.2 Parameters and Statistics

Definition 1.2.3. 1. Given a set of data, any numerical value computed from the data using a formula or a rule is called a quantitative measure of the data.

2. A quantitative measure of a population data is called a **parameter**. In other words, parameters belong to the whole population and are computed (if feasible) from the WHOLE population data. Examples: the average GPA of all KU students, the height of the tallest student in KU, the average income of the entire KU student population.

One way to study a population is to know some of the parameters of the population. Unfortunately, computing such parameters could be expensive or even impossible. In practice, parameters are unknown and the main game of statistics is to try to estimate parameters on the basis of small samples collected from the population.

Definition 1.2.4. A quantitative measure of a sample data is called a **statistic**. Any constant that we compute from a sample is a statistic. We use these statistics to estimate the parameters of the population. For example, the average height computed from a sample is a reasonable estimate for the (parameter) average height of the KU student population. Obviously, we do not expect the value of the statistic to be exactly equal to the parameter value. Hopefully, the error will be small or will exceed our tolerable limit very rarely (say once in a 100 trials).

Why do we need a statistic?

Sometimes it will be impossible to know the actual value of a parameter. For example, let μ be the mean length of the life of light bulbs produced by a company. In this case, the company cannot test all the bulbs it produces to find a mean length. So, the best that that we can do is to we test a few bulbs (the sample), compute the sample mean length $\overline{x}$ (a statistic) of the life of these bulbs and use $\overline{x}$ as an estimate for the mean length (parameter μ) of the life for all the bulbs it produces.

1.2.3 Frequency Distribution

In this section we talk about representation of data organized in tabular form. Such a representation is called a frequency distribution. We are mostly concerned with numerical data (i.e., quantitative data), but also consider some non-numerical data (i.e., qualitative data).

Example. The following is data on the blood group of 72 patients in a hospital:

O	A	O	B	O	A	O	A	A	O	A	O	B	A	O	A	A	O
O	A	O	A	B	B	O	O	AB	O	A	O	O	A	O	A	O	A
A	A	A	O	A	A	O	AB	B	O	A	O	B	A	A	A	O	O
O	A	A	B	A	O	O	O	O	A	A	B	O	O	O	A	A	A

We have four types of blood groups, namely, O, A, B, AB. Each of these blood groups may be referred to as a "**class**." The **frequency** of a class is defined as the number of data members that belong to that class. For example, the frequency of the class O is 31; the frequency of class A is 31. A table that lists the classes and the corresponding frequency is called the frequency distribution of this qualitative data. Following is the frequency distribution of this data:

Bloodgroup	**frequence**
O	31
A	31
B	8
AB	2
Total	72

1.2.4 Ungrouped Data

For the quantitative data, we consider two types of frequency table. When we are working with a large set of data we group the data set into a few classes and construct a "frequency table," which we will discuss later.

If the data set contains only a few distinct values then data would not be grouped. We make a list of all the data-values present and give the corresponding frequency for each data-value in a table. The number of times a data-value appears in the data set is called the **frequency** of the data member. A list that presents the data members and the corresponding frequency in a tabular form is called a **frequency table** or **frequency distribution**. The **relative frequency** and **percentage frequency** of a data member x are defined as follows:

$$\begin{cases} relative\ frequency\ of\ x = \frac{frequency\ of\ x}{Data\ size} \\ percentage\ frequency\ of\ x = \frac{(frequency\ of\ x)100}{Data\ size} \end{cases}$$

Example 1.1.1 To estimate the mean time taken to complete a three-mile drive by a race car, the race car did several time trials, and the following sample of times taken (in seconds) to complete the laps was collected:

50	48	49	46	54	53	52	51	47	56	52	51
51	53	50	49	48	54	53	51	52	54	54	53
55	48	51	50	52	49	51	53	55	54	50	

Note that there are 35 observations here. So we say that the size of the sample (or data) is 35. Also the values present are 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56. Since there

are only 11 distinct values present we can make a frequency table for the ungrouped data. The following is the frequency distribution of this ungrouped data:

Time (*in sec*)	**frequency**	**relative frequency**	**percentage frequency**
46	1	1/35	2.86
47	1	1/35	2.86
48	3	3/35	8.57
49	3	3/35	8.57
50	4	4/35	11.43
51	6	6/35	17.14
52	4	4/35	11.43
53	5	5/35	14.29
54	5	5/35	14.29
55	2	2/35	5.71
56	1	1/35	2.86
Total	35	1	100

(1.1)

1.2.5 Grouped Data

When we are working with a large set of data that contains too many distinct values, then we group the whole set of data into a few class intervals and give the corresponding "frequency" of the class. When the data is presented in this way, the data is called **grouped data**. The number of data members that fall in a class interval is called the **class frequency** and the **relative and percentage frequencies** are computed by the same formula as above. A list that gives various class intervals and the corresponding class frequencies in a tabular form is called a **class frequency table or class frequency distribution** of the data. The frequency distribution may also include the relative and percentage frequencies.

Grouped Data and Loss of Information: Note that when we construct a frequency table of ungrouped data, there is no loss of information. The original data can be reconstructed from the frequency table of ungrouped data. Only loss would be the order in which the original data appeared.

Sometimes it is necessary to group data into class intervals to construct a frequency distribution. This would be the case when there are too many distinct data-values present in the data-too many even to fit into a table in a regular size paper for presentation. In such situations, we group the data in a few class intervals. While class frequency distribution is very good for presentation and may be convenient for other reasons, we lose a lot of information in this process. There would be no way to recover the original data from the class frequency distribution.

Steps to Construct Frequency Distribution: To construct a class frequency table of data-set, the first question would be, how many class intervals should we have? The answer is that it should not be too few nor should it be too many. The fewer the number

of class intervals, more is the loss of information. In the extreme case, if we use only one class interval, all the information would be lost. On the other hand, if we take too many, we will have the problem of having to work with ungrouped data. (In this course we will always tell you how many classes to take.) Although sometimes it may be necessary to take class intervals of varying width, in this course we only consider classes of equal class width. We follow the following steps to construct a class frequency distribution.

1. **Range:** Pick a suitable number L less than or equal to the smallest value present in the data. Pick a suitable number H greater than or equal to the highest value present in the data. The range R that we consider is $R = H - L$.

2. **Number of Classes**: Decide on a suitable number of classes. (In this course we will tell you the number of classes.)

3. Class Width: We have

$$\textbf{class width} = w = \frac{R}{Number\ of\ classes}$$

   We will pick L, H, and the number of classes so that class width is a "round number." Classes: We divide our interval [L,H] into subintervals, to be called classes, as

$$[L, L+w], \quad [L+w, L+2w], \quad [L+2w, L+3w], \quad \cdots, \quad [H-w, H]$$

4. **Frequency:** Find the frequency for each of the classes. You can use an advanced calculator or some software (like Excel) to count frequencies.

   An ambiguous situation may arise, when a data-value falls on a class boundary. Depending on the nature of data or otherwise, we would have to follow a consistent convention whether to count such data members on the left or the right class interval.

A few more important definitions. The above intervals are called **class intervals**. The w above is called the **class width**. The lower end of the class is called lower limit and the upper end of the class is called upper limit. The **class mark** is the midpoint of the class, defined as follows:

$$\textbf{class mark} = \frac{(lower\ limit\ of\ class) + (upper\ limit\ of\ class)}{2}$$

Example 1.2.5. The following is the weight (in ounces), at birth, of a certain number of

babies.

74	105	124	110	119	137	96	110	120	115	140
65	135	123	129	72	121	117	96	107	80	91
74	123	124	124	134	78	138	106	130	97	145
93	133	128	96	126	124	125	127	62	127	92
95	118	126	94	127	121	117	124	93	135	156
143	125	120	147	138	72	119	89	81	113	91
133	127	138	122	110	113	100	115	110	135	141
97	127	120	110	107	111	126	132	120	108	148
143	103	92	124	150	86	121	98	74	85	99

We construct a class frequency table of this data by dividing the whole range of data into class intervals.

Solution: Note that the lowest value is 62 and the highest value is 156. We take $L = 60, H = 160$, so $R = H - W = 100$. We made such a choice of L and H, precisely so that $R = 100$ is a "nice" number. Now we decide to have 5 class intervals and so $w = R/5 = 20$. According to what I said above, our classes should be :

$$[60, 80], \quad [80, 100], \quad [100, 120], \quad [120, 140], \quad [140, 160]$$

But if we do so then there is a risk that some data members (like 80, 100, 120, 140) will fall in two classes. To avoid this we add .5 to all the class boundaries. So, our classes are

$$[60.5, 80.5], \quad [80.5, 100.5], \quad [100.5, 120.5], \quad [120.5, 140.5], \quad [140.5, 160.5].$$

So the frequency distribution is as follows:

Classes	**frequency**	**relative frequency**	**percentage frequency**
$60.5 - 80.5$	9	9/99	9.09
$80.5 - 100.5$	20	20/99	20.20
$100.5 - 120.5$	25	25/99	25.26
$120.5 - 140.5$	37	37/99	37.38
$140.5 - 160.5$	8	8/99	8.08
Total	99	1	100

(1.2)

1.2.6 Use of Calculators

We would avoid hand computations. We will be using calculators (TI-84) in this course.

1. **Entering data in TI-84:** let me explain how you enter data in the TI-84.

(a) Press the button "stat."

(b) Select "Edit" in the Edit menu and enter.

(c) You will find a lists named L1, L2, L3, L4, L5, L6.

(d) Let's say you want to enter your data in L1. If L1 has some data, you clear it by pressing the stat button and selecting ClrList in the Edit menu. ClrList appears then type L1 and hit enter. To type "L1" on your TI-84 simply press 2nd then 1.

(e) Once L1 is cleared, you select Edit in the Edit menu and enter.

(f) Now type in your data; enter one by one.

2. **Sorting and counting frequency:** It is not easy to construct a frequency table of a data set unless you are systematic. Traditionally, we used "tally marks" to count the frequency. Now you can use some software programs (e.g., Excel). Let me show you a method, using a calculator (TI-84).

(a) Press "stat."

(b) To input data, enter "edit."

(c) Enter your data (say in L1).

(d) Press "stat."

(e) Enter "sortA" L1.

(f) Press "stat" and then enter "edit." On L1 you will see that the data is sorted in an increasing order.

(g) Now you can count the frequencies.

1.2.7 Problems on Frequency Distribution

Exercise 1.2.6. Repeat example (1.2.5) with class width 5.

Exercise 1.2.7. The following is the weight (in ounces), at birth, of 96 babies born in Lawrence Memorial Hospital in May 2000.

$$
\begin{array}{cccccccccccc}
94 & 105 & 124 & 110 & 119 & 137 & 96 & 110 & 120 & 115 & 119 & \\
104 & 135 & 123 & 129 & 72 & 121 & 117 & 96 & 107 & 80 & 80 & \\
96 & 123 & 124 & 124 & 134 & 78 & 138 & 106 & 130 & 97 & 134 & \\
111 & 133 & 128 & 96 & 126 & 124 & 125 & 127 & 62 & 127 & 96 & \\
116 & 118 & 126 & 94 & 127 & 121 & 117 & 124 & 93 & 135 & 112 & (1.3) \\
120 & 125 & 120 & 147 & 138 & 72 & 119 & 89 & 81 & 113 & 100 & \\
109 & 127 & 138 & 122 & 110 & 113 & 100 & 115 & 110 & 135 & 120 & \\
97 & 127 & 120 & 110 & 107 & 111 & 126 & 132 & 120 & 108 & 148 & \\
133 & 103 & 92 & 124 & 150 & 86 & 121 & 98 & & & &
\end{array}
$$

Construct a class frequency table of this data by dividing the the whole range of data into class intervals:

$[60.5-70.5], [70.5-80.5], [80.5-90.5], [90.5-100.5], [100.5-110.5], [110.5-120.5], [120.5-130.5], [130.5-140.5], [140.5-150.5]$

Exercise 1.2.8. The following are the length (in inches), at birth, of 96 babies born in Lawrence Memorial Hospital in May 2000.

18	18.5	19	18.5	19	21	18	19	20	20.5
19	19	21.5	19.5	20	17	20	20	19	20.5
18	18.5	20	19.5	20.75	20	21	18	20.5	20
21	19	20.5	19	20	19.5	17.75	20	19.5	20
20.5	17	21	18.5	20	20	20	18.5	19.5	19
18	20.5	18	20	19	19	19.5	20	20.75	21
17.75	19	18	19	20	18.5	20	19	21	19
19.5	20	20	19	19.5	20	19.5	18.5	20.5	19.5
20.25	20	19.5	19.5	20	20	20	21	20	19
18.5	20.5	21.5	18	19.5	18				

Construct a frequency table for this data by dividing the whole range into class intervals:

$$[16-17], [17-18], [18-19], [19-20], [20-21], [21-22].$$

Note: If a data member falls on the boundary, count it in the right/upper class-interval.

Exercise 1.2.9. The following data represents the number of typos in a sample of 30 books published by some publisher.

156	159	162	160	156	162
159	160	156	156	160	162
156	159	162	156	162	158
160	158	159	162	158	158
162	160	159	162	162	160

Construct a frequency table (by sorting in your calculator).

Exercise 1.2.10. Following is data on the hourly wages (paid only in whole dollars) of 99

employees in an industry.

7	11	7	11	10	9	10	10	12	13
7	8	11	11	14	9	7	9	11	7
9	13	12	14	7	8	7	14	15	9
9	7	11	9	12	9	12	11	14	9
12	13	7	9	10	14	11	12	13	7
15	15	16	16	15	16	11	7	18	19
15	16	15	15	16	16	17	16	16	13
15	15	16	15	16	15	15	17	16	12
16	15	15	16	15	15	19	8	16	17
16	16	15	16	16	16	13	12	8	

Construct a frequency table (by sorting in your calculator).

Exercise 1.2.11. Following is data on the hourly wages (paid only in whole dollars) in an industry.

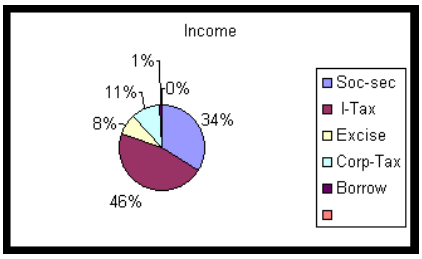
9	11	8	9	10	11	7	10	12	13
7	11	8	11	14	9	10	9	11	7
13	13	14	12	9	8	12	14	15	9
9	7	12	7	12	7	7	11	13	9
11	9	9	9	10	14	11	12	14	7

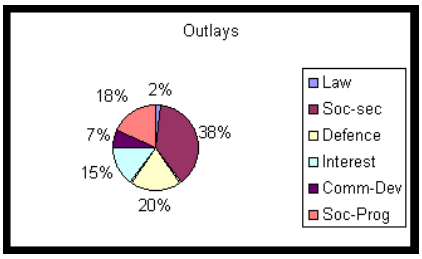
Construct a frequency table (by sorting in your calculator).

1.3 Pictorial Representation of Data

Another way to represent data is to use pictures and graphs. Such pictorial representations are commonly used in newspapers and other media outlets. Pictorial representation is particularly helpful when you have to represent data to people with limited technical background, like newspaper readers or a governmental or congressional body.

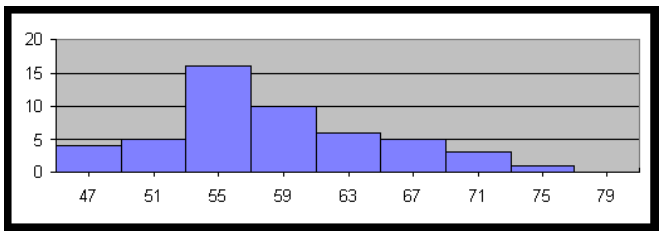
The Pie Chart: The pie chart is a commonly used pictorial representation of data. When you do your tax return every year, you find a few pie charts in the instruction book for form 1040. These charts show what proportion/percentage of each tax dollar goes for particular expenses. I reproduced the following pie charts from the 1040 instruction book of 1999.





The Histogram Among pictorial representations, the most useful in this course is the **histogram**. The histogram of data is the graphical representation of the frequency distribution of the data, where we plot the variable on the horizontal axis and above each class interval, we erect a bar of the height equal to the frequency of the class. Such a histogram is called a frequency histogram.

If, instead, we erect bars of height equal to the relative frequency, then the graph is called a relative frequency histogram. Similarly, we can construct a percentage frequency histogram. The following is a histogram.



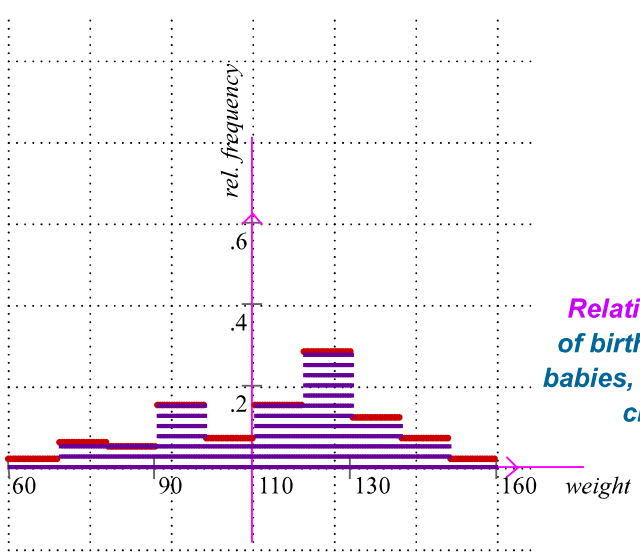
We would avoid unequal class lengths, which makes our discussion of the histogram fairly simple.

1.3.1 More Histograms

In nature, shape of the histograms of almost any kind of numerical data would resemble a bell-shaped pattern. A perfect such bell shape is shown in the following diagram:

Satya Mandal

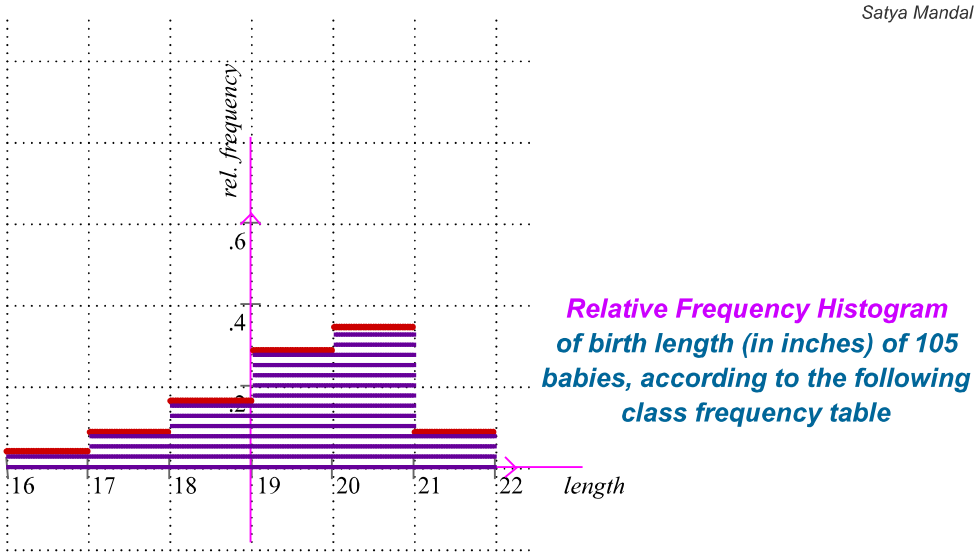
A Perfect Bell Shape Curve



Relative Frequency Histogram of birth weight (in ounces) of 99 babies, according to the following class frequency table

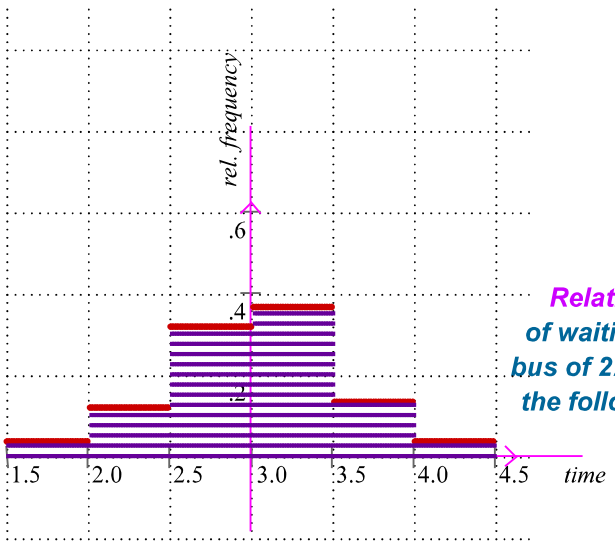
Class	*60-70*	*70 -80*	*80-90*	*90-100*	*100-110*	*110-120*	*120-130*	*130-140*	*140-150*	*150-160*
Frequency	*2*	*6*	*5*	*15*	*7*	*15*	*28*	*12*	*7*	*2*
Rel. Frequencey	*2 /99*	*6 /99*	*5/99*	*15/99*	*7/99*	*15/99*	*28/99*	*12/99*	*7/99*	*2/99*

You can imagine that the histogram roughly fits under a bell shape curve. It may or may not be a great fit, depending upon your expectation.



Class	16-17	17-18	18-19	19-20	20-21	21-22
Frequency	4	9	17	30	36	9
Rel. Frequencey	4 /105	9/105	17/105	30/105	36/105	9/105

You can imagine that the histogram roughly fits under a bell shape curve.



Class	1.5 - 2.0	2.0 - 2.5	2.5 - 3.0	3.0 - 3.5	3.5-4.0	4.0-4.5
Frequency	8	27	72	83	30	8
Rel. Frequencey	8 /227	27/227	72 /227	83//227	30//227	8//227

You can imagine that the histogram roughly fits under a bell shape curve. This one fits better. Normally, it works better when data size is large.

1.3.2 The Cumulative Frequency Distributions and Ogive

We start with this example **Example 1.1.3.** Following is the frequency table of data on height (in inches) of some babies at birth. Sketch the histogram of the following data:

Height	Frequency
$16-17$	3
$17-18$	8
$18-19$	34
$19-20$	60
$20-21$	72
$21-22$	18

For a given value x of a variable, the **cumulative frequency** of the data, for x, is the number of data members that are less than or equal to x.

Definition 1.3.1. Given a frequency distribution of some data, for a class boundary x, the cumulative frequency is the sum of all the class frequencies less or equal to x. The cumulative frequency distribution is a table that gives the cumulative frequencies against some x values (for us the class boundaries). We also define cumulative relative frequency and cumulative percentage frequency as follows:

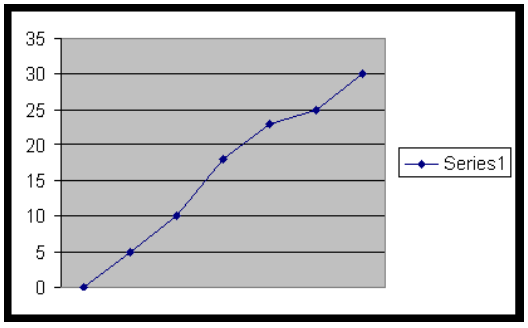
$$\begin{cases} \text{cumulative relative frequency of data , at x} = \frac{cumulative\ \ frequency\ \ of\ \ x}{Data\ \ size} \\ \text{cumulative percent frequency of data , at x} = \frac{(cumulative\ \ frequency\ \ of\ \ x)100}{Data\ \ size} \end{cases}$$

Example 1.1.4 Once again we consider the data on birth weight of babies in Example 1.1.2 that we discussed in the last section. A cumulative frequency distribution can be constructed from the frequency distribution.
Solution: We have seen the frequency distribution before. The following is the cumulative distributions:

Weight	Cumulative Frequency	Cumulative Relative − Cumulative	Cumulative PercentageFrequency
60.5	0	0	0
80.5	9	9/99	9.09
100.5	29	29/100	29.29
120.5	54	54/99	54.55
140.5	91	91/99	91.92
160.5	99	1	100

Definition. The ogive is a line graph, where we plot the variable on the horizontal axis and the cumulative frequency on the vertical axis. If we plot the cumulative relative

frequency on the vertical axis, then the line graph is called the relative frequency ogive.



Chapter 2

Measures of Central Tendency and of Dispersion

2.1 Measure of Central Tendency: Mean

In this lesson we define various numerical measures (constants) for data sets. These numerical measures summarize and describe the data. The average value of the data would be a common example. There are two broad classification such numerical measures that are computed from the data:

1. **measures of central tendency** and
2. **measures of dispersion.**

A measure of central tendency represents an "average value." Mean, median, mode (if you already know these) are measures of central tendency. A measure of dispersion is a measure of how widely the data is scattered around. In the section we discuss (1). The most common measure of central tendencies is the mean or arithmetic mean.

Definition. Given a data set, by Data size, we mean number of sample units present. The **mean** $\overline{x}$ or the arithmetic mean of a set of data x is given by

$$mean = \overline{x} = \frac{sum\ of\ all\ the\ data\ values}{Data\ size}$$

The data size is usually denoted by n. The data variable is usually denoted by x or $x_1, x_2, \ldots, x_n$. So, we can write

$$\overline{x} = \frac{\sum x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

where the symbol $\sum$ (Sigma) denotes sum, and $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$.

Weighted Mean: In the formula (2.1), all the values x_i have equal weight. Sometimes, different values in data carry different weight. Let us consider the following data and the corresponding frequency distribution that we computed earlier:

2.1.1 To estimate the mean time taken to complete a three-mile drive by a race car, the race car did several time trials. The following are sample times taken (in seconds) to complete the laps (same as in (1.1)) :

50	48	49	46	54	53	52	51	47	56	52	51
51	53	50	49	48	54	53	51	52	54	54	53
55	48	51	50	52	49	51	53	55	54	50	

The following is the frequency distribution of this data, we computed above:

Time x_i	46	47	48	49	50	51	52	53	54	55	55
Frequency f_i	1	1	3	3	4	6	4	5	5	2	1

(*Observe, unlike (1.1), we represented the same frequency table, horizontally.*) To compute the mean time of the original data (1.1), we obviously, add all the data values and divide by the data size 35. The frequency distribution tells us that, in the data, 46 was present 1 time, 47 was present 1 time, 48 was present 3, times and so on. So, using the frequency distribution, we compute the mean $\overline{x}$ as follows:

$$\overline{x} = \frac{(46 \cdot 1 + 47 \cdot 1 + 48 \cdot 3 + 49 \cdot 3 + 50 \cdot 4 + 51 \cdot 6 + 52 \cdot 4 + 53 \cdot 5 + 54 \cdot 5 + 55 \cdot 2 + 56 \cdot 1)}{(1 + 1 + 3 + 3 + 4 + 6 + 4 + 5 + 5 + 2 + 1)}$$

$$= \frac{1799}{35} = 51.4$$

The mean of the original data is the "**weighted mean**" of the data values 46, 47, 48, 49, 50, 51, 52, 53, 54, 55 and 56 with the corresponding frequencies as the **weight**. Therefore, when we compute the mean using the frequency table, the formula for the mean would be

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i} \quad \text{where } f_i \text{ is the weight of } x_i$$

The weighted mean is defined in more general context as follows:

Definition. If $x_1, x_2, \ldots, x_n$ in a data set have different weights and the values x_i has weight w_i, then the **weighted** is defined as

$$\textit{weighted mean} = \overline{x} = \frac{\sum w_i x_i}{\sum w_i}$$

Properties of the Mean: Following are some of the obvious properties of means:

1. **Combining two means:** Suppose we have two sets of data.
   The mean of the first set is $\overline{x}$, and the size of the first set is m;
   the mean of the second set is $\overline{y}$, and size of the second set is n.
   Then, the mean of the combined data is

$$\textbf{Combined mean} = \frac{m\overline{x} + n\overline{y}}{m + n}$$

2. **Effect of translation:** Let $\overline{x}$ be the mean of $x_1, x_2, \ldots, x_n$. Then the mean of $y_1 = x_1 + d, y_2 = x_2 + d, \ldots, y_n = x_n + d$ is given by

$$\overline{y} = \overline{x} + d$$

3. **Effect of multiplication by a constant:** Let $\overline{x}$ be the mean of $x_1, x_2, \ldots, x_n$. Then the mean of

$$z_1 = cx_1, z_2 = cx_2, \ldots, z_n = cx_n \quad \text{is given by} \quad \overline{z} = c\overline{x}$$

Example. (**effect of translation**): Your teacher tells you that the mean score for the midterm in your class is 73. After you complained and requested a change, he agreed that all can add 7 points to their score. The new mean score is (old mean + 7) = 73 + 7 = 80. This is what we meant by "**effect of translation**."

Example. (**effect of multiplication by c**): Suppose you have some data $x_1, x_2, \ldots, x_n$ on salaries in an industry in the United States and the mean is \$37000. On a certain day (January 29, 2021), 1 U.S. dollar = 1.28 Canadian dollars (say $c = 1.28$). So, in Canadian dollars the mean is $37000 * c = 37000 * 1.28 = 47360$ Canadian Dollars.

Similarly, any change of units (inches to feet or cm, minutes to seconds) are "multiplication by a constant c." (Like inches to centimeters, pounds to kilograms.)

Example 2.1.2. (**GPA is an exampled of weighted mean**). A student took PHSX 115 (College Physics), PSYC 120 (Personality), FREN 110 (Elementary French), BUS 241 (Managerial Accounting), and MATH 365 (Elementary Statistics). The number of credit hours and the student's grade is given in the following table:

Course	*PHSX* 115	*PSYC*120	*FREN* 110	*BUS* 241	*MATH* 365
Grade(Points)	*B*(3)	*A*(4)	*B*(3)	*C*(2)	*B*(3)
Credit Hours	4	3	5	3	3

What is the student's GPA?
Solution. The GPA is the weighted average of the points (corresponding to the grades), weight being the course-credit hours. So, the

$$GPA = \frac{3*4+4*3+3*5+2*3+3*3}{4+3+5+3+3} = \frac{354}{18} = 3.$$

2.1.1 Other Measures of Central Tendency: Median, and Mode

Definition. The **median represents** the middle value of the data. Half the data will be less than or equal to the median, and half the data will be greater than or equal to the median. You are above the median American income if half the American population is making less than you make. Suppose the data is arranged in an increasing order (i.e., in an array). Then,

1. If the size of data is ODD then the median is the middle value.
2. If it is EVEN, then the median is the mean of the middle two values.

Definition. **The Percentiles**: For a number p with $0 \leq p \leq 100$, the p^{th} percentile x_p of the data is a number such that at least p percent of the data members are below x_p and at least $(100 - p)$ percent of the data members are above x_p. Further,

1. The 25th percentile is called the **first quartile** Q_1.
2. The median is the 50^{th} percentile, also called the **second quartile** Q_2.
3. The 75^{th} percentile is called the **third quartile** Q_3.

Definition. (**The Mode**)
The MODE of the data is the value or values that have the highest frequency. For example, the mode of the set $\{1, 3, 5, 5, 7\}$ is because it has the highest frequency. The mode of $\{1, 1, 3, 5, 5, 7\}$ is $\{1, 5\}$ because 1 and 5 both have the highest frequency. Such a set is said to be bimodal.

Using $TI - 84$: We already mentioned how to enter data.

1. **Sorting data and computing the median:**
   (a) Enter your data in a list, say $L1$.
   (b) Select SortA in the Edit menu and enter.
   (c) The calculator will ask for the list.
   Type in the list ($L1$), close the parentheses, and enter.
   (d) The calculator will say Done.
   (e) Press stat, select edit in the Edit menu, and enter.
   (f) You will see that your data in $L1$ has been sorted in an increasing order.
   (g) If the data size is odd, the median is the middle value.
   (h) If the data size is even, the median is the average of the middle two values.
2. **Computing the mean if only raw data is given:**
   (a) Enter your data in a list, say $L1$.
   (b) Select "1-Var Stats" in the CALC menu and enter.
   (c) The calculator will ask for the list. Type in the list $L1$ and enter.
   (d) The calculator will give a list of numbers; $\overline{x}$ is the mean .
3. **Computing the mean if the frequency table is given:**

(a) Enter the frequency table in the calculator, say, x-values in $L1$ and frequencies in $L2$.

(b) Select "1-Var Stats" in the CALC menu and enter.

(c) The calculator will ask for the lists. Type in the list $L1, L2$ and enter.

(d) The calculator will give a list of numbers; $\overline{x}$ is the mean.

4. **Computing the median:** Do the same as above and scroll down.

2.1.2 Problems on Mean and Median

Exercise 2.1.1. The following is the price (in dollars) of a stock (say, APPL) checked by a trader several times on a particular day.

$$138 \quad 142 \quad 127 \quad 137 \quad 148 \quad 130 \quad 142 \quad 133$$

Find the median price and mean price observed by the trader. **Solution:** Use TI-84.

Exercise 2.1.2. The following figures refer to the GPA of six students.

$$3.0 \quad 3.3 \quad 3.1 \quad 3.0 \quad 3.1 \quad 3.1$$

Find the median and mean GPA.

Exercise 2.1.3. The following data give the lifetime (in days) of light bulbs.

$$138 \quad 952 \quad 980 \quad 967 \quad 992 \quad 197 \quad 215 \quad 157$$

Find the mean and median lifetime of these bulbs. **Solution:** Use TI-84.

Exercise 2.1.4. An athlete ran an event 32 times. The following frequency table gives the time taken (in seconds) by the athlete to complete the events.

Time x_i $(in\ sec.)$	**frequency** f_i
26	3
27	6
28	5
29	6
30	9
31	3
Total	32

Compute the mean and median time taken by the athlete. **Solution:** Use TI-84. **Solution:** Use TI-84.

Exercise 2.1.5. Consider the data, as in (Ex. 1.2.7) on the weight (in ounces), at birth, of 96 babies born in Lawrence Memorial Hospital in May 2000. Compute the mean and median weight, at birth, of the babies. **Solution:** Use TI-84.

Exercise 2.1.6. Consider the data, as in (Ex. 1.2.10) on the hourly wages (paid only in whole dollars) of 99 employees in an industry. **Solution:** Use TI-84.

Exercise 2.1.7. Following is the frequency table on the number of typos in a sample of 30 books published by a publisher.

No. of Typos	156	158	159	160	162
Frequency	6	4	5	6	9

Find the mean and median number of typos in a book. **Solution:** Use TI-84.

Exercise 2.1.8. Consider the data, as in (Ex. 1.2.8), on the length (in inches), at birth, of 96 babies born in Lawrence Memorial Hospital in May 2000. **Solution:** Use TI-84.

2.2 Measures of Dispersion

The measures of central tendencies?mean, median, mode? represent the middle values of the data set. The variability of data would also be of our interest, for a better understanding of the distribution of data. Two data sets may have same mean and median, but they may be spread out differently. Following is an example.

Example 2.2.1. Suppose two sections of the statistics class have the following percentage score distribution at the end of the semester:

Section A	81	84	83	80	82
Section B	72	93	92	82	71

Both these sections have the same mean - 82. Medians of both the sets are same - 82. But the data sets are differently dispersed. In Section A, everybody will get a B grade. In section B, we will have two C, one B and two A.

The **measure of dispersion** is a measure of how widely the data is scattered around. In section A, the data has a very small dispersion or variability, whereas section B has a large dispersion. A very simple (crude) measure of dispersion is the range of the data as we have defined before:

$$range = (largest\ value) - (smallest\ value).$$

We will define three more measures of dispersions:
Mean Deviation, **Variance**, and **Standard Deviation**.

Definition: Suppose $x_1, x_2, \ldots, x_n$ is a set a of data.

1. Then, the **Mean Deviation** of the data is defined to be

$$\textbf{MeanDeviation} = \frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n}$$

   So, the mean deviation is the mean of the absolute deviations $|x_i - \overline{x}|$ from the mean.

2. The **the Variance** s^2 of the data is defined as follows:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} \tag{2.2}$$

   Observe:

   (a) Note that we denote the sample variance as the square of a number s (*and we can do so*).

   (b) Also note that we divide by $n-1$, not by n. For some statistical reason, dividing by $n-1$ works better. (*We justify this, in the remark, below* (7.1).)

   (c) We would like our measure of dispersion to have the same units as our data, but our formula involves squares $(x_i - \overline{x})^2$. Therefore, the unit of the variance, s^2, is the unit of the data squared. If the data is in feet, the variance is in square feet. To solve this problem we define another measure of dispersion, standard deviation denoted s.

3. The **Standard Deviation** is defined as the square root of the variance s^2. So, to compute the sample standard deviation, we have to compute the sample variance first.

If the data corresponds to a sample, then we call the above as **sample Mean Deviation**, **sample Variance**, and **sample Standard Deviation**. If the data corresponds to the population, then we call the above as **population Mean Deviation**, **population Variance**, and **population Standard Deviation**. For now, we work with samples only. (*Populations data is actually unknown, which we try to estimate or model.*)

If we simplify the definition of the variance we get the following formula:

$$s^2 = \frac{(x_1^2 + x_2^2 + \cdots + x_n^2) - n\overline{x}^2}{n-1}$$

We do some computations with the above example 2.2.1.

1. The

$$\begin{cases} \text{mean deviation for section A} = \frac{(1+2+1+2+0)}{5} = \frac{6}{5} \\ \text{mean deviation for section B} = \frac{(10+11+10+0+11)}{5} = \frac{42}{5} \end{cases}$$

   As expected, mean deviation of Section B is higher, because the variability of section B was clearly higher than that of A.

2. The variance s_A^2 of section A is:

$$s_A^2 = \frac{(81-82)^2 + (84-82)^2 + (83-82)^2 + (80-82)^2 + (82-82)^2}{5-1} = 2.5$$

The variance s_B^2 of section B is:

$$s_B^2 = \frac{(72-82)^2 + (93-82)^2 + (92-82)^2 + (82-82)^2 + (71-82)^2}{5-1} = \frac{442}{4}$$

For future discussions, for reasons mentioned above, the standard deviation s would be our choice of our (and for practicing statistician's) measure of dispersions.

2.2.1 Application of Standard deviation

Under normal circumstances, which we will discuss later, the mean and the standard deviation carries an enormous amount of information regarding the distribution of the data.

Chebyshev's Rule. This rule applies for all kinds of data. Suppose $\overline{x}$ is the mean and s is the standard deviation of a set of data $x_1, x_2, \ldots, x_n$. Then we have the following:

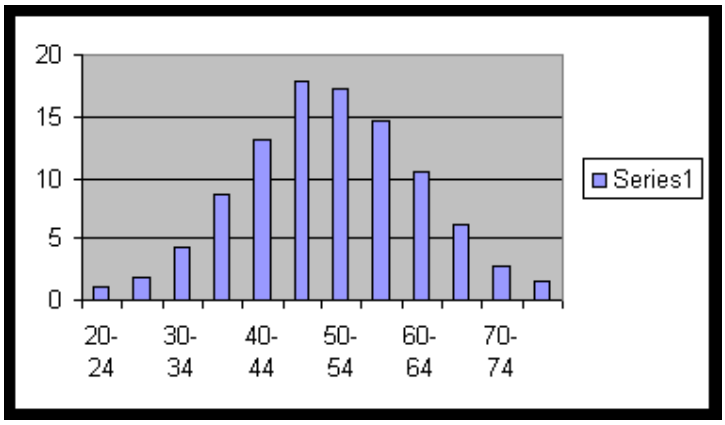
1. At least 0 percent of the observations will fall within 1 standard deviation of the mean, i.e, within $(\overline{x} - s, \overline{x} + s)$. This is clearly of not use.

2. At least **75 percent** of the observations will fall within 2 standard deviations of the mean, i.e., within $(\overline{x} - 2s, \overline{x} + 2s)$.

3. At least **89 percent** of the observations will fall within 3 standard deviations of the mean, i.e., within $(\overline{x} - 3s, \overline{x} + 3s)$.

4. More generally, at least $100\left(1 - \frac{1}{k^2}\right)$ percent of the data will be within k-standard deviations from the mean, i.e. within $(\overline{x} - ks, \overline{x} + ks)$.

Note, this applies to any set of data, without any hypotheses. In that respect, it is very strong. It is mathematically smart, but statistically crude. In fact, in nature, the histogram of data usually fits nicely, under a nice bell curve, as was shown above and below.


20
15
10
5
0
20-24
30-34
40-44
50-54
60-64
70-74
Series1

A Perfect Bell Shape Curve

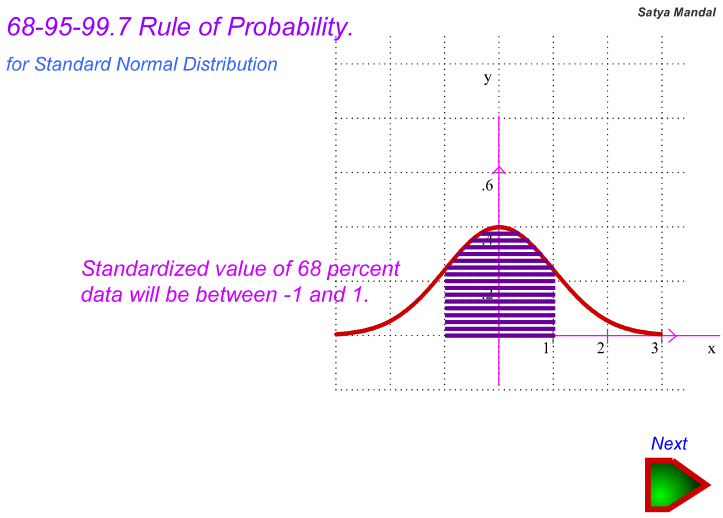
We incorporate this basic nature of data, in to our statistical modeling, and give a improved version of Chebyshev's Rule, which I would call **statistically smarter**. This is called the Empirical Rule, as follows.

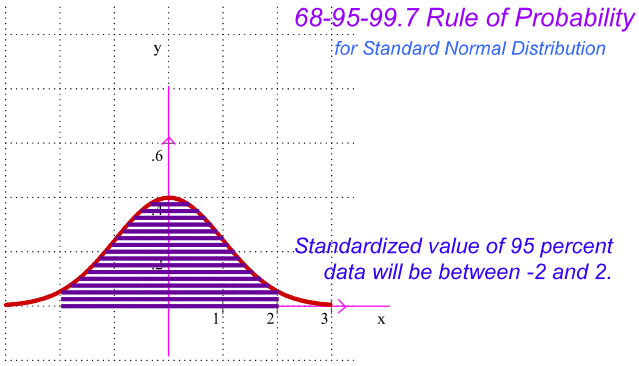
The Empirical Rule: Suppose the histogram of the data fits under a bell curse as above. So, it is symmetric around the vertical line through the mean $x = \overline{x}$: In this case,

1. At least **68.3 percent** of the observations will fall within 1 standard deviation of the mean, i.e, within $(\overline{x} - s, \overline{x} + s)$. This is clearly of not use.

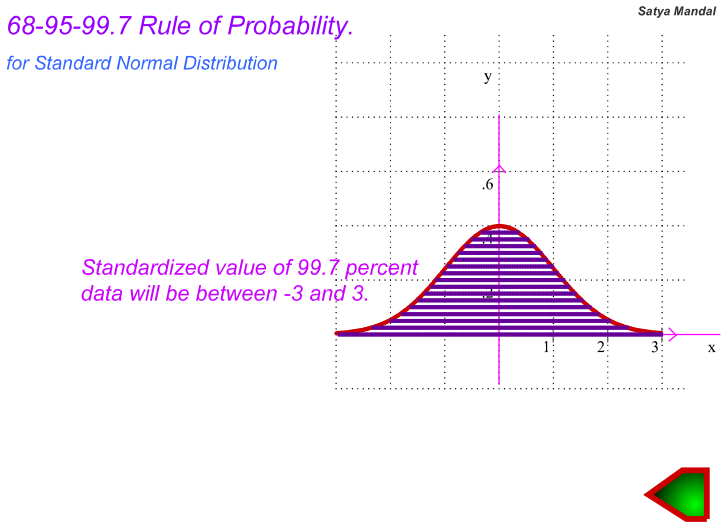
2. At least **95.4 percent** of the observations will fall within 2 standard deviations of the mean, i.e., within $(\overline{x} - 2s, \overline{x} + 2s)$.

3. At least **99.7 percent** of the observations will fall within 3 standard deviations of the mean, i.e., within $(\overline{x} - 3s, \overline{x} + 3s)$.




68-95-99.7 Rule of Probability
for Standard Normal Distribution
y
.6
.2
1
2
3
x
Standardized value of 95 percent
data will be between -2 and 2.

Next



Question: What does it mean when the variance or mean deviation of some data is zero? The answer is that all the data members are EQUAL!

Practice Problem. Consider the all the data set given above. For each problem, compute the mean and standard deviation of the data and find what percentage of the data are within one, two, or three standard deviations from the mean.

2.2.2 Use of the Frequency Table

When a frequency table is given, we can use new formulas to compute the mean and variance of the data.

Formulas: Suppose frequency table of a data set is given, where the data value x_i has frequency f_i. Let $n = \sum f_i$ denote the data size. Then,

1. The mean
$$\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}$$

2. The variance
$$s^2 = \frac{\sum f_i (x - \overline{x})^2}{n-1} = \frac{(\sum f_i x_i^2) - n\overline{x}^2}{n-1}$$

3. If the data is given in a class frequency table of the grouped data, we use the same formula, with x_i as the class mark, which is the average of the class limits. In this case, we only get an estimate of the variance of the original data.

Example 2.2.2. The following table expand the frequency table (1.1) of the time taken to complete a lap by a race car (example 1.1.1) to compute mean and variance using the above formulas.

Time x_i	**frequency** f_i	$f_i x_i$	$f_i x_i^2$
46	1	46	2116
47	1	47	2209
48	3	144	6912
49	3	147	7203
50	4	200	10000
51	6	306	15606
52	4	208	10816
53	5	5265	14045
54	5	270	14580
55	2	110	6050
56	1	56	3136
Total $\sum$	35	1799	92673

So, the mean
$$\overline{x} = \frac{\sum f_i x_i}{n} = \frac{1799}{35} = 51.4$$
and the variance
$$s^2 = \frac{(\sum f_i x_i^2) - n\overline{x}^2}{n-1} = \frac{92673 - 35 * (51.4)^2}{34} = 6.0118$$

Remark: When computing power was not in abundance (20 or 30 years ago), as it is now, we used to use such tables to do the computations. Such methods are out of date by now. We use TI-84 or other tools now.

Example 2.2.3 Consider the class frequency distribution of the data (1.2) on birth weight of some babies (exercise 1. 2. 5). We expand it, and use the above formula to compute

mean and variance.

Classes	**frequency** f_i	*class mark* x_i	$f_i x_i$	$f_i x_i^2$
$60.5 - 80.5$	9	70.5	634.5	44732.25
$80.5 - 100.5$	20	90.5	1810	163805
$100.5 - 120.5$	25	110.5	2762.5	305256.25
$120.5 - 140.5$	37	130.5	4828.5	630119.25
$140.5 - 160.5$	8	150.5	1204	181202
Total $\sum$	99		11239.5	1325114.75

So, the mean (approximate)

$$\overline{x} = \frac{\sum f_i x_i}{n} = \frac{11239.5}{99} = 113.53$$

and the variance (approximate)

$$s^2 = \frac{(\sum f_i x_i^2) - n\overline{x}^2}{n-1} = \frac{1325114.75 - 99 * 113.53^2}{99-1} = 500.997$$

Remarks.

1. Note that we can only get an approximate mean and variance if we use the class mark and with the above formula. If you use the original data, you will notice a difference.

2. As mentioned above, because of the availability of computers, the importance of such approximations has declined.

Use of calculators: We have had detailed discussions of various formulas for defining the mean, variance, and other constants. It is important to understand these concepts and formulas. It is equally important to appreciate the value and necessity of using calculators or other available software (like Excel). It is almost impossible (and unnecessary) to compute these constants manually and correctly, unless one is specially gifted with numerical computations. To compute the variance and standard deviation (with TI-84), do the following:

1. Follow the same steps used for computing the mean (using either raw data or the frequency table).

2. The calculator will give a list of numbers; S_X is the standard deviation.

3. The variance is the square of the standard deviation.

2.2.3 Problems on Variance, Standard Deviation

As before (Sec. 2.1.2), we would have problems of raw data and data in frequency tables. We compute measure of dispersions, for all problems in Sec. 2.1.2.

Exercise 2.2.1. Find the variance and standard deviation of the price, for stock price in Ex. 2.1.1:

138 142 127 137 148 130 142 133

Solution: Use TI-84.

Exercise 2.2.2. Find the variance and standard deviation of GPA, for the data in Ex. 2.1.2:

3.0 3.3 3.1 3.0 3.1 3.1

Find the median and mean GPA.

Exercise 2.2.3. Find the variance and standard deviation of the lifetime of these bulbs, for data in 2.1.3:

138 952 980 967 992 197 215 157

Find the mean and median lifetime of these bulbs. **Solution:** Use TI-84.

Exercise 2.2.4. Compute the variance and standard deviation of time taken by the athlete, for data in 2.1.4

Time x_i *(in sec.)*	**frequency** f_i
26	3
27	6
28	5
29	6
30	9
31	3
Total	32

Solution: Use TI-84.

Exercise 2.2.5. Compute the variance and standard deviation of the weight, at birth, of these babies, with data in Ex. 1.2.7, Ex. 2.1.5. **Solution:** Use TI-84.

Exercise 2.2.6. Compute the variance and standard deviation of the hourly wages, for data in Ex. 1.2.10, Ex. 2.1.6. **Solution:** Use TI-84.

Exercise 2.2.7. Find the variance, and standard deviation of typos in a book, for date in Ex. 2.1.7:

No. of Typos	156	158	159	160	162
Frequency	6	4	5	6	9

Solution: Use TI-84.

Exercise 2.2.8. Compute the variance and standard deviation of the length, at birth, of these babies, for data in Ex. 1.2.8, Ex. 2.1.8. **Solution:** Use TI-84.

Exercise 2.2.9. The following is the frequency table of weight (in pounds) of some salmon in a river.

weight x	31	32	33	34	35	36	37
Frequency f	3	2	4	5	6	5	9

Find the variance and standard deviation. **Solution:** Use TI-84.

Exercise 2.2.10. The following data represents the time (in minutes) taken by students to drive to campus.

23	17	19	24	42	33	20	22	15	9
26	37	29	19	35	18	30	21	11	23
13	27	32	32	23	35	25	33	24	23

Find the mean, variance, and the standard deviation of the data. **Solution:** Use TI-84.

Chapter 3

Probability

3.1 Introduction

The concept of probability is prevalent at a very basic human intuitive and intellectual level. Other synonyms of probability include likelihood and chances. Probabilistic statements are made on daily basis without any awareness that there may be some intuitive mathematical calculation involved behind such statements.

Most people are aware of different kinds of game of chances and gambling. Examples of such games include tossing coins, any dice rolling game and games of cards. It is universally accepted that when you toss a coin, likelihood of head showing up is fifty percent. When you roll a normal die, it is universally accepted that the likelihood of that a particular face will show up is one in six. There is also awareness of loaded coins and loaded dice, in which likelihood of an outcome (say head or the face six) is higher than that of other outcomes. It is a common sense that in a poker game, it is extremely unlikely that one would get three aces in a particular deal. A lot of people would not buy a lottery ticket because they believe, that it is extremely unlikely that he or she would ever win a multi million dollar or would not make money in the long run. These are intuitive or semi-mathematical understanding of probability of occurrence of certain events.

Statements regarding chances of departure of your flight on time would not be so uncommon. Same is true, regarding similar statements on chances of a thunderstorm, rain, snow or other whether related events. Statements regarding the chances of accidents would be another types probabilistic statements. One would not be surprised to hear a five year old making statements like "I will probably invite Aaron for my birthday". Such would mean that the child is aware of uncertainties of parental permission or his/her own indecisiveness.

When such statement that includes word like "likelihood", "chances" or "probability" are made, one is essentially talking about what they have experienced in the past and trying to project that the same pattern will continue in future.

Some of the early development of Probability, as a mathematical theory, originated in gambling. In the last century, this concept received further boost in genetics (more generally biosciences) and other branches of science.

3.1.1 Basic Concept of Probability

The concept of Probability attempts to quantify chances (or likelihood) of the occurrence of an event in a consistent manner. It uses human experience of the past to hypothesize and formulate a model. The model needs to behave in a manner that makes an overall sense. For example the chances of occurrence (say of raining today) and nonoccurrence of the same event should add to 100 percent (or probabilities of the same should add to one).

Our experience with tossing normal coins tells us that if we toss a coin for a large number of times, essentially half the time the head shows up. We assume that same pattern will continue in future. Therefore, we hypothesize that the probability that the head will show up is .50. On the other hand, if through extensive experimentation with a coin, we conclude that the ratio of the number of Heads to the number of tosses remains close to and moves around .49 then we would hypothesize that the probability of heads is .49. Such minor difference would be of serious consideration for the gambling houses.

Similarly, our experience with rolling normal dice tells us that when we roll a die a particular face (say face six) would show up, is essentially once is six times. So, we hypothesize that probability that face six will show up is 1/6. Contrary to this, there are loaded dice. You may have a loaded die that you experimented with and determined face six shows up 40 percent of the times. So, you would hypothesize that probability that face six will show up for this loaded die is .40.

Similar data may be collected for road accidents and probabilistic hypothesis (model) could be made regarding number of daily accidents on a street.

These examples explain the basic notion of probability. The probability of an event is hypothesized (modeled), as the "relative frequency", the ratio of occurrences of the EVENT to the total number of times the EXPERIMENT is repeated (or experienced in the past). Following are some pictures of coin toss experiments:

The first coin is set up to have 1 in 2 chances of Jayhawk

This second coin is set up to have 3 in 4 chances of Jayhawk

This third coin is set up to have 1 in 10 chances of Jayhawk

3.1.2 Basic Set theoretic Definitions

Definition. By a **set** S we mean a collection of objects. The objects in this set S are also called **elements** or **members** of the set. A set E is said to be a **subset** of a set S if each element of E is also an element of S. We write

$$E \subseteq S$$

to mean that E is a subset of S. Obviously, a subset E of S is a smaller collection than or equal to S. The following are some examples. We also explain the **usage of braces** to describe a set.

1. Let D = the collection of all 52 cards in a deck. Then D is a set. Let E be the collection of all the hearts in this deck. Then E is a subset of D. In brace notation

$$E = \{x \in D : x \; is \; a \; Heart\}$$

We read this as "E is equal to the set of all x in D such that x is a heart". We use the symbol $\in$ to abbreviate the word "in" or "is an element of".

2. Let T be the collection of all those who filed a tax return to the IRS for the year 2001. Then T is a set. Let L be the collection of those whose Adjusted Gross Income in the return was less or equal to \$30,000. Then L is a subset of T. Let C be the collection of those who declared capital gains income. Then C is a subset of T. We write

$$L \subseteq T, \qquad C \subseteq T$$

In brace notation

$$L = \{x \in T : \textit{the Adjusted Gross Income of } x \textit{ is less or equal to } \$30,000\}.$$

3. Let N be the collection of all integers, and let E be the collection of even integers. Then N, E are set and $E \subseteq N$ In brace notation

$$N = \{n : n \textit{ is an integer}\} \qquad E = \{n \in N : n \textit{ is even}\}.$$

We also write

$$N = \{\cdots, -2, -1, 0, 1, 2, \cdots\} \qquad E = \{\cdots, -4, -2, 0, 2, 4, \cdots\}$$

4. Let R be the set of all (real) numbers. Let I be the set of all numbers between 0 and 1, not equal to 0,1. Then R, I are sets and I is a subset of R. In brace notation

$$R = \{x : x \textit{ is a real number}\}, \qquad I = \{x \in R : 0 < x < 1\}.$$

There are other (interval) notations, like
$R = (-\infty, \infty)$,
$I = (0, 1)$.

5. $S = \{1, 7, 13, 17, 19\}$ is a set.

6. Let S be the collection of you and your siblings, B be the collection of your brothers, and F be the collection of your sisters. Then S, B, F are sets and we have

$$F \subseteq S \qquad B \subseteq S.$$

3.1.3 Statistical Experiments and Sample Space

We use the above language of set theory, only in the context of random experiments.

Definition. We introduce three fundamental definitions:

1. A **statistical experiment** is a procedure that produces exactly one out of many possible outcomes. All the possible outcomes are known, but which outcome will result when you perform the experiment is not known. A statistical experiment is also called a **random experiment**.

2. Given an experiment, the set of all possible outcomes is called the **sample space**. The sample space is usually denoted by S.

3. Given an experiment, an **outcome** of the experiment is also called a **sample point**. So, the sample space consists of sample points.

Examples. Following are some standard examples of samples spaces.

1. Suppose the experiment is tossing a coin. The outcomes are H (heads) and T (tails). So, the sample space is $S = \{H, T\}$.

2. Suppose the experiment is tossing a coin twice. The sample points (or outcomes) are HH, HT, TH, TT and the sample space is $S = \{HH, HT, TH, TT\}$.

3. Your experiment is rolling a die. The outcomes are $1, 2, 3, 4, 5, 6$ and the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

4. Suppose that the experiment is rolling a die twice. Then the sample space is

$$S = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\} \tag{3.1}$$

   In brace notation, we can write the same as

$$S = \{(i, j) : i = 1, 2, 3, 4, 5, 6 \text{ and } j = 1, 2, 3, 4, 5, 6\}.$$

5. Suppose the experiment is to determine the number of road accidents in Lawrence on a particular day. So, the sample space is $S = \{0, 1, 2, 3, \ldots\}$.

6. Suppose the experiment is to determine the sex of an unborn child. Then the sample space is $S = \{Female, Male\}$.

7. Suppose the experiment is to determine the blood group of a patient in a lab. Then the sample space is $S = \{O, A, B, AB\}$.

8. Suppose the experiment is to observe the annual wheat production in Kansas. Then the sample space is

$$S = \{x : x \text{ is a nonnegative Number}\} = \{x \in R : x \geq 0\} = [0, \infty).$$

9. Suppose that the experiment is rolling a die three times. Then the sample space can be written as

$$S = \{(i, j, k) : i = 1, 2, 3, 4, 5, 6 \text{ and } j = 1, 2, 3, 4, 5, 6 \text{ and } k = 1, 2, 3, 4, 5, 6\}.$$

A a set can be described or written in many different ways, as was done above, in some cases. There is no hard and fast rules. But some jargons, as above, have become standard.

Definition. The sample space S is called a finite sample space if S has only a finite number of outcomes. If S has infinite elements, it is called an infinite sample space. Note that examples 1, 2, 3, 4, 6, 7, 9 have finite sample spaces, and examples 5, 8 have infinite sample space.

We define Events, in the context of a random experiment.

Definition. Given an experiment and its sample space S, the following are important definitions.

1. A subset E of the sample space S is called an **event**. So, an event E consists of outcomes, and we have
$$E \subseteq S$$
We say that E **would have occurred, if the outcome of the experiment would be in** E, when the experiment is performed.

2. There is a special event called the **empty event** or **impossible event**, denoted by ϕ. It is defined as the event with **no outcomes** in it. So, the impossible event consists of no outcome. Therefore, the impossible event would never occur.

3. Since S is also a subset of itself, S is an event. This event S is called the **sure event**. Since the outcome of the experiment would always be in S, the sure event S is sure to occur.

4. A **simple event** consists of a single outcome.

Examples. The following are some examples of events.

1. Consider example 2 above -the experiment on the coin toss. Let E be the event that at least one of the tosses gave T, and let F be the event that both tosses gave the same face. Then
$$E = \{HT, TH, TT\}, \quad \textit{and} \quad F = \{HH, TT\}.$$

2. Look at example 4 above-the experiment on rolling a die. Let E_5 be the event that first die showed 5. Then
$$E_5 = \{(5,1), (5,2), (5,3), (5,4), (5,5), (5,6)\}.$$
Let T_5 be the event that the sum of the two "rolls" is 5. Then
$$T_5 = \{(1,4), (2,3), (3,2), (4,1)\}.$$
Let T_1 be the event that the sum of the two rolls is 1. Because T_1 has no outcome, $T_1 = \phi$ is the impossible event. Likewise, let T_{13} be the event that the sum of the two rolls is 13. Then $T_{13} = \phi$ is also the impossible event.

3. Look at the example 5 above-the experiment on road accidents. Let E be the event that there is no accident on that day. Then

$$E = \{0\}.$$

4. Look at example 8 above-the experiment on annual wheat production. Let E be the event that there will be more than 1000 units of wheat production in 1998. Then

$$E = (1000, \infty).$$

3.1.4 The Definition of Probability

Given a sample space S, in the MATHEMATICS of probability we have hypotheses and rules for how to compute the probability of an event E. Although the MATHEMATICS of probability was modeled based on our past experiences, we do not derive anything from our intuitive ideas. We would fully be guided by the precise hypotheses, rules and laws that we set up.

For now we would be dealing with finite sample spaces.

Definition. Suppose we have a random experiment, and it has a finite sample space

$$S = \{e_1, e_2, \ldots, e_N\} \quad \text{with } N \text{ outcome.}$$

The **probability** of a simple event $\{e_i\}$ is a number (possibly given) denoted by $P(\{e_i\})$ or loosely written as $P(e_i)$, which has the following properties:

1. $0 \leq P(e_i) \leq 1$ for all $i = 1, 2, \ldots, N$

2. The sum of the probabilities of all the simple events is 1:

$$P(e_1) + P(e_2) + \ldots + P(e_N) = 1.$$

3. If E is an event, then the probability $P(E)$, of occurrence of E, is defined as the sum of the probabilities of all the simple events $e_i \in E$:

$$P(E) = \sum_{e_i \in E} P(e_i) \tag{3.2}$$

So, we also have

$$P(\textit{impossible Event}) = P(\phi) = 0 \qquad P(\textit{Sure Event}) = P(S) = 1$$

3.2 Probability Tables and Equally Likely

Remark. If we know the probabilities $P(e_i)$ of all the simple events $\{e_i\}$, we will be able to compute the probability of any event E using Formula 3.2. The probabilities of the simple events would

1. either be given explicitly, say by a table, or
2. OR, a rule or a formula will be given, to compute it.

One of the most frequently used models to compute probabilities of simple events is called **equally likely outcomes**, as follows.

Definition. Let $S = \{e_1, e_2, , e_N\}$ be a finite sample space. We say that all the outcomes are **equally likely**, if all the outcomes have the same probability. So, in this case, we have

$$P(e_1) = P(e_2) = \cdots = P(e_N) = \frac{1}{N}$$

Also, in this case, for an event E, from Formula 3.2

$$P(E) = \sum_{e_i \in E} P(e_i) = \sum_{e_i \in E} \frac{1}{N} = \frac{no.\ of\ outcomes\ in\ E}{N}$$

If $n(E)$ denotes the number of outcomes in E then

$$P(E) = \frac{n(E)}{n(S)} \tag{3.3}$$

3.2.1 Problems: on Tabular probability

We work out some problems in this section, where probability of simple events is given in a table.

Exercise 3.2.1. The following table gives the blood group distribution of a certain population.

Blood Group Distribution				
Blood Group	O	A	B	AB
percent of population	47	42	8	3

Find the probability that a random sample of blood will be of Blood Group A or B or AB.
Solution: Here $S = \{O, A, B, AB\}$ and we want to compute the probability $P(E)$ of the event $E = \{A, B, AB\}$. From the table

$$P(O) = .47, \quad P(A) = .42 \quad P(B) = .08 \quad P(AB) = .03$$

$$\text{So,} \quad P(E) = P(\{A, B, AB\}) = P(A) + P(B) + P(AB) = .42 + .08 + .03 = .53$$

Exercise 3.2.2. A student wants to pick a school based on its grade distribution. Following is the most recent grade distribution in a school:

Grade Distribution (*Unreal Data*)					
Grades	A	B	C	D	F
percent of students	19	33	31	14	3

Find the probability that a randomly picked student will have at least a B average.
Solution: Here $S = \{A, B, C, D, F\}$ and we want to compute the probability $P(E)$ of the event $E = \{A, B\}$. From the table

$$P(A) = .19, \quad P(B) = .33 \quad P(C) = .31 \quad P(D) = .14, \quad P(F) = .03$$

$$\text{So,} \quad P(E) = P(\{A, B\}) = P(A) + P(B) = .19 + .33 = .52$$

Exercise 3.2.3. The following table gives the probability distribution of a loaded die.

Probability Distribution for the Die						
Face	1	2	3	4	5	6
probability	0.20	0.15	.015	0.10	0.05	0.35

Find the probability that the face 2 or 3 or 6 will show up when you roll the die.
Solution: Here $S = \{1, 2, 3, 4, 5, 6\}$ and we want to compute the probability $P(E)$ of the event $E = \{2, 3, 6\}$. From the table

$$P(1) = 0.20, \quad P(2) = 0.15, \quad P(3) = 0.15, \quad P(4) = 0.10, \quad P(5) = 0.05, \quad P(6) = 0.35$$

$$\text{So,} \quad P(E) = P(\{2, 3, 6\}) = P(2) + P(3) + P(6) = 0.15 + 0.15 + 0.35 = .65$$

Exercise 3.2.4. An arbitrary spot is selected in a swamp. The depth (in feet) of water in the swamp has the following probability distribution:

Depth Distribution									
depth	0+	1+	2+	3+	4+	5+	6+	7+	8+
probability	.1	.2	.09	.17	.13	.11	.08	.07	.05

1. What is the probability that the depth at an arbitrary spot is less than three feet?
2. What is the probability that the depth at an arbitrary spot is 3 feet or higher?

Solution: Here, the sample space is $S = \{0+, 1+, 2+, 3+, 4+, 5+, 6+, 7+, 8+\}$.
From the table, probability $P(0+) = .1, P(1+) = .2, P(2+) = .09$ and so on.

1. Let E be the event that depth at an arbitrary spot is less than three feet.
   Then $E = \{0+, 1+, 2+\}$. So,
   $P(E) = P(0+) + P(1+) + P(2+) = .1 + .2 + .09 = .39$.

2. Let F be the event that the spot is 3 feet or higher.
   Then $F = \{3+, 4+, 5+, 6+, 7+, 8+\}$. So,
   $P(F) = P(3+) + P(4+) + P(5+) + P(6+) + P(7+) + P(8+)$
   $= .17 + .13 + .11 + .08 + .07 + .05 = .61$

Exercise 3.2.5. Van pool can carry 7 people. Following is the distribution of number of riders in the van on a given day.:

Distribution of number of passengers							
number of passengers	1	2	3	4	5	6	7
probability	0	.12	.22	.23	.28	.08	.07

1. What is the probability that there will be at most 4 riders?

2. What is the probability that there will be less than 4 riders?

3. What is the probability that there will be more than 4 riders?

4. What is the probability that the van will not be full on a particular day?

Solution: Here, the sample space is $S = \{1, 2, 3, 4, 5, 6, 7\}$.
From the table, probability $P(1) = 0, p(2) = .12, P(3) = .22$ and so on.

1. Let E be the event that there will be at most 4 riders.
   Then $E = \{0, 1, 2, 3, 4\}$.
   So, $P(E) = P(0) + P(1) + P(2) + P(3) + P(4) = 0 + .12 + .22 + .23 = .57$.

2. Let F be the event that 3 there will be less than 4 riders.
   Then $F = \{0, 1, 2, 3\}$.
   So, $P(F) = P(0) + P(1) + P(2) + P(3) = 0 + .12 + .22 = .34$.

3. Let G be the event that that there will be more than 4 riders.
   Then, $G = \{5, 6, 7\}$.
   So, $P(G) = P(5) + P(6) + P(7) = .28 + .08 + .07 = .43$.

4. Let H be the event that there the van will not be full.
   Then, $H = \{0, 1, 2, 3, 4, 5, 6\}$.
   $P(H) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6)$
   $= 0 + .12 + .22 + .23 + .28 + .08 = .93$

Exercise 3.2.6. Following is the distribution of hourly wages (in whole dollars) earned by workers in an industry:

Wage Distribution														
Wage in USD	7	8	9	10	11	12	13	14	15	16	17	18	19	20
probability	.04	.06	.07	.09	.11	.12	.14	.11	.09	.08	.04	.03	.01	.01

An employee is selected at random.

1. What is the probability that the randomly selected worker makes less than 10 dollars an hour?

2. What is the probability that the randomly selected worker makes at least \$10 an hour?

3. What is the probability that the randomly selected worker makes between \$12-\$16 an hour?

Solution: Here, the sample space is $S = \{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. From the table, probability $P(7) = .04, p(8) = .06, P(9) = .07$ and so on.

1. Let E be the event that the randomly selected worker makes less than 10 dollars an hour. Then $E = \{7, 8, 9\}$.
   So, $P(E) = P(7) + P(8) + P(9) = .04 + .06 + .07 = .17$

2. Let F be the event that the randomly selected worker makes at least \$10 an hour. So,
   Then $F = \{10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$.
   $P(F) = P(10) + P(11) + P(12) + P(13) + P(14) + P(15) + P(16) + P(17) + P18) +$
   $P(19) + P(20)$
   $= .09 + .11 + .12 + .14 + .11 + .09 + .08 + .04 + .03 + .01 + .01 = .82.$

3. Let G be the event that that worker makes between \$12-\$16 an hour. Then, $G = \{12, 13, 14, 15, 16\}$.
   So, $P(G) = P(12) + P(13) + P(14) + P(15) + P(16) = .12 + .14 + .11 + .09 + .08 = .54$

Exercise 3.2.7. In a school district, the distribution of number of students in a class has the following probability distribution:

Distribution of number of students																
no of students	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
probability	.03	.04	.06	.07	.10	.12	.13	.11	.09	.07	.06	.04	.03	.02	.02	.01

A child is selected at random from the school district.

1. What is the probability that the child will be in a class of at least 20?
2. What is the probability that the child will be in a class of at most 10?
3. What is the probability that the child will be in a class of less than 10?

Solution: Here, the sample space is $S = \{8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23\}$. From the table, probability $P(8) = .03, p(9) = .04, P(10) = .06$ and so on.

1. Let E be the event that the the child will be in a class of at least 20.
   Then $E = \{20, 21, 22, 23\}$. So,
   $P(E) = P(20) + P(21) + P(22) + P(23) = .03 + .02 + .02 + .01 = .08$
2. Let F be the event that the child will be in a class of at most 10.
   Then $F = \{8, 9, 10\}$. So,
   $P(F) = P(8) + P(9) + P(10) = .03 + .04 + .06 = .13.$
3. Let G be the event that that child will be in a class of less than 10.
   Then, $G = \{8, 9\}$. So,
   $P(G) = P(8) + P(9) = .03 + .04 = .07.$

3.2.2 Problems: on Equally Likely model

We work out some problems in this section, where all outcomes are equally likely. As mentioned above (3.3), we use the formula

$$P(E) = \frac{n(E)}{n(S)}$$

Exercise 3.2.8. An urn contains 7 apples and 3 oranges and 5 pears. One piece of fruit is picked at random. Find the probability that

1. the fruit is an apple,

2. the fruit is either an apple or a pear, and

3. the fruit is an orange.

Solution: Here, the sample space is
$S = \{A1, A2, A3, A4, A5, A6, A7, O1, O2, O3, P1, P2, P3, P4, P5\}$.
So, $n(S) = 7 + 3 + 5 = 15$

1. Let E be the event that the selected fruit is an apple.
   Then $E = \{A1, A2, A3, A4, A5, A6, A7\}$.
   So, $n(E) = 7$
   $P(E) = \frac{n(E)}{n(S)} = \frac{7}{15}$

2. Let F be the event that the selected fruit is an apple or a pear.
   Then $F = \{A1, A2, A3, A4, A5, A6, A7, P1, P2, P3, P4, P5\}$.
   So, $n(F) = 7 + 5 = 12$
   $P(F) = \frac{n(F)}{n(S)} = \frac{12}{15}$

3. Let G be the event that the selected fruit is an orange.
   Then $G = \{O1, O2, O3\}$.
   So, $n(G) = 3$
   $P(G) = \frac{n(G)}{n(S)} = \frac{3}{15}$

Exercise 3.2.9. A die is rolled twice. Find the probability that

1. the sum is 8,

2. only 2 or 3 showed up in both the rolls, and

3. the first roll produced a bigger number.

4. that the sum of two numbers is 1.

Solution: The sample space S is given above (3.1).
So, $n(S) = 36$.

1. Let E be the event that the sum is 8.
   Then $E = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ and $n(E) = 5$.
   So, $P(E) = \frac{n(E)}{n(S)} = \frac{5}{36}$.

2. Let F be the event that only 2 or 3 showed up in both the rolls.
   Then, $F = \{(2,2),(2,3),(3,2),(3,3)\}$ and $n(F) = 4$
   So, $P(F) = \frac{n(F)}{n(S)} = \frac{4}{36}$

3. G be the event that the first roll produced a bigger number.
   Then,
$$G = \left\{ \begin{array}{lllll} (2,1) & & & & \\ (3,1) & (3,2) & & & \\ (4,1) & (4,2) & (4,3) & & \\ (5,1) & (5,2) & (5,3) & (5,4) & \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) \end{array} \right\}$$
   So, $n(G) = 15$.
   $P(G) = \frac{n(G)}{n(S)} = \frac{15}{36}$.

4. Let H be the event that the sum is 1.
   In fact, sum of the two numbers is never equal to 1.
   So, $H = \varphi$ is the impossible event .
   So, $n(H) = 0$.
   So, $P(H) = \frac{n(H)}{n(S)} = \frac{0}{36} = 0$.

Exercise 3.2.10. A letter is chosen at random from the letters of the English alphabet. Find the probability that

1. the letter is either I or U,

2. the letter is in the word ALWAYS, and

3. the letter is not in the word NEVER.

Solution: Here, the sample space is $S = \{A, B, C, D, E, \ldots, X, Y, Z\}$.
So, $n(S) = 26$.

1. Let E be the event that the selected letter is I or U.
   Then $E = \{I, U\}$.
   So, $n(E) = 2$
   $P(E) = \frac{n(E)}{n(S)} = \frac{2}{26}$.

2. Let F be the event that letter is in the word ALWAYS.
   Then $F = \{A, L, W, Y, S\}$.
   So, $n(F) = 5$.
   $P(F) = \frac{n(F)}{n(S)} = \frac{5}{26}$.

3. Let G be the event that the letter is not in the word NEVER.
   Then $G =$ all letters, except N, E, V, R.
   So, $n(G) = 26 - 4 = 22$.
   $P(G) = \frac{n(G)}{n(S)} = \frac{22}{26}$.

3.3 Laws of Probability

In the above sections, we defined probability for finite sample spaces. Some of the laws of probability will be discussed in this section. In fact, probability laws are similar to that of law of area, volume or weight.

Notations from Set Theory Following notations from the set theory will be useful in our context of sample spaces and events.

Notations. Let S be a set and E, F be two subsets of S.

1. The **union** $E \cup F$, of E and F is the set defined as follows:
$$E \cup F = \{x \in S : x \in E \text{ or } x \in F\}.$$
   So, if you put together the elements of E and F in a single collection, you get the union $E \cup F$.

2. The **intersection** $E \cap F$, of E and F is defined as follows:
$$E \cap F = \{x \in S : \text{both } x \in E \text{ and } x \in F\}.$$
   So, if you take all the elements common to both E and F, you get the intersection of E and F.

3. The **complement** E^c, of E is defined as follows:
$$E^c = \{x \in S : x \notin E\}.$$
   So, the complement E^c of E is the collection of all the elements in S that are not in E.

Remark. It would be enough, for our purpose, if we can interpret the above definitions in our context of sample spaces and events.
For us, S will be a fixed sample space and E, F will be events.

1. $E \cup F$ is the event that consists of all outcomes that are either in E or in F (or both). So the occurrence of either E or of F is the same as the occurrence of $E \cup F$. That is why $E \cup F$ is also denoted by $(E \ or \ F)$.

$$E \cup F = (E \ or \ F).$$

2. $E \cap F$ is the event that consists of all the outcomes that are both in E and F. So the simultaneous occurrence of E and F is the same as the occurrence of $E \cap F$. That is why $E \cap F$ is also denoted by $(E \ and \ F)$. Notationally,

$$E \cap F = (E \ and \ F)$$

3. Similarly, E^c is the event that consists of all the outcomes in S that are not in E. So, the occurrence of E^c is the same as the nonoccurrence of E. Notationally,

$$E^c = (not \ E)$$

3.3.1 Laws of Probability

Following are some of the laws of probability. First, probability behaves like area and the laws of probability are like that of area.

Some formulas: Let S be sample space and let E and F be two events.

1. Then,
$$\begin{cases} P(E \cup F) = P(E) + P(F) - P(E \cap F) & \text{which is :} \\ P(E \ or \ F) = P(E) + P(F) - P(E \ and \ F) \end{cases} \tag{3.4}$$
We subtract $P(E \cap F)$ because we counted it twice: once in $P(E)$ and once in $P(F)$.

2. **Definition.** We say E and F are **mutually exclusive**, if E and F cannot occur simultaneously. This is if E and F have no outcome or $E \cap F = \phi$. Since $P(\phi) = 0$, it follows from (3.4) that if E and F are mutually exclusive then
$$P(E \cup F) = P(E) + P(F)$$

3. We also have
$$P(E^c) = P(not \ E) = 1 - P(E). \tag{3.5}$$
Definition. Let E be an event. We say that the **odds of an event E occurring are a to b**, if
$$P(E) = \frac{a}{a+b}$$

Remark: This concept of ODDS is used often in gambling and horse races. When the odds in favor of a horse are 2 to 3, essentially this means that the probability the horse will win is 2/5. We say "essentially" because in actual betting, the probability is actually slightly less than 2/3, so that in the long run the gambling establishment makes more money than it gives. (This instructor is not particularly experienced in such betting or horse races.)

3.3.2 Problems: on Laws of Probability

Exercise 3.3.1. Let E, F, G be three events. It is given

$$P(E) = 0.3, \quad P(F) = 0.7, \quad P(G) = 0.6, \quad P(E \cap F) = 0.2, \quad P(E \cup G) = 0.7$$

Find the probability that

1. E or F occur,
2. both E and G occur, and
3. E does not occur.

Solution:

1. $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F) = 0.3 + 0.7 - .02 = 0.8$
2. $P(E \text{ and } G) = P(E) + P(G) - P(E \text{ or } G) = 0.3 + 0.6 - 0.7 = 0.2$
3. $P(\text{not } E) = 1 - P(E) = 1 - 0.3 = 0.7$

Exercise 3.3.2. Let E, F, G be events.

1. If the odds in favor of E are 3 to 5, find the probability that E occurs.
2. If the odds against F are 3 to 4, find $P(F)$.
3. If $P(G) = 7/10$, what are the odds in favor of G?

Exercise 3.3.3. The probability that a Christmas tree is taller than 6 feet is .30; the probability that a Christmas tree weighs more than sixty pounds is 0.25; and the probability that a Christmas tree is either taller than 6 feet or more than sixty pounds is .4.

1. Find the probability that a Christmas tree is both taller than 6 feet and weighs more than sixty pounds.

2. Find the probability that a Christmas tree is not taller than 6 feet.

3. Find the probability that a Christmas tree is either less than 6 feet tall or less than sixty pounds in weight.

Solution: Let E be the event that the selected tree is taller than 6 feet.
Let F be the event that the selected tree is heavier than sixty pounds. We are given

$$P(E) = 0.30, \quad P(F) = 0.25, \quad P(E \text{ or } F) = 0.4$$

1. $P(E \text{ and } F) = P(E) + P(F) - P(E \text{ or } F) = 0.30 + 0.25 - 0.4 = 0.15$

2. $P(\text{not } E) = 1 - P(E) = 1 - 0.3 = 0.7$

3. In fact $\text{not}(E \text{ and } F)$ is the event that the selected tree is either less than 6 feet tall or less than sixty pounds in weight
$P(\text{not}(E \text{ and } F)) = 1 - P(E \text{ and } F)1 - 0.15 = 0.85.$

Exercise 3.3.4. The probability that a student majors in liberal arts is .44; the probability that a student majors in business is .33; and the probability that a student majors in either liberal arts or business is .65. Find the probabilities

1. that a student majors in both liberal arts and business.

2. that a student majors in neither liberal arts nor business.

Solution: Let E be the event that the selected student majors in liberal arts.
Let F be the event that the selected student majors in business. We are given

$$P(E) = 0.44, \quad P(F) = 0.33, \quad P(E \text{ or } F) = 0.65$$

1. $P(E \text{ and } F) = P(E) + P(F) - P(E \text{ or } F) = 0.44 + 0.33 - 0.65 = 0.12$

2. In fact, $\text{not}(E \text{ or } F)$ is the event that the selected student majors neither in liberal arts nor in business.
$P(\text{not}(E \text{ or } F)) = 1 - P(E \text{ or } F) = 1 - 0.65 = 0.35$

Exercise 3.3.5. In a restaurant menu, entrees are served with rice product or potato product or others. Probability that an entree is served with rice product is .35 , probability that an entree is served with potato product is .40, probability that an entree is served with both is .15. What is the probability that an entree is served with either rice product

or potato product?
Solution: Let E be the event that an entree is served with rice product
and F be the event that an entree is served with potato product. We are given

$$P(E) = .35, \quad P(F) = .40, \quad P(E \text{ and } F) = .15.$$

We need to compute $P(E \text{ or } F)$.
$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F) = .35 + .40 - .15 = .60$

Exercise 3.3.6. Usual infections can be bacterial or viral. Probability that a person will get a bacterial infection (next winter) is .35, probability that a person will get a viral infection is .65, probability that a person will get either a bacterial or a viral infection is .85.

1. What is the probability that a person will get both next winter?

2. What is the probability that a person will not get an infection next winter?

Solution: Let E be the event that a person will get a bacterial infection
and F be the event a person will get a viral infection. We are given

$$P(E) = .35, \quad P(F) = .65, \quad P(EorF) = .85.$$

1. We need to compute $P(E \text{ and } F)$.
   $P(E \text{ and } F) = P(E) + P(F) - P(E \text{ or } F) = .35 + .65 - .85 = .15$

2. We need to compute $P(not(E \text{ and } F))$.
   $P(not(E \text{ and } F)) = 1 - P(E \text{ or } F) = 1 - .85 = .15$

Exercise 3.3.7. You go for an examination of upper stomach (EGD) and lower stomach (colonoscopy). Probability that some problem in upper stomach would be found is .15, probability that some problem in lower stomach would be found is .20 and probability that some problem both in lower and upper stomach would be found is .07.

1. What is the probability that some problem would be found in either upper of lower stomach?

2. What is the probability that it would found that both upper and lower stomach are healthy?

Solution: Let E be the event that some problem would be found in the upper stomach and F be the event that some problem would be found in the lower stomach. We are given

$$P(E) = .15, \quad P(F) = .20, \quad P(E \text{ and } F) = .07.$$

1. We need to compute $P(E \text{ or } F)$.
   $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F) = .15 + .20 - .07 = .28$

2. We need to compute $P(not(E \text{ and } F))$.
   $P(not(E \text{ or } F)) = 1 - P(E \text{ or } F) = 1 - .28 = .72$

Exercise 3.3.8. Probability that a person owns a domestic car is .55, probability that a person owns an import is .55 and the probability that a person owns both is .20.

1. What is the probability that a person owns either a domestic or an import?

2. Also what is the probability that a person owns none?

Solution: Let E be the event that a person owns a domestic
and F be the event that a person owns an import. We are given

$$P(E) = .55, \quad P(F) = .55, \quad P(EandF) = .20.$$

1. We need to compute $P(E \text{ or } F)$.
   $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F) = .55 + .55 - .20 = .90$

2. We need to compute $P(not(EandF))$.
   $P(not(E \text{ or } F)) = 1 - P(E \text{ or } F) = 1 - .90 = .10$

Problems on the second formula: (3.5)

Exercise 3.3.9. In a county, 38 percent of the community is a minority. What is the probability that a randomly selected jury will not be a minority?

Solution: Let E be the event that the jury will be a minority. Then $P(E) = .38$.
Therefore, the answer is $P(not\ E) = 1 - P(E) = 1 - .38 = .62$.

Exercise 3.3.10. In a school district, probability that a student will be in a class of less than 10 students is .27. The probability that a student will be in a class of less than 20 students is .38. What is the probability that a randomly selected student will be a class 10 or more?

Solution: Let E be the event that the student will be in a class of less that 10 students. Then, $P(E) = .27$.
So, the answer is $P(not\ E) = 1 - P(E) = 1 - .27 = .73$.

Exercise 3.3.11. In a swamp, probability that the depth at a random spot is higher than 4 feet is .17. What is the probability that at a random spot, the depth is four feet or less?

Solution: Let E be the event that at a random spot the depth is higher than 4 feet.
So, $P(E) = .17$. So, the answer is
$P(not\ E) = 1 - P(E) = 1 - .17 = .83$.

Exercise 3.3.12. It is known that 43 percent of the work force in a town earns more than $37,000 annually. What is the probability that a randomly selected working person would make at most $37,000 annually?

Solution: Let E be the event that a randomly selected working person would make more than $37,000 annually.
Therefore, $P(E) = .43$.
So, the answer is $P(not\ E) = 1 - P(E) = 1 - .43 = .57$.

Exercise 3.3.13. It is known that you can get an empty seat in the bus 64 percent of the rides. What is the probability that on a particular ride would not get a seat?

Solution: Let E be the event that you get an empty seat in the bus.
So, $P(E) = .64$. Therefore, the answer is
$P(not\ E) = 1 - P(E) = 1 - .64 = .36$.

3.4 Counting Techniques and Probability

Counting techniques are important and useful. The following are some interesting examples:

1. the number of English words (formal) of 5 letters, (A formal word is any sequence of letters from the English alphabet. For example, eefzq is a formal word.)
2. the number of ways you can deal a hand of 13 cards from a deck of 52 cards, or
3. the number of ways you can assign the first row of 11 seats to 231 guests.

Before we go further into counting, let us recall the factorial notation.
Notations. Let n be a positive integer. Then the n! (read as n-factorial) is defined as

$$n! = 1 \cdot 2 \cdots (n-2) \cdot (n-1) \cdot n \qquad 0! = 1.$$

n-factorial is the product of all integers from 1 up to n.

One of the main tools for counting is the following principle:

The Basic Counting Principle. Suppose we have an experiment that is a combination of r sub-experiments, performed one after the other, such that

1. **[Exp.1]** the first sub-experiment has $\mathbf{n_1}$ outcomes;

2. **[Exp.2]** corresponding to each outcome of the first sub-experiment, the second sub-experiment has $\mathbf{n_2}$ outcomes;

3. **[Exp.3]** corresponding to each outcome of the first and the second sub-experiments, the third sub-experiment has $\mathbf{n_3}$ outcomes;

4. $\cdots$

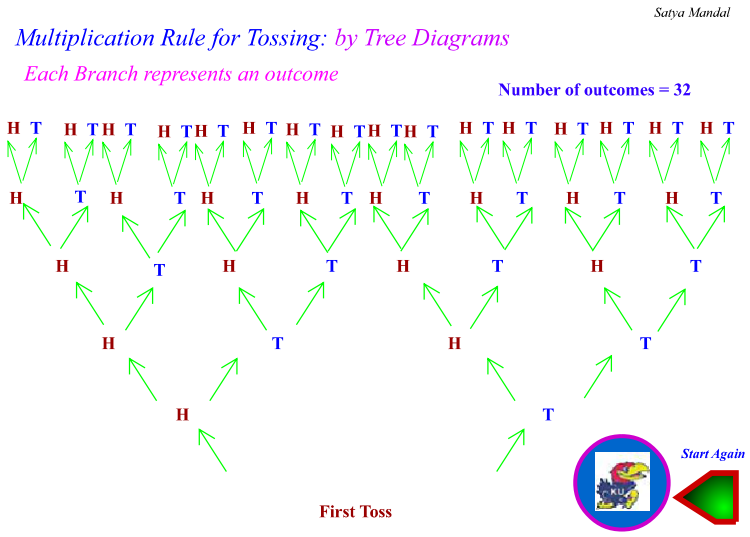
5. **[Exp.r]** corresponding to each outcome of each of the previous $r-1$ sub-experiments, the r^{th} sub-experiment has $\mathbf{n_r}$ outcomes.

Then our original (combined) experiment will have

$$\mathbf{n_1 n_2 \cdots n_r} \qquad \text{outcomes.}$$

This Counting Principle is also referred to as the multiplication rule, for counting.

Remark. Here we have used the word "experiment" in a slightly different sense than the statistical experiments. The basic counting principle will be used to count the number of outcomes in sample spaces and events.



Examples. 3.4.1. Count the number of words of length five that can be constructed from the English alphabet.
Answer:$= 26 * 26 * 26 * 26 * 26$
To see this, use the counting principle by dividing this job of constructing a word of length five into five sub-jobs:

Stage	*Job to do*	*no. of Ways*
1.	*Pick the* 1^{st} *letter*	26
2.	*Pick the* 2^{nd} *letter*	26
3.	*Pick the* 3^{rd} *letter*	26
4.	*Pick the* 4^{th} *letter*	26
5.	*Pick the* 5^{th} *letter*	26
Answer $=$	*Product* $=$	$11,881,376$ *words*

3.4.2. Count the number of ways you can assign the 11 seats in the first row in a concert hall to 231 guests. The job of assigning 11 seats can be divided into 11 jobs of assigning

each 11 seats. We use the product rule, as follows:

Stage	*Job to do*	*no. of Ways*
1.	*Assign seat* 1	231
2.	*Assign seat* 2	230
3.	*Assign seat* 3	229
4.	*Assign seat* 4	228
5.	*Assign seat* 5	227
6.	*Assign seat* 6	226
7.	*Assign seat* 7	225
8.	*Assign seat* 8	224
9.	*Assign seat* 9	223
10.	*Assign seat* 10	222
11.	*Assign seat* 11	221
Answer =	*Product* =	$221 * 222 * \ldots * 230 * 231$

Example 3.4.3. Contrast: How many ways can you form a committee of 11 members from a group of 231 people? Unlike assigning seats, here the order of selection of the members will be ignored. The 11 members, when permuted around, will have different seat assignments but in the same committee. Forming the committee is a "combination" problem that comes below.

Remark. The difference between assigning 11 seats in a row and forming a committee of 11 is that in the first case the **order of assignment** is important. Assigning the first row to the same 11 guests in two different ways will count as two different outcomes. When we form a committee, the order in which we pick 11 members does not make any difference.

Definition. Suppose we have n objects. We pick r of them one by one (without ever putting them back) and arrange them in a row. Such an ordered arrangement will be called a **permutation** of n objects taken r at a time. The number of permutations of n objects taken r at a time is denoted by ${}_nP_r$. It follows from the basic counting principle that

$$ {}_nP_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} $$

So, the number of such permutations ${}_nP_r$ = product of r integers starting from n **downward**.

Definition. In contrast, we can pick r objects from a collection of n objects one by one but place the object back in the collection before the next pick, and arrange all of them in a row. Such selection and arrangement is called selection **with replacement**. Constructing a formal word of length 5 is an experiment of picking **with replacement**. By the same token, permutations are selection **without replacement**.

Definition. Suppose we have n objects in a container. We pick r of them all at a time. In this case the order of selection does not come into consideration. Such a selection is called

a **combination** of n objects taken r at a time. The number of combinations of n objects taken r at a time is denoted by ${}_nC_r$ and is given by

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

Examples.

1. Count the number of ways you can form
   a committee of 11 from a group of 231 people.
   Answer:$=_{231} C_{11}$

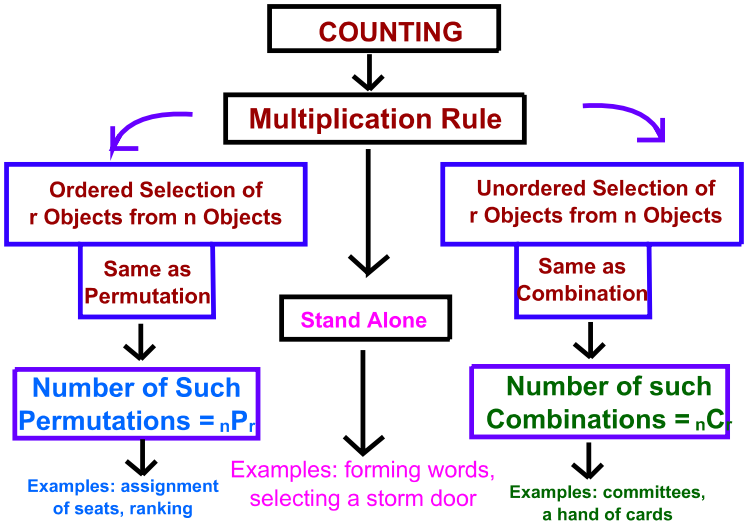
2. Count the number of ways you can deal
   a hand of 13 cards from a deck of 52 cards.
   Answer:$=_{52} C_{13}$

Remarks:

1. Note that n-factorial $n!$ grows very fast. In fact, $n!$ is beggar than 2^{n-1}. Like 2^{n-1}, the computers crashes soon if you try to run a program to commute n!. If and when you take a course on programming, one of first program you will be asked to write would be the computation of $n!$.

2. Likewise, ${}_nP_r$, ${}_nC_r$ can also be very large.

Following chart illustrates different classification of counting:



Use of Calculators (TI-84): To compute ${}_nP_r$, ${}_nC_r$, using TI-84, do the following:

1. To compute 4! do the following:
   a) type in 4, b) hit the MATH key and scroll to right to PRB, c) scroll down to !, d) ENTER .

2. To compute ${}_{13}P_4$ do the following:
   a) type in 13, b) hit the MATH key and scroll to right to PRB,
   c) scroll down to ${}_nP_r$, d) type in 4 and ENTER .

3. To compute ${}_{13}C_4$ do the following:
   a) type in 13, b) hit the MATH key and scroll to right to PRB,
   c) scroll down to ${}_nC_r$, d) type in 4 and ENTER .

3.4.1 Problems: Counting Techniques and Probability

Exercise 3.4.1. Find 5! **Solution:** $5! = 1 * 2 * 3 * 4 * 5 = 120$

Exercise 3.4.2. A homeowner would like to install a new storm door. The local store offers 2 brand names; each brand has 4 different styles and 3 colors. How many choices does the homeowner have? **Solution:** Use of multiplication rule:

Stage	*To select*	*no. of Ways*
1.	*Pick brand name*	2
2.	*Pick the styles*	4
3.	*Pick the colors*	3
Answer =	*Product* =	24

Exercise 3.4.3. Suppose in the World Cup soccer tournament, group A has 8 teams. Each team of group A has to play all the other teams in the group.
How many games will be played among the group A teams.
Solution: **Answer:**$=_8 C_2 = 28$

Exercise 3.4.4. Suppose there are 14 tennis players are competing is a tournament for the top three positions. How many outcomes are possible?
Solution: Solution: Since, order matters here, this is an ordered selection of 3 from 14. So, **Answer:**$_{14}P_3 = 2184$

Exercise 3.4.5. Suppose there are 14 applications for three positions in the college office. The positions are all at the same level and pay. How many selection is possible?
Solution: Solution: Since all positions are alike, this is an unordered selection of 3 from 14. So, So, **Answer:**$_{14}C_3 = 364$

Exercise 3.4.6. Psychology department has funds for awards, of cash value $2000 each, for best 5 teachers. There are 44 teachers in the math department. How many selection of five winners is possible?
Solution: Since, order of selection is irrelevent, this is an unordered selection of 5 from 44. So, **Answer:**$=_{44} C_5 = 1086008$

Exercise 3.4.7. Nursing school has funds for awards for the best four teachers this year. Awards have cash values of $5,000, $3000, $2000 and $1000. There are 37 teachers in the math department. How many selection of four winners is possible?
Solution: Since, order matters here, this is an ordered selection of 4 from 37. So, **Answer:**$_{37}P_4 = 1585080$

Exercise 3.4.8. How many ways can you deal a hand of 13 cards from a deck of 52 cards?
Solution: **Answer:**$=_{52} C_{13}$.
This would be too large, to simplify!
Think about this number, in the context of card games.

Exercise 3.4.9. How many ways can you deal a hand of 4 spades, 3 hearts, 3 diamonds, and 3 clubs? **Solution:** Use of multiplication rule:

Stage	*To select*	*no. of Ways*
1.	*Pick 4 spades*	${}_{13}C_4 = 715$
2.	*Pick 3 hearts*	${}_{13}C_3 = 286$
3.	*Pick 3 diamonds*	${}_{13}C_3 = 286$
4.	*Pick 3 clubs*	${}_{13}C_3 = 286$
Answer =	*Product* =	$715 * (286)^3$

It is too large to simplify.

Exercise 3.4.10. We have 13 students in a class. How many ways can we assign the 4 seats in the first row? **Solution:** This is a ordered assignment of 4 seas to 13.
Answer:$={}_{13}\ P_4 = 17160$.

Alternately, we could use the multiplication rule, and assign the 4 seats, one by one: $13 * 12 * 11 * 10$ ways.

Exercise 3.4.11. Programming languages sometimes use a hexadecimal system (also called "hex") of numbers. In this system, 16 digits are used and denoted by 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F. Suppose you form a 6-digit number in a hexadecimal system.

1. What is the probability that the number will start with a letter digit?
2. What is the probability that the number is divisible by 16 (i.e., ends with 0)?

Solution: Here the sample space is the collection of all the 6-digit hex numbers.
Using the counting principle, the number of hex $= n(S) = (16)^6$.

1. Let E be the event that the number starts with a letter digit.
   Again, by the counting principle, the number of hex in $E = n(E) = 6 * (16)^5$.
   So, $P(E) = \frac{n(E)}{n(S)} = \frac{6*(16)^5}{(16)^6} = \frac{6}{16}$.
2. Let F be the event that the number is divisible by 16.
   Since a number is divisible by 16 means, in hex, the first digit is 0.
   So, the number of hex in $F = n(F) = (16^5) * 1 = 16^5$.
   So, $P(F) = \frac{n(F)}{n(S)} = \frac{16^5}{16^6} = \frac{1}{16}$.

Exercise 3.4.12. You are playing Bridge and you are dealt a hand of 13 cards.

1. What is the probability that you will get a hand of 4 spades, 3 hearts, 3 diamonds and 3 clubs?
2. What is the probability that you will get all 4 aces?
3. What is the probability that you will get all 13 spades?

Solution: Here the sample space S is the set of all possible hands of 13 cards, out of a deck of 52 cards. So, $n(S) =_{52} C_{13}$.

1. Let E be the event that you get a hand of
   4 spades, 3 hearts, 3 diamonds and 3 clubs. In Ex. 3.4.9, we computed
   $n(E) = 715 * (286)^3$. So,
   $P(E) = \frac{n(E)}{n(S)} = \frac{715*(286)^3}{{}_{52}C_{13}}$.
2. Let F be the event that you get all 4 aces (and 9 other card).
   So, $n(F) = ({}_4C_4) * ({}_9C_{39}) =_9 C_{39}$. So,
   $P(F) = \frac{n(F)}{n(S)} = \frac{{}_9C_{39}}{{}_{52}C_{13}}$.
3. Let G be the event that you get all 13 spades
   Then, $n(G) =_{13} C_{13} = 1$. So,
   $P(G) = \frac{n(G)}{n(S)} = \frac{1}{{}_{52}C_{13}}$.

These numbers are too large, to simplify.

Exercise 3.4.13. A committee of 9 is selected at random from a group of 11 students, 17 mothers and 13 fathers.

1. What is the probability that the committee has 3 students, 3 mothers, and 3 fathers, i. e., is a balanced committee?
2. What is the probability that the committee has 4 mothers and 5 fathers?
3. What is the probability that the committee has all students?

Solution: Here total number of people in $n = 11 + 17 + 13 = 41$.
The sample space S is that set of all possible committees 9 out of these 41.
Forming committees are unordered selection.
So, $n(S) =_{41} C_9 = 350,343,565$

1. Let E be the event that the committee consists of 3 students, 3 mothers, and 3 fathers.

   To, count $n(E)$, we make a table for multiplicative principle:

Stage	*Job to do*	*no. of Ways*
1.	*Select* 3 *students from* 11	${}_{11}C_3 = 165$
2.	*Select* 3 *mothers from* 17	${}_{17}C_3 = 680$
3.	*Select* 3 *fathers from* 13	${}_{13}C_3 = 286$
$n(E) =$	*Product* $=$	32089200

   So, $P(E) = \frac{n(E)}{n(S)} = \frac{32089200}{350,343,565}$

2. F be the event that the committee has students only. Then,
   $n(F) = 11C_9 = 55$. So,
   $P(F) = \frac{n(F)}{n(S)} = \frac{55}{350,343,565}$

3.5 Conditional Probability and Independent Events

Sometimes when new information becomes available, the probability of an event may have to be reevaluated in light of this new information. Suppose we have a sample space S and an event E. Now suppose we have new information that an event C has occurred. We will have to reevaluate the conditional probability of E, given the knowledge that C has occurred. The conditional probability of E given that C has occurred, is denoted by $P(E|C)$. Clearly, $P(E|C)$ may be different from $P(E)$. In fact, now that C has occurred, our old sample space is no longer relevant. And C assumes the role of the new sample space. We give the following definition and formulas.

Definition. Let S be a sample space and E, C be two events.

1. The conditional probability of E given that C has occurred is

$$P(E|C) = \frac{P(E \cap C)}{P(C)} \qquad \text{if} \quad P(C) \neq 0.$$

2. , So, we get the following formula

$$P(E \cap C) = P(E|C)P(C). \tag{3.6}$$

3.5.1 Independent Events

If the conditional probability $P(E|F) = P(E)$ the "simple" probability, then we say that E and F are independent. In this case,

$$P(E \cap F) = P(E)P(F).$$

Definition. Two events E and F are defined to be **independent** if

$$P(E \cap F) = P(E)P(F).$$

If two events are not independent, then they are said to be **dependent**.

Remark. Let us also describe what we mean by independence of 3 or more events. For events $E_1, E_2, \ldots, E_n$, we say they are independent if the "multiplication rule" applies for any number of them. For example, four events E, F, G, H are defined to be **independent** if all of the following holds:

1. With two events:

$$\begin{array}{ll} P(E \cap F) = P(E)P(F), & P(E \cap G) = P(E)P(G), \\ P(E \cap H) = P(E)P(H), & P(F \cap G) = P(F)P(G) \\ P(F \cap H) = P(F)P(H), & P(G \cap H) = P(G)P(H) \end{array}$$

2. With three events:

$$\begin{array}{ll} P(E \cap F \cap G) = P(E)P(F)P(G), & P(E \cap F \cap H) = P(E)P(F)P(H), \\ P(E \cap G \cap H) = P(E)P(G)P(H), & P(F \cap G \cap H) = P(F)P(G)P(H) \end{array}$$

3. With all four events:

$$P(E \cap F \cap G \cap H) = P(E)P(F)P(G)P(H)$$

Example. (Justification). Suppose we pick a KU student at random and let E be the event that the student is taller than 6 feet.
We have the following:

1. The sample space S is the whole KU student population.

2. Since all the outcomes are equally likely, we have

$$P(E) = \frac{number\ of\ KU\ students\ who\ are\ taller\ than\ 6\ feet}{Total\ number\ of\ KU\ students} = \frac{n(E)}{n(S)}$$

3. Now suppose we know that the student selected is a male. Let us denote the event that the student is a male by C. The probability that the student is taller than 6 feet, given that the student is a male, is higher than "simple" $P(E)$. In fact, our new sample space is C, which is the whole KU male student population, not S, which is the whole KU student population.

4. We now have the probability that the student is taller than 6 feet in height given that the student is a male

$$P(E|C) = \frac{number\ of\ MALE\ students\ who\ are\ taller\ than\ 6\ feet}{Total\ number\ of\ male\ KU\ students} = \frac{n(E \cap C)}{n(C)}$$

$$= \frac{n(E \cap C)/n(S)}{n(C)/n(S)} = \frac{P(E \cap C)}{P(C)}$$

3.5.2 Problems: Conditional Probability and Independent Events

Exercise 3.5.1. Let A, B be two events. It is given

$$P(A) = .66, \quad P(A \cap B) = .11$$

Compute $P(B|A)$ **Solution:** $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.11}{.66} = \frac{1}{6}$

Exercise 3.5.2. It is given

$$P(A|B) = .8, \qquad P(B) = .1$$

Find $P(A \cap B)$.

Solution: By Formula (3.6):

$$P(A \cap B) = P(A|B)P(B) = .8 * .1 = .08$$

Exercise 3.5.3. In a certain county, the probability that a person took a flu shot is .45 and the probability that a person will get flu, given that he/she took a flu shot is .06. What is the probability that a randomly selected person took a flu shot and will get flu?

Solution: Let E be the event that the person will get a flu. Let C be the event that the person took a flu shot. We have

$$P(C) = .45, \qquad P(E|C) = .06$$

We need to find $P(C \cap E)$. By Formula (3.6):

$$P(C \cap E) = P(E|C)P(C) = .06 * .45$$

Exercise 3.5.4. Following are data (*unreal*) from a hospital emergency room:

1. The probability that a patient in the emergency room will have health insurence is 0.75.
2. The probability that a patient in the emergency room will survive the treatment 0.85.
3. The probability that a patient in the emergency room will have health insurance and will also survive is 0.7.

What is the conditional probability that a patient in the emergency room will survive, given that he/she has health insurance.

Solution: Let H = event that that patient has health insurance and
S = event that the patient will survive.
Given

$$P(H) = .75, \quad P(S) = .85, \quad P(H \text{ and } S) = .7$$

So,

$$P(S|H) = \frac{P(S \text{ and } H)}{P(H)} = \frac{.7}{.75} = .9333$$

Exercise 3.5.5. Following are some statistics about pneumonia:

1. Probability that an individual will get a pneumonia vaccine shot is .58.
2. Probability that an individual will get a Pneumonia shot and will get pneumonia in winter is .04.
3. Probability that an individual will get pneumonia is .13.

What is the conditional probability that a randomly selected person will get pneumonia given that he/she took pneumonia shot?

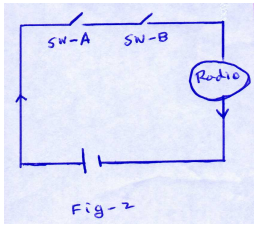
Solution: E =event that a randomly selected person ill get a pneumonia vaccine shot, F =event that a randomly selected person will get a pneumonia We are given

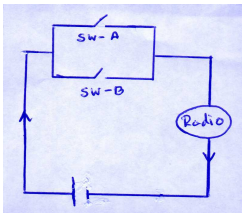
$$P(E) = .58, P(E \text{ and } F) = .04, \quad P(F) = .13$$

We have

$$P(F|E) = \frac{P(F \text{ and } E)}{P(E)} = \frac{.04}{.58} = .0690$$

Exercise 3.5.6. Consider the following two circuit diagrams:





For each of the two circuits do the following: As you can see, current flows through two switches A and B to the radio and back to the battery. It is given that the probability that the switch A is closed is 0.91 and the probability that the switch B is closed is 0.83. Assume that the two switches function independently. Find the probability that the radio is playing. Solution

Solution: E =the event that Switch A is on (closed) and F= the event that Switch B is on (closed). We are given

$$P(E) = .91, P(F) = .83$$

1. Consider the first circuit: In this case:

$$P(radio\ is\ playing) = P(E\ and\ F) = P(E)P(F) = .91 * .83 = .7553$$

2. Consider the second circuit: In this case,

$$P(radio\ is\ playing) = P(E\ or\ F) = P(E) + P(F) - P(E\ and\ F)$$

$$= P(E) + P(F) - P(E)P(F) = .91 + .83 - .91 * .83 = .9847$$

Exercise 3.5.7. An airplane has two engines. The probability that engine 1 fails is 0.023 and the probability that engine 2 fails is 0.06. Assume that the engines function independently.

1. What is the probability that both engines fail?
2. What is the probability that at most one will fail?
3. What is the probability that neither will fail?

Solution:

1. Let E = event that the engine 1 fails and F = event that the engine 2 fails.
   Then, the event that both fail is $(E \; and \; F)$.
   We have $P(E) = .023$ and $P(F) = .06$
   Assuming independence,
   $P(E \; and \; F) = P(E)P(F) = .023 * .06 = .00138$
2. $P(at \; most \; one \; will \; fail) = P(not(E \; and \; F))$
   $= 1 - P(E \; and \; F) = 1 - .00138 = .99862$
3. $P(neither \; fail) = P[(not \; E) \; and \; (notF)]$
   $= P(not \; E)P(not \; F) = (1 - .023)(1 - .06) = .91838$

Exercise 3.5.8. The probability that you will receive a wrong number call this week is 0.3; the probability that you will receive a sales call this week is 0.8; and that the probability that you will receive a survey call this week is 0.5. What is the probability that you will receive one of each this week? (Assume that all these calls are independent.)

Solution: Let E = event that you will receive a wrong number call this week,
F =event that you will receive a sells number call this week,
G = event that you will receive a survey number call this week.

Then, $(E \; and \; F \; and \; G)$ = the event that you will one of each this week.
We have $P(E) = .3, \quad P(F) = 0.8, \quad P(G) = 0.5.$
So, $P(E \; and \; F \; and \; G) = P(E)P(F)P(G) = .3 * .8 * .5 = .12$

Exercise 3.5.9. Suppose you went for a job interview in Lawrence and another one in Kansas City. Probability of that you will get the job in Lawrence is .25 and the probability of that you will get the job in Kansas City is .33. It is reasonable to assume independence.

1. What is the probability that you will get both the jobs?
2. What is the probability that you will get neither?

Solution: Let E = event that you will get the job in Lawrence and
F = event that you will get the job in Kansas City.

1. Then, the event that you will get both the jobs is $(E\ and\ F)$.
   We have $P(E) = .25, \quad P(F) = .33$

   Assuming independence,
   $P(E\ and\ F) = P(E)P(F) = .25 * .33 = .0825$
2. The event that you will get neither is $((not\ E) and (not\ F))$.
   $P(not\ E) = 1 - P(E) = 1 - .25 = 75$, and $P(not\ F) = 1 - .33 = .67$
   $P(Neither) = P((not\ E)\ and\ (not\ F)) = P(not\ E)P(not\ F) = .75 * .67 = .5025$

Exercise 3.5.10. You are taking the Elementary Statistics course in KU and your brother is taking the same course in MU. The probability that you will get an A is .18 and the probability that your brother will get an A is .21.

1. What is the probability that both of you will get an A.
2. What is the probability that none of you will get an A.

Solution: Let E be the event that you will get an A and
F be the event that your brother will get an A.

1. Then, the event that both of you will get A is $(E\ and\ F)$.
   We have $P(E) = .18, \quad P(F) = .21.$

   Assuming independence,
   $P(Both) = P(E\ and\ F) = P(E) * P(F) = .18 * .21 = .0378.$
2. The event that neither will get A is $((not\ E)\ and\ (not\ F))$.
   $P(notE) = 1 - .18 = .82$ and $P(not\ F) = 1 - .21 = .79.$
   $P(Neither) = P((not\ E)\ and\ (not\ F)) = P((not\ E) * P(not\ F)) = .82 * 79 = .6478$

Exercise 3.5.11. Probability that you will receive a call from a sibling this week is .35 and the that you will receive a call from a parent this week is .43.(Assume independence.)

1. What is the probability that you receive a call from both, this week.

2. What is the probability that you receive a call from neither, this week.

Solution: Let E be the event that you will receive a call a sibling this week and F be the event that you will receive a call a parent.

1. Then, the event that you will receive a call from both is $(E \; and \; F)$.
   We have $P(E) = .35$ and $P(F) = .43$.

   Assuming independence, $P(Both) = P(E \; and \; F) = P(E) * P(F) = .35 * .43 = .1505$.

2. The event that neither will call you is $((not \; E) \; and \; (not \; F))$.
   $P(not \; E) = 1 - .35 = .65$ and $P(not \; F) = 1 - .43 = .57$
   $P(Neither) = P((not \; E) \; and \; (not \; F)) = P(not \; E) * P(not \; F)) = .65 * .57 = .3705$

Exercise 3.5.12. Probability that it will rain in Lawrence today is .22 and probability that it will rain today at your home town is .40. (Assume independence.)

1. What is the probability that it will rain in both places?

2. What is the probability that it will neither rain in Lawrence nor in your hometown?

Solution: Let E be the event that it will rain in Lawrence today and F be the event that it will rain at your home town today.

1. Then, the event that it will rain in both places is $(E \; and \; F)$.
   We have $P(E) = .22, \quad P(F) = .40$.

   Assuming independence,
   $P(Both) = P(E \; and \; F) = P(E) * P(F) = .22 * .40 = .088$.

2. The event that it will neither rain in Lawrence nor in your hometown is $((not \; E) \; and \; (not \; F))$.
   $P(not \; E) = 1 - .22 = .78$ and $P(not \; F) = 1 - .40 = .60$
   $P(Neither) = P((not \; E) \; and \; (not \; F)) = P(not \; E) * P(not \; F)) = .78 * .60 = .468$

Exercise 3.5.13. According to the poll, probability that a person would vote for Candidate A is .43.

1. What is the probability that both you and I would vote for Candidate A?
   (We can assume independence because you and I do not influence each other.)

2. What is the probability that neither you nor I would vote for Candidate- A?

Solution: Let E be the event that you will vote for Candidate A and
let F be the event that I will vote for Candidate A.

1. Then, the event that it will rain in both places is $(E \; and \; F)$.
   We have $P(E) = P(F) = .43$.

   Assuming independence,
   $P(Both) = P(E \; and \; F) = P(E) * P(F) = .43 * .43 = .1849$

2. The event that neither you nor I will vote for Candidate A is $((not \; E) \; and \; (not \; F))$.
   $P(not \; E) = 1 - .43 = .57$ and $P(not \; F) = 1 - .43 = .57$.
   $P(Neither) = P((not \; E) \; and \; (not \; F))$
   $= P(not \; E) * P(not \; F)) = .57 * .57 = .3249$

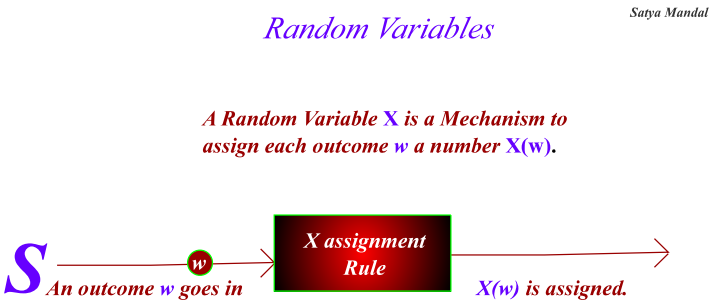
Chapter 4

Random Variables

4.1 Random Variables

A random variable is a complete description of a numerical characteristic of the population. Among the examples would be (1) Weight of the fish population in a lake, (2) Number of typos in the textbooks used in KU, (3) GPA of the student population, (4) Delay in departure time of all the commercial flights originating from US. The following is a formal definition.

Definition. A **random variable** is a rule or a formula or a mechanism that associates a numerical value to each member (outcome) of the the sample space S. So, given a member (outcome) w of S, a variable X assigns a numerical value $X(w)$ to w. For us, $X(w)$ will be a characteristic (like height, weight, time, salary) of the population.



Example. Suppose the KU student population is under study. Therefore, the sample space S is the whole collection of KU students population. A KU student is a sample unit. To this sample space, we can associate multiple random variables, as follows:

1. Let G = the GPA of the students . Then G is a random variable. The GPA is the "characteristic" that G describes. So, given a student, G has a value. For example:

$$G(\text{Donald Smith}) = 3.25, \quad G(\text{Sam Donaldson}) = 3.11,$$
$$G(\text{Karen Currie}) = 3.89, \quad G(\text{King Who}) = 2.13$$

2. Define Y as follows:

$$Y(w) = 0 \quad \text{If } w \text{ is Male}$$
$$Y(w) = 1 \quad \text{If } w \text{ is Female}$$

   Then, Y is a random variable. We have, for example,

$$G(\text{Donald Smith}) = 0, \quad G(\text{Sam Donaldson}) = 0,$$
$$G(\text{Karen Currie}) = 1, \quad G(\text{King Who}) = 0$$

3. Let Z = the total expenses of the students this year. Then Z is a random variable.

4. Let H = height of the students. The H is a random variable. Here, H describes the HEIGHT-characteristic of the sample space.
5. Let W = the weight of the students. Then W is a random variable.
6. Let C = the number of course credit hours completed by the student. Then C is a random variable.
7. Let T = Tuition paid by the students this year. Then T is a random variable.
8. Similarly, given any other characteristic like annual income, distance from KU to the students' residence, a random variable can be defined for this sample space or population.

Definition. Random variables are classified to two different types: (1) the continuous random variables and (2) discrete random variables.

1. If a random variable X can assume any numerical value over an interval, then it would be called a **continuous random variable**. The random variables H and W are continuous random variable.
2. A random variable X is said to be a discrete random variable, if the possible values of X the variable can be written in a (finite or infinite) list:

$$x_1, x_2, x_3, \ldots$$

In other words, a discrete random variable assumes only finitely many or countably infinitely many distinct values. In the above example, G, Y, C, T are discrete random variables.

Examples of Continuous and Discrete Variables

1. The examples of continuous random variables are weight, length, volume, area, and time.
2. For this course, most of the examples of discrete random variables would be the number of something?number of typos, number of road accidents, number of phone calls.

You may not need anything more than these examples of continuous and discrete random variables.

Two Examples.

1. Let X be the number of wrong number calls you receive in a day. Then X is a discrete random variable.
2. Let X be the waiting time before you receive the next wrong number call. Then X is a continuous random variable.

4.2 Probability Distribution

A Random variable X represents a numerical feature (say weight) of the whole sample space (or the population). As far as statistics is concerned, the distribution of X (i.e. distribution of the values of X) is unknown. Goal of this course is to model (or make realistic hypotheses) and estimate the actual distribution of X.

The **probability distribution** of a random variable X is a table or a rule or a method that answers probability-related questions regarding X. We would first be concerned with the discrete random variables.

Definition. Suppose X is a discrete random variable that assumes the values

$$x_1, x_2, x_3, \cdots$$

The **probability distribution** of X can be described by giving

$$p(x_i) = P(X = x_i) \qquad \text{for} \quad i = 1, 2, 3, \ldots$$

in a table or by a formula. This function $p(xi_)$ is called the **probability function** of X.

When the probability distribution of X is given in a table, it would look like the following:

Value X	x_1	x_2	x_3	$\cdots$
$p(x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\cdots$

Properties of Probability function. Suppose X is a discrete random variable that assumes values

$$x_1, x_2, x_3, \cdots$$

and let $p(x)$ be the probability function. Then we have the following:

$$0 \leq p(x_i) \leq 1, \qquad \text{and} \qquad \sum p(x_i) = 1.$$

We introduce few more important definitions. **Definitions.** Let X be a discrete random variable that assumes the values

$$x_1, x_2, x_3, \cdots$$

and let $p(x)$ be the probability function.

1. The **mean** μ of X is defined as

$$\mu = \sum x_i p(x_i)$$

The mean μ of X is also called the **expectation** $E(X)$ of X.

2. The **variance** σ^2 of X is defined as

$$\sigma^2 = Var(X) = \sum (x_i - \mu)^2 p(x_i) = \sum x_i^2 p(x_i) - \mu^2$$

3. The **standard deviation** σ of X is defined as the positive square root of the variance of X. So,

$$\textbf{St. Dev. of } X = \sigma = \sqrt{Variance(X)}$$

Recall, in chapter 2, we defined mean $\overline{x}$ (2.1) and variance s^2 (2.2) of data. Here the mean μ corresponds to the mean of the whole population, and the mean $\overline{x}$ would correspond to mean of some sample data from this X-population. Similarly, variance σ^2 and s^2 defers from, and are related, to each other. This is way:

1. the mean $\overline{x}$ would be referred to as **sample mean**,
   s^2 would be referred to as **sample variance**, and
   s would be referred to as **sample st. deviation**.
   These are **statistic**, as defined in section 1.2.2.

2. And the mean μ of X would be referred to as the **population mean**,
   σ^2 would be referred to as **population variance**, and
   σ would be referred to as **population st. deviation**.
   These are **parameters**, as defined in section 1.2.2.

Remark. As was mentioned above, a random variable X represents numerical characteristics of the population. Statistics has a place in life, only because the population is unknown and it needs to be modeled and estimated. In particular,

1. The distribution of the random variable X would be unknown.

2. The mean μ and standard deviation σ of

3. represent the population mean and the population standard deviation of the corresponding characteristic. If X represents the weight of the fish population in a lake, then the mean μ of

4. represents the mean weight of the fish population. Similarly, the standard deviation σ of X represents the standard deviation weight of the fish population.

5. The population parameters μ, σ are unknown. Goal of this course would be to estimate them, using the statistics $\overline{x}$, s which are computed from collected samples.

6. Under reasonable (Bell curve distribution) assumptions about the population, we would see in future, mean μ and standard deviation σ would determine the complete distribution of X.

Example. Suppose you design a coin toss game . In this game, you give the opponent \$3 if a head comes and you collect \$1 if a tail comes. Let X be the money you receive. Then X assumes the values -3 and 1. You also have a loaded coin so that

$$P(H) = \frac{1}{9}, \qquad P(T) = \frac{8}{9}.$$

Then the probability distribution of X is given by

Value X	-3	1
$p(x_i)$	$\frac{1}{9}$	$\frac{8}{9}$

So, the mean μ of X is given by

$$\mu = \sum x_i p(x_i) = (-3) * \frac{1}{9} + 1 * \frac{8}{9} = \frac{5}{9}$$

The variance

$$\sigma^2 = \sum x_i^2 p(x_i) - \mu^2 = \left((9) * \frac{1}{9} + 1 * \frac{8}{9}\right) - \left(\frac{5}{9}\right)^2 = 1.5802$$

The standard deviation is given by

$$\sigma = \sqrt{Variance(X)} = \sqrt{1.5802} = 1.2571.$$

Interpretation of mean μ of X:

1. In this example, the mean μ tells us your average win per game if you play for a long time, which is $\frac{5}{9}$dollars per game. Similarly, the standard deviation $\sigma = 1.2571$ dollars is a measure of variability (or uncertainty) of your win per game.

2. Similarly, if Z represents the height of the KU student population, then the mean $\mu = E(Z)$ is the actual mean height of the KU student population. If we take a large sample from the KU student population and compute the sample mean, it should approximate μ.

4.2.1 Problems: on Probability Distribution

Exercise 4.2.1. The number of passengers X in a car on a freeway has the following probability distribution.

$X = x$	1	2	3	4	5
$p(x)$	0.35	0.30	0.15	0.15	0.05

1. Find the expected number of passengers in a car;

2. Find the Variance σ^2 of the number of passengers;

3. Find the probability that the number of passengers in a car is at least 3.

Solution:

1. The mean μ of X is given by

$$\mu = \sum x_i p(x_i) = 1 * 0.35 + 2 * 0.30 + 3 * 0.15 + 4 * 0.15 + 5 * 0.05 = 2.25$$

The variance

$$\sigma^2 = \sum x_i^2 p(x_i) - \mu^2 = (1*0.35+4*0.30+9*0.15+16*0.15+25*0.05) - (2.25)^2 = 1.4875$$

The standard deviation is given by

$$\sigma = \sqrt{Variance(X)} = \sqrt{.0475} = 1.2196$$

Exercise 4.2.2. Karin is a plumber who works for 3 different employers. Employer A pays her \$120 a day, employer B pays her \$70 dollars a day, and employer C pays her \$180 a day. She works for whoever calls her first. The probability that employer A calls her first is 0.30; the probability that employer B calls first is .20; and the probability that employer C calls her first is 0.40 (the probability that no one calls is .10). What is the expected income and variance of Karin per day?

Solution: Suppose X =Karin daily earnings.
Some days $X = 0$, the days she has no work. $P(X = 0) = 1 - (0.30 + 0.20 + 0.40) = .10$
So, the probability distribution of X is

$X = x$	120	70	180	0
$p(x)$	0.30	0.20	0.40	0.10

1. The mean μ of X is given by

$$\mu = \sum x_i p(x_i) = 120 * .30 + 70 * .20 + 180 * .40 + 0 * .10 = 122$$

The variance

$$\sigma^2 = \sum x_i^2 p(x_i) - \mu^2 = (120^2 * .30 + 70^2 * .20 + 180^2 * .40 + 0^2 * .10) - (122)^2 = 3376$$

The standard deviation is given by

$$\sigma = \sqrt{Variance(X)} = \sqrt{3376} = 58.1034$$

Exercise 4.2.3. An insurance company sells a flight insurance policy at a flat rate of \$500 per flight. If a policyholder dies in flight, the insurance company pays \$100,000 to the survivors. The probability that a policyholder will die in flight is .003. What is the expected gain and variance of the company per sale?

Solution: Let X =gain per policy sell

So, the probability distribution of X is

$X = x$	500	-99500
$p(x)$	$1 - .003$	0.003

1. The mean μ of X is given by

$$\mu = \sum x_i p(x_i) = 500 * (1 - .003) - 99500 * .003 = 200$$

The variance

$$\sigma^2 = \sum x_i^2 p(x_i) - \mu^2 = 500^2 * (1 - .003) + (-99500^2) * .003$$

Exercise 4.2.4. The following table gives the proportion of credit hours that earned grades F, D, C, B and A in KU:

Grade	*A*	*B*	*C*	*D*	*F*
Proportion	.15	.35	.30	.15	.05

Let X represent the points earned for grades A,B,C,D and F. Write down the probability distribution of X and compute the mean (or the ex-pected value $E(X)$ and the standard deviation. **Solution:** We have $X = 0, 1, 2, 3, 4$ respectively, when the grades are F, D, C, B, A. Therefore, the distribution of X is given by

$X = x$	4	3	2	1	0
$p(x)$	.15	.35	.30	.15	.05

1. The mean μ of X is given by

$$\mu = \sum x_i p(x_i) = 0 * .05 + 1 * .15 + 2 * .30 + 3 * .35 + 4 * .15 = 2.4$$

The variance

$$\sigma^2 = \sum x_i^2 p(x_i) - \mu^2 = 0^2 * .05 + 1^2. * 15 + 2^2 * .30 + 3^2 * .35 + 4^2 * 2 - (2.4)^2 = 1.14$$

2. The standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.14} = 1.0677.$$

4.3 The Bernoulli and Binomial Experiments

There are many random variables that we encounter fairly often. The first one that we discuss is called a **Bernoulli random variable**.

Definition. There are many statistical experiments that have only two outcomes. In such cases, the outcomes may be called a success or a failure. So the sample space is

$$S = \{s, f\}$$

Here s means success and f means failure. Such an experiment is called a **Bernoulli trial**. Given a Bernoulli trial, we can define a random variable as

$$\begin{cases} X = 1 & \text{if success} \\ X = 0 & \text{if failure} \end{cases}$$

If the probability, $P(success) = p$, then we have $P(failure) = 1 - p$. So, the probability distribution of a Bernoulli random variable is given by

$X = x$	0	1
$p(x)$	$1 - p$	p

So, the mean μ, variance σ^2 and the st. deviation σ of X, is given by

$$\begin{cases} \mu = 0 * (1 - p) + 1 * p = p \\ \sigma^2 = (0^2 * (1 - p) + 1^2 * p) - \mu^2 = p - p^2 = p(1 - p) \\ \sigma = \sqrt{\sigma^2} = \sqrt{p(1 - p)} \end{cases}$$

It is interesting to note that the variance $\sigma^2 = p(1 - p) = P(success) * P(failure)$.

Examples. While it a simple random experiment, there are abundance of examples Bernoulli random experiments (or trial) would be. Here are some:

1. Testing an item (a lamp) in a factory for defectiveness. Here, if the item were defective, the outcome of the experiment would be called a "success".
2. A medical test, like a test for pregnancy. If the woman were pregnant, the outcome of the experiment would be called a success.
3. Tossing a coin. Head is a "success".

4.3.1 Binomial Random Variable

Bernoulli random variables (experiments) are too simple. But combining certain number of them leads to something very useful, as follows.

Definition. A combination of n "**identical and independent**" Bernoulli trials, is called a **Binomial experiment**. More formally, given a positive integer n and a number p with $0 \leq p \leq 1$ a $Binomial(n, p)$ experiment (or $B(n, p)$ experiment) is characterized as follows:

1. A binomial experiment consists of n identical and independent Bernoulli trials.

2. The probability of success in each trial remains fixed and is equal to p.

Definition. Given a $B(n,p)$-experiment, let

$$X = \text{total number of successes in these n trials.}$$

Then X is called a **binomial** $(n,p)-$**random** (or $B(n,p)$**-random**) variable.
Following are some important facts about a $B(n,p)$-random variable X. (We will not try to prove them in this course.)

1. X assumes values $0, 1, \ldots, n$.

2. The probability distribution is given by

$$p(r) = P(X = r) = P(r \ success) =_n C_r p^r (1-p)^{n-r} \quad \text{where} \quad r = 0, 1, 2, \ldots, n. \tag{4.1}$$

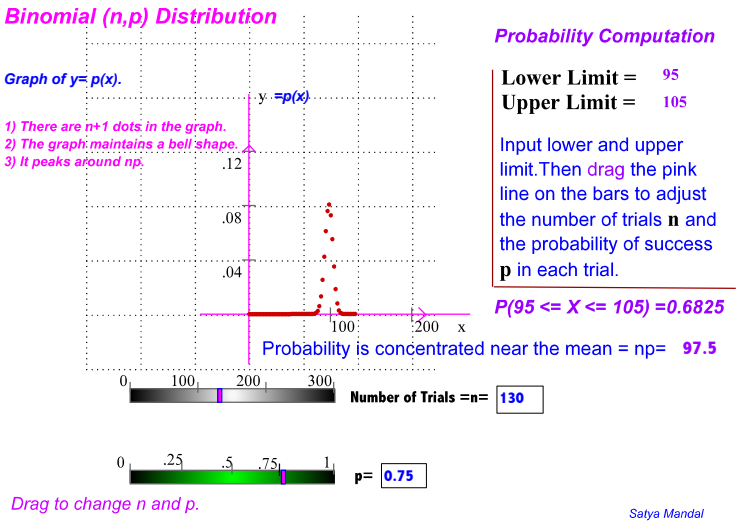
3. The mean μ, variance σ^2 and st. deviation σ, of X is given by

$$\begin{cases} \mu = E(X) = np \\ \sigma^2 = np(1-p) \\ \sigma = \sqrt{\sigma^2} = \sqrt{np(1-p)} \end{cases}$$

It is interesting to note, the mean

$$\mu = (\text{number of trial}) * (\text{probability of success in each trial}).$$

Following shows the graph of the probability function of a binomial random variable.



4.3.2 Use TI-84: to compute $B(n,p)$-probabality

We use TI-84, to compute probability, for Binomial variabels.
Suppose X is $B(12,.6)$ random variable.

1. he **binomialcdf**-function:

   (a) To compute $P(X \textit{ is at most } 8) = P(0 \leq X \leq 8)$ do the following:
   a) Press 2nd and then Distr (VARS)
   b) Scroll down to binomialcdf and ENTER
   c) type in "12, .6, 8)" and ENTER.
   TI will give the answer.
   In "binomilacdf" stands for cumulative density function.
   (It was not proper for them to use the word "density".)

   (b) The **binomialpdf**-function, can also be used.

i. To compute $P(X = 4)$:, do the following:
a) Press 2nd and then Distr (VARS)
b) Scroll down to binomialpdf and ENTER
c) type in "12, .6, 4)" and ENTER.
TI will give the answer.

ii. To compute $P(2 \leq X \leq 7)$, find
$P(X = 2)$, $P(X = 3)$, $P(X = 4)$, $P(X = 5)$, $P(X = 6)$, $P(X = 7)$
as above and then add.
Use at least four decimal points.

In "binomilapdf", stands for probability density function.
(It was not proper for them to use the word "density".) binomilacdf function in TI-84:

4.3.3 Problems: on Binomial Experiments

In a more mathematically demading class, the formula (4.1) would be used to compute probability. Some of you may try to do so. We use TI-84.

Exercise 4.3.1. Let X be a $B(6, .3)$-random variable. Find $P(X = 2)$. Also find the probability that X is at least 2.

Solution by TI-84.
Here $n = 6$, $p = .3$. We use TI-84:
$P(X = 2) = binomialpdf(6, .3, 2) = .324135$
Alternately. use formula (4.1)

$$P(X = 2) =_n C_r p^r (1-p)^{n-r} =_6 C_2 (.3)^2 (1-.3)^{6-2} = \frac{6!}{(6-2)!2!}(.3)^2(.7)^{6-2} = \frac{6*5}{2*1}(.3)^2(.7)^4$$

Exercise 4.3.2. According to a report entitled "Pediatric Nutrition Surveillance" published by Centers for Disease Control (CDC), 18 percent of children younger than 2 years had anemia in 1997. On a particular day, a pediatrician examined 11 children.

1. What is the probability that none will have anemia?
2. What is the probability that exactly 5 will have anemia?
3. What is the probability that all will have anemia?
4. Compute the expectation and variance of the number of children with anemia.
5. What is the probability that at least 7 will have anemia?

Solution by TI-84.

Here $n = 11$, $p = .18$.

X = Number of children with anemia.

X is $B(11, .18)$ random variable. We use TI-84 :

1. probability that none will have anemia
   $= P(X = 0) = binomialpdf(11, .18, 0) = .1127$

   Alternately. use formula (4.1)

$$P(X = 0) =_n C_r p^r (1-p)^{n-r} =_{11} C_0 (.18)^0 (1-.18)^{11-0} = \frac{11!}{0!11!}.82^{11}$$

2. probability that exactly 5 will have anemia
   $= P(X = 5) = binomialpdf(11, .18, 5) = .0265$

   Alternately. use formula (4.1)

$$P(X = 5) =_n C_r p^r (1-p)^{n-r} =_{11} C_5 (.18)^5 (1-.18)^{11-5} = \frac{11!}{(11-5)!5!}.18^5.82^7$$

3. probability that all will have anemia
   $= P(X = 11) = binomialpdf(11, .18, 11) = 0$ (approximately)

   Alternately. use formula (4.1)

$$P(X = 11) =_n C_r p^r (1-p)^{n-r} =_{11} C_{11} (.18)^{11} (1-.18)^{11-11} = \frac{11!}{(0)!11!}.18^{11} = .18^{11}$$

4. Expected number of children with anemia
   $= E(X) = \mu = np = 11 * .18 = 1.98$.
   $Variance(X) = \sigma^2 = np(1-p) = 11 * .18 * (1 - .18) = 1.6236$;

5. probability that at least 7 will have anemia $= P(X = 7 \text{ or more})$
   $= P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11)$
   $= binomialpdf(11, .18, 7) + binomialpdf(11, .18, 8) + binomialpdf(11, .18, 9) + binomialpdf(11, .18, 10) +$
   $binomialpdf(11, .18, 11)$

$=.0010213525$

Alternately, use the binomialcdf function.
probability that at least 7 will have anemia
$= P(X = 7 \text{ or more}) = 1 - P(0 \leq X \leq 6)$
$= 1 - binomialcdf(11, .18, 6) = 1 - .9989786475 = .0010213525$

Exercise 4.3.3. A gardener planted 15 seeds. The probability that a seed will germinate is 0.1.

1. What is the probability that exactly 3 seeds will germinate?
2. What is the probability that exactly 4 seeds will germinate?
3. What is the probability that exactly 9 seeds will germinate?
4. Compute the expected number of seeds that will germinate.
5. Compute the standard deviation of the number of seeds that will germinate.
6. What is the probability that at most 4 seeds will germinate?

Solution by TI-84.
Here $n = 15$, $p = .1$.
$X =$ Number of seed that will germinate.
X is $B(15, .1)$ random variable. We use TI-84 :

1. probability that exactly 3 seeds will germinate
   $= P(X = 3) = binomialpdf(15, .1, 3) = .1285$
2. exactly 4 seeds will germinate$= P(X = 4) = binomialpdf(15, .1, 4) = .0428$
3. exactly 9 seeds will germinate$= P(X = 9) = binomialpdf(15, .1, 9) = 2.6599 * 10^{-6}$.
4. Expected number of seeds that will germinate
   $= E(X) = \mu = np = 15 * .1 = 1.5$
   Standard deviation of $X = \sigma = \sqrt{np(1-p)} = \sqrt{15 * .1 * (1 - .1)} = 1.6119$
5. probability that at most 4 seeds will germinate $= P(X = 4 \text{ or less})$
   $= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$
   $= binomialpdf(15, .1, 0) + binomialpdf(15, .1, 1) + binomialpdf(15, .1, 2) + binomialpdf(15, .1, 3) + binomialpdf(15, .1, 4) = .9873$

Alternately, probability that at most 4 seeds will germinate
$= P(X = 4 \text{ or less}) = P(0 \leq X \leq 4) = binomialcdf(15, .1, 4) = .9873.$

Exercise 4.3.4. In a particular county, 60 percent of the population is Hispanic.

1. What is the probability that a jury of 12 will have exactly 6 Hispanic members?
2. What is the probability that a jury of 12 will have more than 6 Hispanic members?

Solution by TI-84.
Here $n = 12$, $p = .6$.
X =Number of Hispanic juries. X is $B(15, .6)$ random variable. We use TI-84 :

1. probability that will be exactly 6 Hispanic members
   $= P(X = 6) = binomialpdf(12, .6, 6) = .1766$

2. probability that a jury of 12 will have more than 6 Hispanic members
   $= P(7 \leq X \leq 12) = 1 - P(0 \leq X \leq 6) = 1 - P(0 \leq X \leq 6)$
   $= 1 - binomialcdf(12, .6, 6) = 1 - .3347914423 = .6652085577$

Exercise 4.3.5. From the hiring statistics of a corporation (say IBM), it is known that for every 4 interviews they give, they make 1 job offer. Suppose that the corporation interviews 8 candidates each time it comes to campus. What is the mean and standard deviation of the number of job offers made each time?

Solution:
Here $n = 8$, $p = .25$.
X = Number of job that will be offered each time.
So, X is a $B(8..25)$ random variable.

1. Expected number of jobs offered $= E(X) = \mu = np = 8 * .25 = 2$
2. Standard deviation of $X = \sigma = \sqrt{np(1-p)} = \sqrt{8 * .25 * (1 - .25)} = 1.2247$

Remark. In the some of the problems above, sometimes we had to add only less than 10 terms with the binomialpdf function. In a real life situation, one may have to add a large number of such terms. In those cases, it is better to use binomialcdf function. Following are some such problems.

Exercise 4.3.6. It is believed proportion of voters (in a county) who vote by absentee ballot is $p = .18$. You sample 725 voters.

1. Compute the mean μ and standard deviation σ of the number of absentee votes among these 725.
2. What is the probability that at least 160 in this sample will vote by absentee ballot?
3. Compute the probability that the number of absentee votes among these 725 would be at least 120 and at most 150.

Solution by TI-84.
Here $n = 725$, $p = .18$.
Let X = Number of absentee votes among this sample of 725.
So, X is a $B(725, .18)$ random variable. We use TI-84 :

1. Expected number of absentee votes $= E(X) = \mu = np = 725 * .18 = 130.5$
   Standard deviation of $X = \sigma = \sqrt{np(1-p)} == \sqrt{725 * .18 * (1 - .18)} = 10.3446.$
2. probability that at least 160 will vote by absentee ballot
   $= P(160 \leq X) = 1 - P(0 \leq X \leq 159)$
   $= 1 - binomialcdf(725, .18, 159) = 1 - .9969 = .0031.$
3. probability that the number of absentee votes would be at least 120 and at most 150
   $= P(120 \leq X \leq 150) = P(X \leq 150) - P(X \leq 119)$
   $= binomialcdf(725, .18, 150) - binomialcdf(725, .18, 119)$
   $= .9718 - .1435 = .8283.$

Exercise 4.3.7. About 27 percent of the population take flu shots. You are in a class of 750 students.

1. Compute the mean μ and standard deviation σ of the number students who took a flu shot.
2. compute the probability that at most 200 would have taken a flu shot.
3. Compute the probability that between 190 and 215 students would have taken a flu shot.

Solution by TI-84.

Here $n = 750$, $p = .27$.

Let X =Number of students in this class who took the flu shot.

So, X is a $B(750, .27)$ random variable. We use TI-84 :

1. Expected number $= E(X) = \mu = np = 750 * .27 = 202.5$
   Standard deviation of $X = \sigma = \sqrt{np(1-p)} = \sqrt{750 * .27 * (1 - .27)} = 12.1583$.

2. probability that at at most 200 would have taken a flu shot
   $= P(X \leq 200) = binomialcdf(750, .27, 200) = .4371$

3. probability that between 190 and 215 students would have taken a flu shot
   $= P(190 \leq X \leq 215) = P(X \leq 215) - P(X \leq 189)$
   $= binomialcdf(750, .27, 215) - binomialcdf(750, .27, 189)$
   $= .8573 - .1423 = .715$

Exercise 4.3.8. It is believed that 35 percent of the population in a county shop in health food market. If you sample 800 individuals, what is the probability that at least 400 would shop in health food market?

1. Compute the mean μ and standard deviation σ of the number those in this sample who shop in health food market.

2. compute the probability that at least 300 in this sample shop in health food market..

3. Compute the probability that between 270 and 315 shop in in health food market.

Solution by TI-84.

Here $n = 800$, $p = .35$.

Let X =Number of shop in health food market.

Then, X is a $B(750, .27)$ random variable. We use TI-84 :

1. Expected number $= E(X) = \mu = np = 800 * .35 = 280$
   Standard deviation of $X = \sigma = \sqrt{np(1-p)} = \sqrt{800 * .35 * (1 - .35)} = 13.4907$

2. probability that at least 300 in this sample shop in health food market
   $= P(300 \leq X) = 1 - P(X \leq 299)$
   $= 1 - binomialcdf(800, .35, 299) = 1 - .9253 = .0747$

3. probability that between 270 and 315 shop in in health food market
   $= P(270 \leq X \leq 315) = P(X \leq 315) - P(X \leq 269)$
   $= binomialcdf(800, .35, 315) - binomialcdf(800, .35, 269)$
   $= .9955 - .2186 = .7769$

Exercise 4.3.9. It is known that 78 percent of the microwave ovens last more that five years. A SQC inspector sampled 600 microwaves.

1. Compute the mean μ and standard deviation σ of the number of microwave ovens in this sample that would last more that five years.

2. Compute the probability that at least 470 in this sample microwave ovens would last more that five years.

3. Compute the probability that between 460 and 480 of microwaves in this sample would last more than five years.

Exercise 4.3.10. It is known that a vaccine may cause fever as side effect, after one takes the shot. The producer of the vaccine claims that only 11 percent of those who take the shot experience such side effects. You sample 1000 individuals who took the shot.

1. Compute the mean μ and standard deviation σ of the number of those this sample who would get a fever as side effect.

2. Compute the probability that at least 110 in this sample would get a fever as side effect.

3. Compute the probability that between 105 and 135 would get a fever as side effect.

Remark. When the number of trials n in a binomial experiment is too large, TI-84 (even computers) may fail. Foe example, if we try Ex.4.3.10, with 10,000 individuals, TI-84 $Binomial(10000, .11, 450)$ would give erratic answer. We will provide a remedy to this problem in chapter 5.

Chapter 5

Continuous Random Variables

5.1 Probability Density Function (pdf)

In chapter 4 a **continuous random variable** was defined as a random variable X that can assume any value on an interval. The probability distribution of a continuous random variable defined and behaves in a different manner than a discrete random variable, as follows.

Definition. Let X be a continuous random variable on a sample space S. The probability distribution of X is defined as follows:

1. There is a function $f(x)$, of real numbers x, associated to X. This function $f(x)$ would be called the probability density function, (abbreviated as pdf) of X. T he pdf $f(x)$ of X must satisfy the following two properties:

   (a) $f(x)$ is always non-negative (i.e $f(x) \geq 0$ for all x). Geometrically, the graph of the function $y = f(x)$ always lies on or above the x-axis.

   (b) The total area under the graph of $y = f(x)$ and above the x-axis is one.

2. For any two real numbers $a \leq b$ (also for $a = -\infty$ and $b = \infty$) the probability that X will be between a and b is given by the area under the graph of $y = f(x)$, above the x-axis and between the vertical lines $x = a$ and $x = b$. Notationally, we write

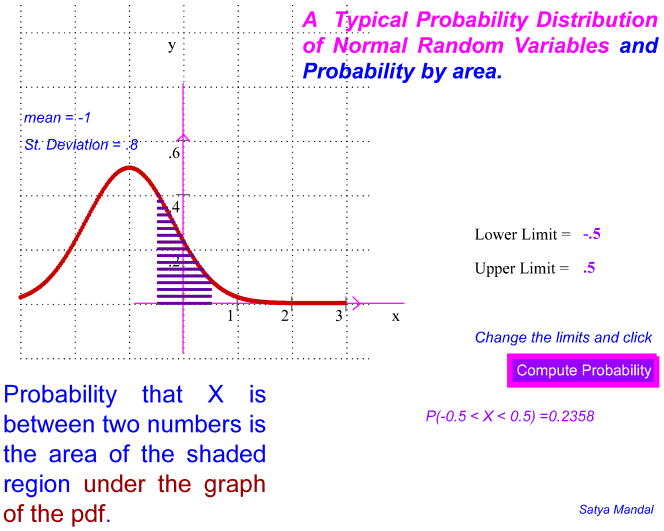
$$\begin{cases} P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b) = \\ \textbf{the area of the region under the graph of } y = f(x), \\ \textbf{above x} - \textbf{axis}, \\ \textbf{between the vertical lines } x = a \text{ and } x = b. \end{cases}$$

3. (**A property**:) It follows from this that, for any real number a

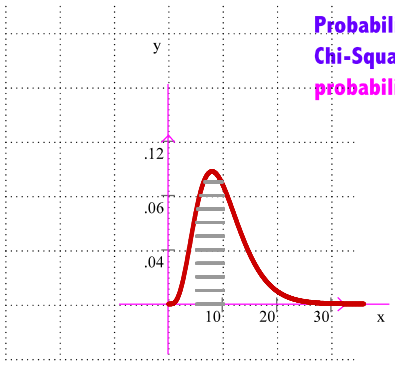
$$P(X = a) = P(a \leq X \leq a) = 0$$

This clearly distinguishes continuous random variable, from discrete random variables (like Binimial), where probability is concentrated at certain points.

The following, four diagrams are the graphs of the pdf of four different random variables:



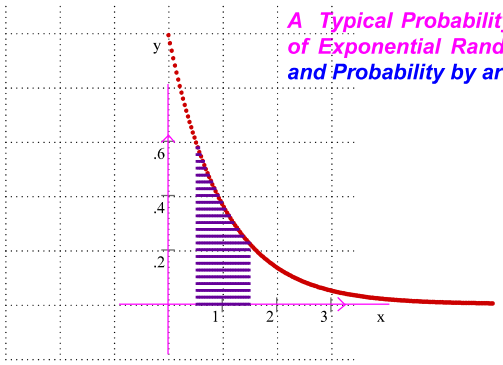
Probability that X is between two numbers is the area of the shaded region under the graph of the pdf.



Probability Distribution of typical Chi-Square random variables and probability by area.

Probability that X is between two numbers is the area of the shaded region under the graph of the pdf.

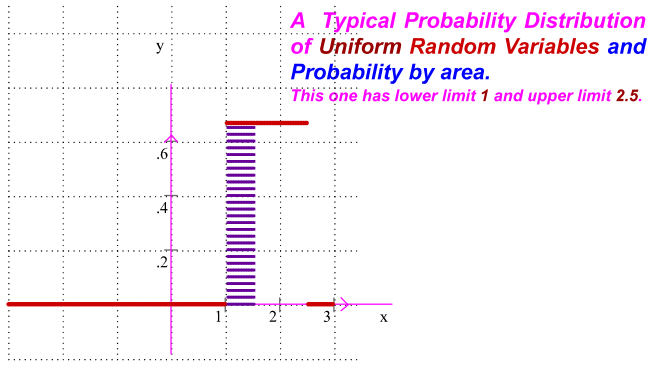
P(5 < X < 10) =0.4508



A Typical Probability Distribution of Exponential Random Variables and Probability by area.

Probability that X is between two numbers is the area of the shaded region under the graph of the pdf.

P(0.5 < X < 1.5) =0.3838



Remark. Given a continuous random variable X, to get a model for the pdf $f(x)$, we look at large samples and their relative frequency histograms. We try to encase the histogram, under the graph of a suitable function $y = f(x)$, which is accepted a the model for the pdf $f(x)$.

5.1.1 The Mean, Variance and Standard Deviation

Suppose X is a continuous random variable. The mean μ, variance σ^2 and standard deviation σ of continuous random variables X represent the same thing as in the case of discrete random variables. The mean μ of X represents the average value of X. The mean μ is also called the Expected value $E(X)$ of X. The standard deviation? of X is a measure of variability of X.

A formal definition involves some knowledge of Calculus (integration), which is not a prerequisite for this course. For this course, a formal definition will not be necessary. However, we try to give a flavor, which some of you may ignore (if you did not take

calculus).

Knowledge of of integration (taught in calculus classes), is necessary to formally define various concepts for a continuous random variables. Since Calculus is not a prerequisite for this course, let me comment that **integration takes the place of summation** in case of continuous random variables. Summation $\sum$ sign was used to define the mean and variance of discrete random variables. If you replace the summation sign $\sum$ in those definitions, in the case of discrete random variables, by the integration sign $\int$, you would get the corresponding definitions for continuous random variables.

Suppose X is a continuous random variable and $f(x)$ is the pdf of X. Then,

1. The probability
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

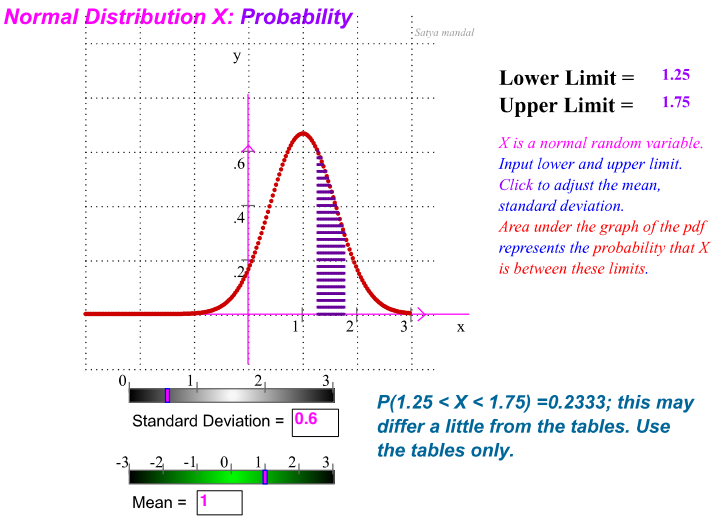
2. The mean $\mu = E(X)$ of X is
$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

3. The variance σ^2 of X is
$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \left(\int_{-\infty}^{\infty} x^2 f(x)dx\right) - \mu^2$$

5.1.2 Examples of Continuous random (read only)

This subsection is devoted to showcase the graph of the pdfs of some of the well known continuous random variables.

Example. The **Normal random variables** will be used most extensively, throughout the rest of this course. We would pursue detailed discussion on it, in the next section. You will notice, that two unknown parameters the mean μ and the standard deviation σ determine the pdf of a normal random variable. A goal of this course would be to estimate these parameters.



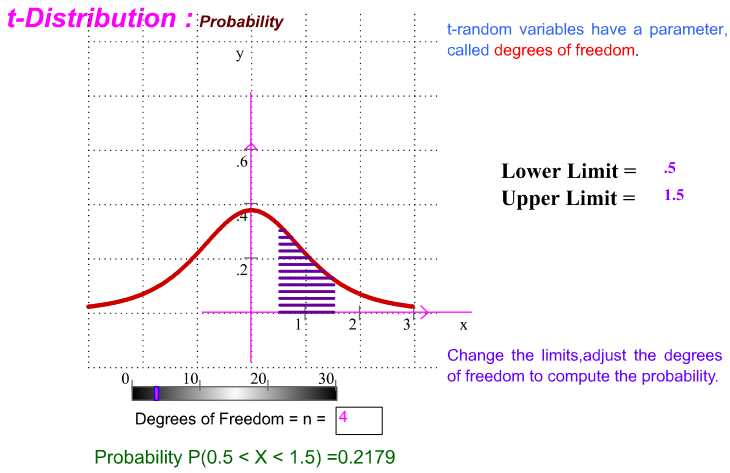
The pdf of a normal random variable X with mean μ and st. deviation σ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty \tag{5.1}$$

So, the pdf is completely determined **by two parameters**, μ and σ^2. We can prove, the expectation $\mu = E(X)$ and variance of X is:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{\infty} x \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \right) dx = \mu \tag{5.2}$$

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} (x-\mu)^2 \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \right) dx = \sigma^2 \tag{5.3}$$



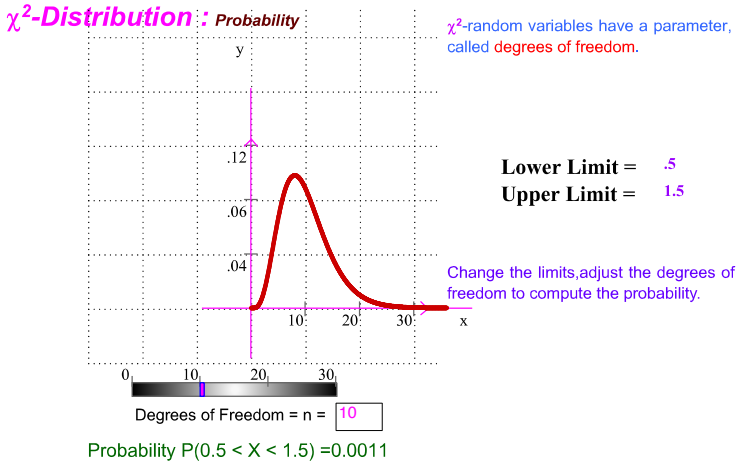
The pdf of a T-random variable X with **degrees of freedom** $n \geq 2$ is:

$$f(x) = A(n)\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad -\infty < x < \infty \quad A(n) = \begin{cases} \frac{3\cdot 5\cdot\cdots\cdot(n-1)}{2\sqrt{n}(2\cdot 4\cdot\cdots\cdot(n-2))} & n \text{ even} \\ \frac{2\cdot 4\cdot\cdots\cdot(n-1)}{\pi\sqrt{n}(3\cdot 5\cdot\cdots\cdot(n-2))} & n \text{ odd} \end{cases} \tag{5.4}$$

$$\text{So} \quad E(X) = \mu = \int_{-\infty}^{\infty} x f(x)dx = \int_{-\infty}^{\infty} x\left(A(n)\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\right)dx = 0 \quad n \geq 2$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2\left(A(n)\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\right)dx = \frac{n}{n-2} \quad n \geq 3$$

The pdf of T-random variable look fairly similar to normal, with mean $\mu = 0$. It is evident from (5.4), it has a only **one parameter** n, known as **degrees of freedom**. We would have some use of T-random variable in Chapter 7, 8, 9.



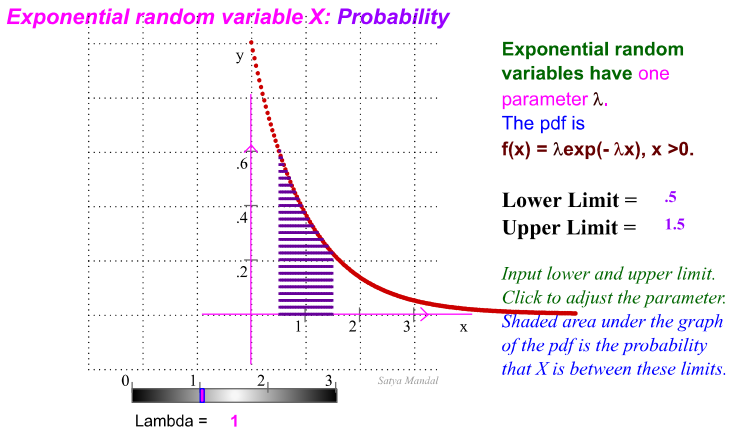
The pdf of a χ^2-random variable X with **degrees of freedom** $n \geq 1$ is:

$$f(x) = \frac{1}{2\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\frac{n}{2}-1} \quad 0 \leq x < \infty, \; \Gamma\left(\frac{n}{2}\right) = \begin{cases} (r-1)! & n = 2r \; even \\ 0.5 * 1.5 \cdots (r-.5) \int_0^\infty \frac{e^{-y}}{\sqrt{y}} dy & n = 2r+1 \; odd \end{cases} \tag{5.5}$$

$$\text{So} \quad E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx = \int_0^\infty x \left(\frac{1}{2\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\frac{n}{2}-1} \right) dx = n$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_0^\infty (x-n)^2 \left(\frac{1}{2\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\frac{n}{2}-1} \right) dx = 2n$$

The pdf of χ^2-random variable has a bell shape, but it is skewed. It has a only one **parameter** n, known as **degrees of freedom**. Further, a χ^2-random variable is always **non negative**. We would have some use of χ^2-random variable in Chapter 7, 8, 9.



The pdf of an exponential random variable X with **parameter** $\lambda > 0$ is:

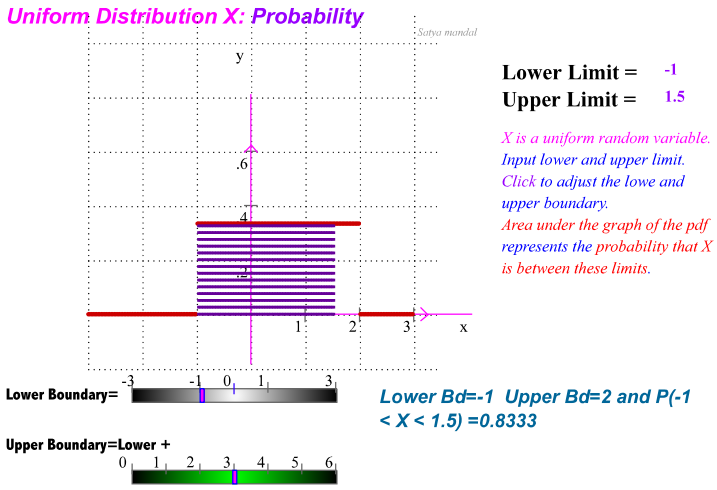
$$f(x) = \lambda e^{-\lambda x} \qquad 0 \leq x < \infty \tag{5.6}$$

In this case, we say X is $\exp(\lambda)$-random variable.

$$\text{So} \quad E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{\infty} x \left(\lambda e^{-\lambda x}\right) dx = \frac{1}{\lambda}$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{0}^{\infty} \left(x - \frac{1}{\lambda}\right)^2 \left(\lambda e^{-\lambda x}\right) dx = \frac{1}{\lambda^2}$$

We do not have any plan to use exponential random variable, beyond this point.



Uniform random variables are simplest and basic, among the continuous random variables. Give two numbers $L < U$, the pdf of an uniform random variable X, is defined to be,

$$f(x) = \begin{cases} \frac{1}{U-L} & if\ L \le x \le U \\ 0 & otherwise. \end{cases}$$

In this case, we say X is $Uniform(L, U)$-random variable.

$$\text{So} \quad E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx = \int_{L}^{U} x \left(\frac{1}{U-L}\right) dx = \frac{L+U}{2} \qquad \text{the mid point}$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{L}^{U} \left(x - \frac{L+U}{2}\right)^2 \left(\frac{1}{U-L}\right) dx = \frac{L^2 - LU + U^2}{12}$$

Remark. Main point that distinguishes continuous random variables from discrete random variables is the following:

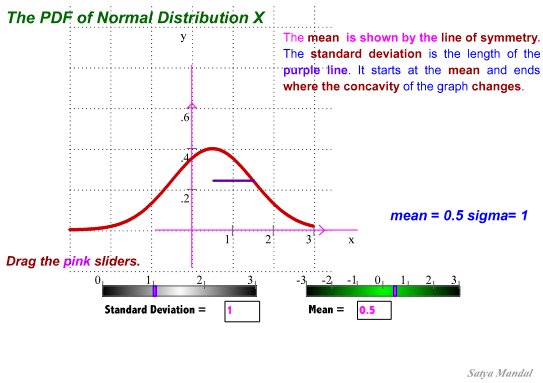
1. A continuous random variable X has a probability density function (pdf) $f(x)$.
2. A discrete random variable X has a probability function $p(x)$, also known as probability **mass** function.

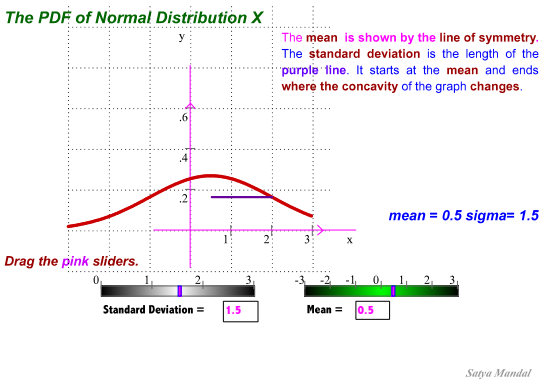
5.2 The Normal Random Variable

The most commonly encountered random variable in nature and real life is the **normal random variable**.

Definition and Properties:

1. Given real numbers μ, $\sigma > 0$, there is a continuous random variable X, to be called the **normal random variable**, whose pdf is given above (5.1).
2. We say X have $N(\mu, \sigma)$-distribution, or say X is a $N(\mu, \sigma)$-random variable. We also write $X \sim N(\mu, \sigma)$, to mean the same thing.
3. Main point it, the graph of the pdf $y = f(x)$ is symmetric around, the vertical line $x = \mu$. As the mean increases or decreases, the graph of the pdf only translates to the right or left.
4. As the standard deviation σ decreases, the whole probability-area concentrates near the mean $x = \mu$. To see this phenomenon, compare the following two diagrams.
5. The standard deviation σ is the distance from the line of symmetry to the point where the concavity of the graph changes from cupped down to cupped up. To see this compare the following two diagram.





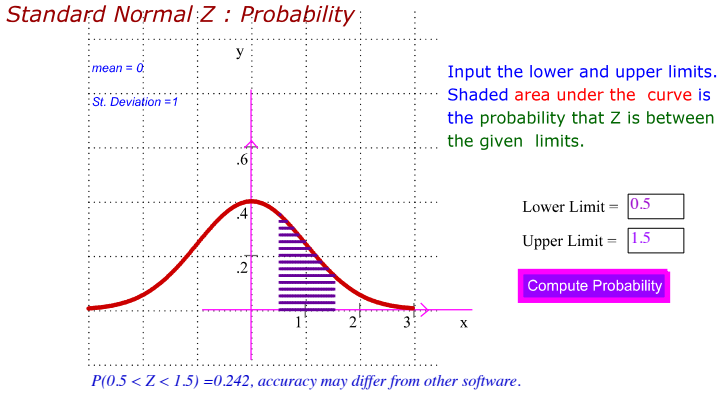
Let us keep in mind, that

1. The total area under the graph, above x-axis, is 1.

2. $P(a \leq X \leq b)$ is the area, under the graph, between vertical lines $x = a$ and $X = b$, above x-axis.

To compute probability, for $X \sim N(\mu, \sigma)$, we standardize things, as follows.

Definition. The normal random variable Z with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **Standard Normal Random Variable**. So, $Z \sim N(0, 1)$. The notation Z is used fairly consistently, to denote a standard normal random variable. The graph of its pdf, is as follows:



Following are some of the properties and usage of $Z \sim N(0, 1)$:

1. Reinterpreting the above properties of normal random variables, the graph of the pdf $y = f(x)$ of the standard random variable Z is symmetric around the y-axis.

2. Since the total area under the graph and above the x-axis is one, on each side of the y-axis the corresponding area is 0.5.

3. Probability tables for the standard normal variable Z are commonly used to compute probability, in college level statistics courses. Such tables for Z would be available

in any standard textbook or Internet. Because of the availability of simple and sophisticated tools (like TI-84, Excel, SAS, Minitab) the usage in probability tables may have **become outdated**. In this course, TI-84 will be using compute probability.

4. The distribution of the standard normal random variable Z is **the most basic tool** used, to deal with problems on any normal random variable X. (In turn, **normal random variables are most basic in nature and real life**.) Any problem on normal random variables $X \sim N(\mu, \sigma)$, is reduced to a problem on standard normal random variable $Z \sim N(0, 1)$. This process of reduction is called **standardization**. The following theorem makes this precise.

Theorem. Let X $N(\mu, \sigma)$-random variable.
Then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ is a standard normal. So,

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \textbf{normalcdf}\left(\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right) \tag{5.7}$$

where **normalcdf**$(-,-)$ is the function in TI-84, that comes under **"Distr"-key**.

Ubiquity of Normal Random Variables: Any random variable that we encounter in nature is, almost certainly, either normal or approximately normal. If there is one piece of information that you want to take from this course, it is this: *nature's random variables are normal or approximately normal. You will hear about normal random variables and the bell curve in your workplace or anywhere you may have to use statistics.*

Problem Solving: There are two steps involved in solving problems in this section:

1. Standardizing the problem to a Z-problem.
2. Use TI-84 to compute the Z-probability.
3. **Example:** Suppose $X \sim N(2, .5)$ random variable.
   What is the probability that $P(1 \leq X \leq 2.5)$?
   Solution: Here $\mu = 2$, $\sigma = .5$.
   $P(1 \leq X \leq 2.5)$
   $= P\left(\frac{1-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{2.5-\mu}{\sigma}\right)$
   $= P\left(\frac{1-\mu}{\sigma} \leq Z \leq \frac{2.5-\mu}{\sigma}\right)$
   $= P\left(\frac{1-2}{.5} \leq Z \leq \frac{2.5-2}{.5}\right)$
   $= P(-2 \leq Z \leq 1) = normalcdf(-2, 1) = .8186$

Use TI-84 to comute $Z \sim N(0, 1)$ Probability:

1. To compute$P(1 \leq Z \leq 2.5)$do the following.
   a) Press 2nd and then Distr (VARS)
   b) Scroll down to normalcdf and ENTER
   c) type in " $-2, 1.5)$" and ENTER. TI will give the answer.

2. To compute $P(Z \leq 1.5)$ do the following.
   a) Press 2nd and then Distr (VARS)
   b) Scroll down to normalcdf and ENTER
   c) type in " $-5, 1.5)$" and ENTER. TI will give the answer.
   (*We used -5 instead of minus infinity.*
   There is no infinity key in TI and this will be precise enough.
   You can check that $P(-5 \leq Z \leq 5) = normalcdf(-5, 5) \approx 1$.)

3. To compute $P(1.53 \leq Z)$ do the following.
   a) Press 2nd and then Distr (VARS)
   b) Scroll down to normalcdf and ENTER
   c) type in "$1.53, 5)$" and ENTER. TI will give the answer.
   (*We used 5 instead of infinity.This will be precise enough.*)

4. Also note, for continuous random variables, It will not make any difference, if $\leq$ is replaced by $<$, and conversely.

5.2.1 Problems: on $X \sim N(\mu, \sigma)$

Exercise 5.2.1. Let $Z \sim N(0, 1)$ be the standard normal random variable.

1. Find the probability $P(-1.1 < Z < 2.5)$.

2. Find the probability $P(Z < -2.1)$.

3. Find the probability $P(-2.1 < Z < -1.5)$.

4. Find the probability $P(1.5 < Z)$.

Solution:

This is already standard normal. Therefore, no standardization is needed. We use TI-84 :

1. $P(-1.1 < Z < 2.5) = normalcdf(-1.1, 2.5) = 8581$

2. $P(Z < -2.1) = normalcdf(-5, -2.1) = .01786$

3. $P(-2.1 < Z < -1.5) = normalcdf(-2.1, -1.5) = .0489$

4. $P(1.5 < Z) = normalcdf(1.5, 5) = .0668$

Exercise 5.2.2. Let X be a normal random variable with mean $\mu = 3$ and standard deviation $\sigma = 1.5$.

1. Find the probability $P(-1.1 < X < 2.5)$.
2. Find the probability $P(X < -2.1)$.
3. Find the probability $P(-1.2 < X < -0.5)$.
4. Find the probability $P(1.5 < X)$.

Solution:
Here the mean $\mu = 3$ and standard deviation $\sigma = 1.5$
$X \sim N(3, 1.5)$ -random variable. We use TI-84 :

1. $P(-1.1 < X < 2.5)$
   $= P\left(\frac{-1.1-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{2.5-\mu}{\sigma}\right)$
   $= P\left(\frac{1.1-\mu}{\sigma} < Z < \frac{2.5-\mu}{\sigma}\right)$
   $= P\left(\frac{-1.1-3}{1.5} < Z < \frac{2.5-3}{1.5}\right)$
   $= P\left(-2.7333 < Z < -.3333\right) = normalcdf(-2.7333, -.3333) = 0.3663.$
2. $P(X < -2.1)$
   $= P(\frac{X-\mu}{\sigma} < \frac{-2.1-\mu}{\sigma})$
   $= P(Z < \frac{-2.1-\mu}{\sigma})$
   $= P(Z < \frac{-2.1-3}{1.5})$
   $= P(Z < -3.4) = normalcdf(-5, -3.4) = 3.3669 * 10^{-4}$
3. $P(-1.2 < X < -0.5)$
   $= P(\frac{-1.2-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{-0.5-\mu}{\sigma})$
   $= P(\frac{-1.2-\mu}{\sigma} < Z < \frac{-0.5-\mu}{\sigma})$
   $= P(\frac{-1.2-3}{1.5} < Z < \frac{-0.5-3}{1.5})$
   $= P(-2.8 < Z < -2.3333) = normalcdf(-2.8, -2.3333) = 0.0073$
4. $P(1.5 < X) = P(\frac{1.5-\mu}{\sigma} < \frac{X-\mu}{\sigma})$
   $= P(\frac{1.5-\mu}{\sigma} < Z)$
   $= P(\frac{1.5-3}{1.5} < Z) = P(-1 < Z) = normalcdf(-1, 5) = .8413$

Exercise 5.2.3. The length of life of some light bulbs produced in a factory is normally distributed with mean 8640 hours and standard deviation 1440 hours. Find the probability that a bulb will last

1. less than 5040 hours;

2. between 5040 hours and 8640 hours.

Solution:
Here the mean $\mu = 8640$ and standard deviation $\sigma = 1440$.
Let X = length of life of light bulbs produced in the factory.
Then $X \sim N(8640, 1440)$-random variable. We use TI-84 :

1. $P(X < 5040)$
   $= P(\frac{X-\mu}{\sigma} < \frac{5040-\mu}{\sigma})$
   $= P(Z < \frac{5040-\mu}{\sigma})$
   $= P(Z < \frac{5040-8640}{1440})$
   $= P(Z < -2.5) = normalcdf(-5, -2.5) = .0062$

2. $P(5040 < X < 8640)$
   $= P(\frac{5040-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{8640-\mu}{\sigma})$
   $= P(\frac{5040-\mu}{\sigma} < Z < \frac{8640-\mu}{\sigma})$
   $= P(\frac{5040-8640}{1440} < Z < \frac{8640-8640}{1440})$
   $= P(-2.5 < Z < 0) = normalcdf(-2.5, 0) = 0.4939$

Exercise 5.2.4. The length X of a fish in a lake has normal distribution with mean 67 cm and standard deviation 21 cm. What proportion (i.e, probability) of fish are between 44 cm and 110 cm long?

Solution:
Here the mean $\mu = 67$ and standard deviation $\sigma = 21$.
Let X = length of fish in the lake. Then, $X \sim N(67, 21)$-random variable. We use TI-84 :

$P(44 < X < 110)$
$= P(\frac{44-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{110-\mu}{\sigma})$
$= P(\frac{44-\mu}{\sigma} < Z < \frac{110-\mu}{\sigma})$
$= P(\frac{44-67}{21} < Z < \frac{110-67}{21})$
$= P(-1.0952 < Z < 2.0476) = normalcdf(-1.0952, 2.0476) = 0.8430$
If the answer is asked in percent, it would be 84.30 percent.

Exercise 5.2.5. The diameter of the pumpkins in my patch has normal distribution with mean 13 inches and standard deviation 4.5 inches. What proportion (i.e., probability) of pumpkins is above 22 inches?

Solution:

Here the mean $\mu = 13$ and standard deviation $\sigma = 4.5$

Let $X =$ diameter of the pumpkins. Then, $X \sim N(13, 4.5)$-random variable. We use TI-84:

$P(22 < X)$
$= P(\frac{22-\mu}{\sigma} < \frac{X-\mu}{\sigma})$
$= P(\frac{22-\mu}{\sigma} < Z)$
$= P(\frac{22-13}{4.5} < Z)$
$= P(2 < Z) = normalcdf(2, 5) = .0227$

Since the answer is asked in percent, it would be 2.27 percent.

Exercise 5.2.6. The annual expenditure X of a student is approximately normally distributed with mean $\mu = 11,000$ dollars and standard deviation $\sigma = 1500$ dollars. What percent of students spend less than 10,000 dollars?

Solution:

Here the mean $\mu = 11,000$ and standard deviation $\sigma = 1500$. Let $X =$ annual expenditure of students. Then, $X \sim N(11000, 1500)$-random variable. We use TI-84 :

$P(X < 10000)$
$= P(\frac{X-\mu}{\sigma} < \frac{10000-\mu}{\sigma})$
$= P(Z < \frac{10000-\mu}{\sigma})$
$= P(Z < \frac{10000-11000}{1500})$
$= P(Z < -.6667) = normalcdf(-5, -.6667) = .2525$

Since the answer is asked in percent, it would be 25.25 percent.

Exercise 5.2.7. Suppose the annual production X of milk by cows in a farm is normally distributed with $\mu = 5500$ liters and standard deviation $\sigma = 150$ liters. What percent of cows have annual yield less than 5155 liters?

Solution:

Here the mean $\mu = 5500$ and standard deviation $\sigma = 150$

Let $X =$ annual production by cows in the farm.

Then $X \sim N(5500, 150)$-random variable. We use TI-84:

$P(X < 5155)$
$= P(\frac{X-\mu}{\sigma} < \frac{5155-\mu}{\sigma})$
$= P(Z < \frac{5155-\mu}{\sigma})$
$= P(Z < \frac{5155-5500}{150})$
$= P(Z < -2.3) = normalcdf(-5, -2.3) = .0107$
Since the answer is asked in percent, it would be 1.07 percent.

Exercise 5.2.8. The amount of vegetable oil X produced by a machine in a day is normally distributed with $\mu = 130$ liters and standard deviation $\sigma = 25$ liters. What is the probability that a machine will produce between 120 liters and 150 liters on a day?

Solution:
Here the mean $\mu = 130$ and standard deviation $\sigma = 25$.
Let $X =$ vegetable oil produced by the machines.
Then $X \sim N(130, 25)$-random variable. We use TI-84:

$P(120 < X < 150)$
$= P(\frac{120-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{150-\mu}{\sigma})$
$= P(\frac{120-\mu}{\sigma} < Z < \frac{150-\mu}{\sigma})$
$= P(\frac{120-130}{25} < Z < \frac{150-130}{25})$
$= P(-.4 < Z < .8) = 0.4436$
Since the answer is asked in percent, it would be 44.36 percent.

Exercise 5.2.9. The weight X at birth of babies is normally distributed with mean $\mu = 114$ oz and standard deviation $\sigma = 18$ oz. What percent of babies will have birth weight below 141 oz?

Solution:
Here the mean $\mu = 114$ and standard deviation $\sigma = 18$.
Let $X =$ vegetable oil produced by the machines.
Then $X \sim N(114, 18)$-random variable. We use TI-84 :

$P(X < 141)$
$= P(\frac{X-\mu}{\sigma} < \frac{114-\mu}{\sigma})$
$= P(Z < \frac{141-\mu}{\sigma})$
$= P(Z < \frac{141-114}{18})$
$= P(Z < 1.5) = normalcdf(-5, 1.5) = .9332$
Since the answer is asked in percent, it would be 93.32 percent.

5.2.2 Inverse Probability (Cut-Off Values)

In certain situation, the probability is known and the variable values would have to be determined. This concept is called inverse probability. These are also called problems of **Cut-Off Values**. Following is an example of such a situation.

Example. Most common examples of inverse probability or cut-off values would be the **percentiles**. Suppose X is the birth weight of babies in a county. When you may want to determine the 75-**percentile** of the birth weight (or of weight at any age), it means that it is given that the probability $P(X < u) = .75$ and the seventy five percentile u is to be determined.

Pediatricians have charts of percentile weights. These charts have age on the horizontal axis and weight on the vertical axis. Several graphs are given in the chart, representing various percentiles (I guess, one for each 5 percent). So, there may be one graph for each of 5-percentile, 10-percentile and up to 95-percentile.

Other such example would be **income percentiles** of the US population. This is available in Wiki.

Example. The annual income X in a county is normally distributed with mean $\mu = \$37,000$ and standard deviation \$15,000. Your annual income is \$75,000. Do you think that your annual income is above 90 percentile?
Solution.
Here the mean $\mu = 37,000$ and standard deviation $\sigma = 15000$
Let $X =$ annual income of the members of the county.
Then $X \sim N(37000, 15000)$ -random variable.
We would use **invNorm** function is TI-84, as follws:

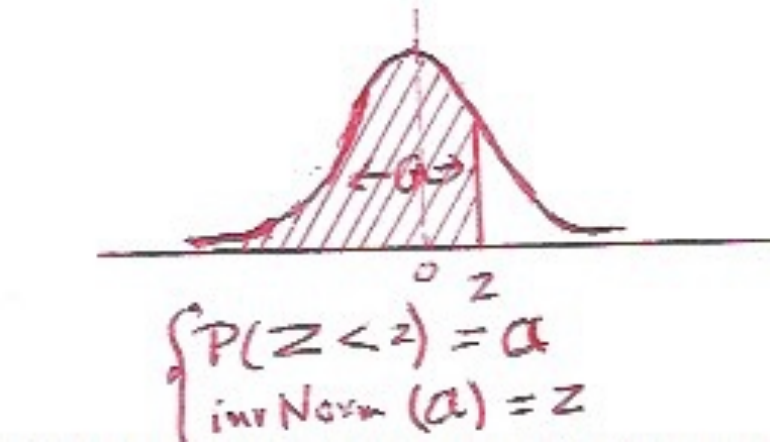
Let $u = 90$ percentile of the annual income. So, $P(X < u) = .90$
$P(\frac{X-\mu}{\sigma} < \frac{u-\mu}{\sigma}) = .90$
$\begin{cases} P(Z < \frac{u-\mu}{\sigma}) = .90 & \text{Also,} \\ P(Z < 1.2816) = .90 & \text{Because , } invNorm(.90) = 1.2816 \end{cases}$
By comparing these two equations

$\frac{u-\mu}{\sigma} = 1.2816$ So,
$u = \mu + 1.2816 * \sigma = 37000 + 1.2816 * 15000 = 56224.$
That means that the 95-percentile of the annual income is \$56224.
We conclude your annual income \$75,000 is above 90^{th} percentile.

TI-84: the inNorm function
This falls under inverse probability. By inverse probability, we mean that the probability is given and we want to find certain variable value. For example, suppose $Z \sim N(0, 1)$, and suppose

$\begin{cases} P(Z \leq u) = .90. & \textbf{what is u?} \quad In\ fact, \\ P(Z \leq 1.2816) = .90. & \text{This information is given by} \quad \textbf{invNorm(.90) = 1.2816} \end{cases}$

where **invNorm**$(-)$ is a function in TI-84, under Distr. key.
Following is how we use **invNorm**$(-)$:
to compute the variable value (or Z-value) u such that $P(Z < u) = .90$ do the following,
a) Press 2nd and then Distr (VARS)
b) Scroll down to invNorm and ENTER
c) type in ".90)" and ENTER. TI will give the answer.



More generally, suppose $X \sim N(\mu, \sigma)$. For

$$0 < \mathfrak{a} < 1 \quad \text{to find u} \quad \text{such that} \quad P(X \leq u) = \mathfrak{a}$$

do the following:

1. Use the invNorm function from TI-84 and Find **invNorm**$(\mathfrak{a}) = \mathbf{z}$.
2. Standardize and reduce the equation $P(X \leq u) = \mathfrak{a}$ to $P\left(Z \leq \frac{u-\mu}{\sigma}\right) = \mathfrak{a}$.
   So, we have
   $$\begin{cases} P\left(Z \leq \frac{u-\mu}{\sigma}\right) = \mathfrak{a} \\ P(Z \leq z) = \mathfrak{a} \end{cases}$$
3. Compare two equations and we have
   $$\frac{u-\mu}{\sigma} = z$$

5.2.3 Problems on Cut-off values

Exercise 5.2.10. Let Z be the standard normal random variable.

1. Given that $P(-1.1 < Z < c) = .6881$, find c.
2. Given that $P(Z < c) = 0.0222$, find c.

3. Given that $P(c < Z < 1.5) = 0.0919$, find c.

4. Given that $P(c < Z) = 0.102$, find c.

Solution.

This is a problem in standard normal Z. Standardization would not be needed for this problem, because Z is already standard normal. We use TI-84 :

1. $P(-1.1 < Z < c) = .6881.$
   $P(Z < c) - P(Z < -1.1) = .6881.$
   $P(Z < c) - normalcdf(-5, -1.1) = .6881.$
   $P(Z < c) - .1357 = .6881.$
   $\begin{cases} P(Z < c) = .8238. \quad \text{Also,} \\ P(Z < .9299) = .8238. \quad \text{Because } invNorm(.8238) = .9299 \end{cases}$
   By comparing these two equations
   $c = .9299.$

2. $\begin{cases} P(Z < c) = 0.0222 \quad \text{Also,} \\ P(Z < -2.0103) = 0.0222 \quad \text{Because } invNorm(.0222) = -2.0103. \end{cases}$
   by comparing these two equations
   $c = -2.0103$

3. $P(c < Z < 1.5) = 0.0919$
   $P(Z < 1.5) - P(Z < c) = .0919.$
   $normalcdf(-5, 1.5) - P(Z < c) = .0919.$
   $.9332 - P(Z < c) = .0919.$
   $\begin{cases} P(Z < c) = .8413. \quad \text{Also,} \\ P(Z < .9998) = .8413. \quad \text{Because } invNorm(.8413) = .9998]. \end{cases}$
   By comparing these two equations
   $c = .9998$

4. $P(c < Z) = 0.102$ So,
   $1 - P(Z < c) = 0.102$
   $\begin{cases} P(Z < c) = .898. \quad \text{Also,} \\ P(Z < 1.2702) = .898. \quad \text{Because } invNorm(.898) = 1.2702. \end{cases}$
   By comparing these two equations
   $c = 1.2702$

Exercise 5.2.11. The length X of a fish in a lake has normal distribution with mean 67 cm and standard deviation 21 cm. On a fishing trip to the lake, you are instructed to keep

only those in the upper 33 percent in length. What is the cut-off length, above which you are permitted keep?

Solution.

Here the mean $\mu = 67$ and standard deviation $\sigma = 21$.
Let $X =$ the length of in the lake.
Then $X \sim N(67, 21)$ random variable. We use TI-84:

Let $L =$ the cut-off length.
$P(L < X) = .33$ So, $1 - P(X < L) = .33$
$P(X < L) = .67$
$P\left(\frac{X-\mu}{\sigma} < \frac{L-\mu}{\sigma}\right) = .67$
$\begin{cases} P\left(Z < \frac{L-\mu}{\sigma}\right) = .67 & \text{Also,} \\ P(Z < .4399) = .67 & \text{Because } invNorm(.67) = .4399 \end{cases}$
By comparing the above two equations
$\frac{L-\mu}{\sigma} = .4399$ So,
$L = \mu + .4399 * \sigma = 67 + .4399 * 21 = 76.2379$ cm.

Exercise 5.2.12. The telephone company's data shows that length X of their international calls has normal distribution with mean 11.5 minutes and standard deviation 4.3 minutes. The company decided to give a special rate for the longest 20 percent calls. What is the cut-off time length?

Solution.

Here the mean $\mu = 11.5$ and standard deviation $\sigma = 4.3$
Let $X =$ the length of the international calls in minutes.
Then $X \sim N(11.5, 4.3)$ random variable. We use TI-84:

Let $L =$ the cut-off length above which the special rate applies.
$P(L < X) = .20$. So, $1 - P(X < L) = .20$, and $P(X < L) = .80$.
$P\left(\frac{X-\mu}{\sigma} < \frac{L-\mu}{\sigma}\right) = .80$
$\begin{cases} P\left(Z < \frac{L-\mu}{\sigma}\right) = .80 & \text{Also,} \\ P(Z < .8416) = .80 & \text{Because } invNorm(.80) = .8416 \end{cases}$
By comparing the above two equations
$\frac{L-\mu}{\sigma} = .8416$
So,
$L = \mu + .8416 * \sigma = 11.5 + .8416 * 4.3 = 15.1189$ minutes.

Exercise 5.2.13. The weight X of babies (of a fixed age) is normally distributed with

with mean $\mu = 212$ oz and standard deviation $\sigma = 25$ oz. Doctors would be concerned (not necessarily alarmed) if a baby is among the lower 5 percent in weight. Find the cut-off weight L below which the doctors will be concerned.

Solution.

Here the mean $\mu = 212$ and standard deviation $\sigma = 25$.
Let $X =$ the weight of the babies of this age group.
Then $X \sim N(212, 25)$ random variable. We use TI-84:

Let $U =$ the cut-off length below which doctors will be concerned.
$P(X < U) = .05$
$P\left(\frac{X-\mu}{\sigma} < \frac{U-\mu}{\sigma}\right) = .05$
$P\left(Z < \frac{U-\mu}{\sigma}\right) = .05$
$\begin{cases} P\left(Z < \frac{U-\mu}{\sigma}\right) = .05 & \text{Also,} \\ P(Z < -1.6449) = .05 & \text{Because } invNorm(.05) = -1.6449 \end{cases}$
By comparing the above two equations
$\frac{U-\mu}{\sigma} = -1.6449$ So,
$U = \mu + (-1.6449) * \sigma = 212 + (-1.6449) * 25 = 170.8775$ oz.

Exercise 5.2.14. Let X be a normal random variable with mean $\mu = 4$ and standard deviation $\sigma = 2.5$. Find the values of c from the following equations.

1. Suppose $P(L \leq X \leq 6.5) = .7787$. What is L?
2. Suppose $P(2.5 \leq X \leq U) = .6000$. What is U?
3. Suppose $P(X \leq u) = .0548$. What is u
4. Suppose $P(l \leq X) = .7775$. What is l?

Solution.

Here the mean $\mu = 4$ and standard deviation $\sigma = 2.5$.
We use TI-84 :

1. $P(L \leq X \leq 6.5) = .7787$
   $P(\frac{L-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{6.5-\mu}{\sigma}) = .7787$
   $P(\frac{L-\mu}{\sigma} \leq Z? \leq \frac{6.5-\mu}{\sigma}) = .7787$
   $P(Z \leq \frac{6.5-\mu}{\sigma}) - P([Z \leq \frac{L-\mu}{\sigma}) = .7787$

$P(Z \leq \frac{6.5-4}{2.5} - P(Z \leq \frac{L-\mu}{\sigma}) = .7787$
$P(Z \leq 1) - P([Z \leq \frac{L-\mu}{\sigma}) = .7787$
$normalcdf(-5, 1) - P(Z \leq \frac{L-\mu}{\sigma}) = .7787$
$.8413 - P(Z \leq \frac{L-\mu}{\sigma}) = .7787$
$\begin{cases} P(Z \leq \frac{L-\mu}{\sigma}) = .0626 & \text{Also,} \\ P(Z \leq -1.5333) = .0626 & \text{Because } invNorm(.0626) = -1.5333 \end{cases}$
Therefore, comparing
$\frac{L-\mu}{\sigma} = -1.5333$
$L = \mu + \sigma * (-1.5333) = 4 + 2.5 * (-1.5333) = 0.1668$

2. $P(2.5 \leq X \leq U) = .6000.$
$P(\frac{2.5-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{U-\mu}{\sigma}) = .6000$
$P(\frac{2.5-\mu}{\sigma} \leq Z \leq \frac{U-\mu}{\sigma}) = .6000$
$P(Z \leq \frac{U-\mu}{\sigma}) - P(Z \leq \frac{2.5-\mu}{\sigma}) = .6000$
$P(Z \leq \frac{U-\mu}{\sigma}) - P(Z \leq \frac{2.5-4}{2.5}) = .6000$
$P(Z \leq \frac{U-\mu}{\sigma}) - P([Z \leq -.6) = .6000$
$P(Z \leq \frac{U-\mu}{\sigma}) - normalcdf(-5, -.6) = .6000$
$P(Z \leq \frac{U-\mu}{\sigma}) - .2743 = .6000$
$\begin{cases} P(Z \leq \frac{U-\mu}{\sigma}) = .8743 & \text{Also,} \\ P(Z \leq 1.1470) = .8743 & \text{Because } invNorm(.8743) = 1.1470 \end{cases}$
Therefore, comparing
$\frac{U-\mu}{\sigma} = 1.1470$
$U = \mu + \sigma * (1.1470) = 4 + 2.5 * (1.1470) = 6.8675$

3. $P(X \leq u) = .0548.$
$P(\frac{X-\mu}{\sigma} \leq \frac{u-\mu}{\sigma}) = .0548$
$\begin{cases} P(Z \leq \frac{u-\mu}{\sigma}) = .0548 & \text{Also,} \\ P(Z \leq -1.6000) = .0548 & \text{Because } invNorm(.0548) = -1.6000 \end{cases}$
Therefore, comparing
$\frac{u-\mu}{\sigma} = -1.6000$
$u = \mu + \sigma * (-1.6000) = 4 + 2.5 * (-1.6000) = 0$

4. $P(l \leq X) = .7775$, So, $1 - P(X \leq l) = .7775$
$P(X \leq l) = .2225$
$P(\frac{X-\mu}{\sigma} \leq \frac{l-\mu}{\sigma}) = .2225$
$\begin{cases} P(Z \leq \frac{l-\mu}{\sigma}) = .2225 & \text{Also,} \\ P(Z \leq -.7638) = .2225 & \text{Because } invNorm(.2225) = -.7638 \end{cases}$
Therefore, comparing

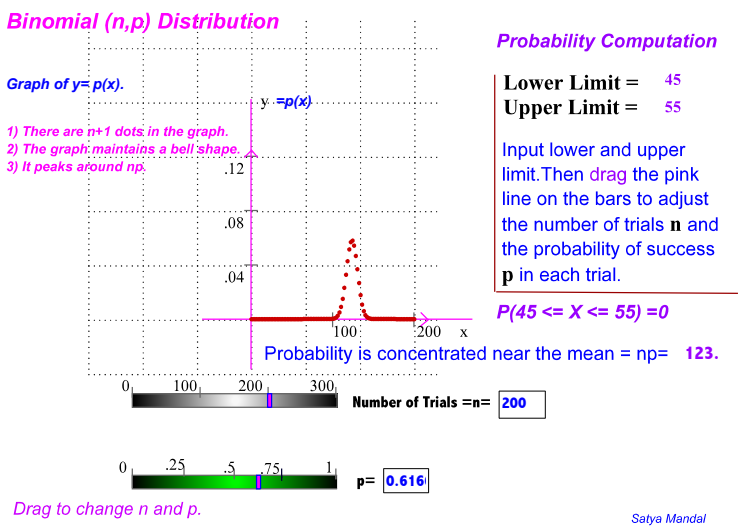
$\frac{l-\mu}{\sigma} = -.7638$
$l = \mu + \sigma * (-.7638) = 4 + 2.5 * (-.7638) = 2.0905$

Exercise 5.2.15. Monthly water consumption X per household, in a subdivision in Kansas City, has normal distribution with mean 15000 gallons and standard deviation 3000 gallons. It has been decided that a surcharge will be imposed for those in the top 25 percent. Find the cut-off consumption u in gallons.

Solution.
Here the mean $\mu = 15000$ and standard deviation $\sigma = 3000$.
u =Cut-off
$P(u \leq X) = .25$, So, $P(X \leq u) = .75$
$P(\frac{X-\mu}{\sigma} \leq \frac{u-\mu}{\sigma}) = .75$
$\begin{cases} P(Z \leq \frac{u-\mu}{\sigma}) = .75 & \text{Also,} \\ P(Z \leq .6745) = .75 & \text{Because } invNorm(.75) = .6745 \end{cases}$
Therefore, comparing
$\frac{u-\mu}{\sigma} = .6745$
$u = \mu + \sigma * (.6745) = 15000 + 3000 * (.6745) = 17023.5$

5.3 Normal Approximation to Binomial

Like most random variables in nature, $Binomial(n, p)$-random variables also behave approximately like Normal random variables.



Visually, the graph of the probability function of the $B(n,p)$-random variables looks similar to that the graph has properties similar to that of Normal:

1. The graph maintains a bell-shape.
2. The graph also peaks at (or near) the mean $\mu = np$.

Below Ex. 4.3.10, we indicated the difficulty to deal with Binomial random variables with large number of trials n. We would approximate Binomial probability, by Normal probability distribution, as follows.

Theorem. Suppose $X \sim B(n,p)$ is a Binomial random variable. Assume n is large and p is not too close to 0 or 1. Then X behaves, approximately, like a $N(\mu,\sigma)$ random variable,

$$\begin{cases} \text{mean} & \mu = np \\ \text{St. Dev.} & \sigma = \sqrt{np(1-p)} \end{cases}$$

More precise rules to approximate is as follows:

1. For $r = 0, 1, 2, \ldots, n$, we have

$$P(X = r) = P(r - .5 \leq X \leq r + .5) \approx P(L \leq Z \leq U) \quad \text{with} \quad \begin{cases} L = \frac{(r-.5)-\mu}{\sigma} \\ U = \frac{(r+.5)-\mu}{\sigma} \end{cases}$$

2. More generally (and usefully)

$$\begin{cases} P(r \leq X \leq s) = P(r - .5 \leq X \leq s + .5) \\ \approx P(L \leq Z \leq U) \end{cases} \quad \text{with} \quad \begin{cases} L = \frac{(r-.5)-\mu}{\sigma} \\ U = \frac{(s+.5)-\mu}{\sigma} \end{cases} \tag{5.8}$$

3. **Remark.** This adjustment by .5 on two sides is called **continuity correction**. Recall $P(Y = r) = 0$ for any continuous random variable Y. Because of this, we could not have treated, naively, X as Normal, without this continuity "correction". On the other hand, when n is very large, it would not make any significant difference. *In such cases, when the computations are completed, your final answer may not show any difference, whether you do the continuity correction or not.*

5.3.1 Problems: On Normal Approximation of $B(n, p)$

Exercise 5.3.1. A Lawrence bank knows that 35 percent of its customers will visit the drive-through window. If 400 customers visit the bank, what is the approximate probability that more than 120 will visit the drive-through window?

Solution.
Here $p = .35$ and $n = 400$.
First step is to compute the mean μ and the standard deviation σ :
$\mu = np = 400 * .35 = 140$ and $\sigma = \sqrt{np(1-p)} = \sqrt{400 * .35 * (1 - .35)} = 9.5394$
Let $X =$ number of customers who will visit the drive-through window.
Then $X \sim B(400, .35)$ random variable and we will use $N(140, 9.5394)$ to approximate.

Now "X is more than 120" means "$X > 120$", which is "$X \geq 121$".
$P(121 \leq X) = P(120.5 \leq X)$ (This is continuity correction.)
$= P(\frac{120.5-\mu}{\sigma} \leq \frac{X-\mu}{\sigma}) \approx P(\frac{120.5-\mu}{\sigma} \leq Z)$
$= P(\frac{120.5-140}{9.5394} \leq Z)$
$= P(-2.0442 < Z) = normalcdf(-2.0442, 5) = .9795$

Exercise 5.3.2. It is known that the probability that a household owns a food processor is 0.1. If 190 households are interviewed, find the approximate probability that

1. more than 26 households own a food processor;

2. less than 30 households own a food processor.

Solution.
Here $p = .1$ and $n = 190$.
First step is to compute the mean μ and the standard deviation σ :
$\mu = np = 190 * .1 = 19$ and $\sigma = \sqrt{np(1-p)} = \sqrt{190 * .1 * (1 - .1)} = 4.1352$
Let X = number of households who own a food processor.
Then $X \sim B(190, .1)$ random variable and we will use $N(19, 4.1352)$ to approximate.

1. Now "X is more than 26" means "$X > 26$", which is "$X \geq 27$".
   $P(27 \leq X) = P(120.5 \leq X)$ (This is continuity correction.)
   $= P(\frac{26.5-\mu}{\sigma} < \frac{X-\mu}{\sigma}) \approx P(\frac{26.5-\mu}{\sigma} < Z)$
   $= P(\frac{26.5-19}{4.1352} < Z) = P(1.8137, < Z) = normalcdf(1.8137, 5) = .0349$

2. Again, "X is less than 30" means "$X < 30$", which is "$X \leq 29$".
   $P(X \leq 29) = P(X \leq 29.5)$ $(Thisiscontinuitycorrection.)$
   $= P(\frac{X-\mu}{\sigma} < \frac{29.5-\mu}{\sigma}) \approx P(Z < \frac{29.5-\mu}{\sigma})$
   $= P(Z < \frac{29.5-19}{4.1352}) = P(Z < 2.5392) = normalcdf(-5, 2.5392) = .9944$

Exercise 5.3.3. The campaign committee of a candidate claims that sixty percent of the voters are in favor of the candidate. You interview 150 voters. Assuming that the campaign committe's claim is accurate, what is the approximate probability that less than 77 will favor the candidate?

Solution.
Here $p = .6$ and $n = 150$.
First step is to compute the mean μ and the standard deviation σ :
$\mu = np = 150 * .6 = 90$ and $\sigma = \sqrt{np(1-p)} = \sqrt{150 * .6 * (1 - .6)} = 6$
Let X = number of voters in favor of the candidate.
Then $X \sim B(150, .6)$ random variable and we will use $N(90, 6)$ to approximate.

Now, "X is less than 77" means "$X < 77$", which is "$X \leq 76$".
$P(X \leq 76) = P(X \leq 76.5)$ (This is continuity correction.)
$= P\left(\frac{X-\mu}{\sigma} < \frac{76.5-\mu}{\sigma}\right) \approx P(Z < \frac{76.5-\mu}{\sigma})$
$= P(Z < \frac{76.5-90}{6}) = P(Z < -2.25) = normalcdf(-5, -2.25) = .0122$

Exercise 5.3.4. A technique is used to fertilize eggs in a fertility clinic laboratory. It is known that the probability that an egg will be fertilized by this technique is 0.1. If 500 eggs are treated, what is the probability that at least 60 eggs will be fertilized?

Solution.
Here $p = .1$ and $n = 500$.
First step is to compute the mean μ and the standard deviation σ :
$\mu = np = 500 * .1 = 50$ and $\sigma = \sqrt{np(1-p)} = \sqrt{500 * .1 * (1 - .1)} = 6.7082$.
Let X = number of egg will be fertilize.
Then $X \sim B(500, .1)$ random variable and we will use $N(50, 6.7082)$ to approximate.

Now "X is at least 60" means "$X \geq 60$".
$P(60 \leq X) = P(59.5 < X)$ (This is continuity correction.)
$= P(\frac{59.5-\mu}{\sigma} < \frac{X-\mu}{\sigma}) \approx P(\frac{59.5-\mu}{\sigma} < Z)$
$= P(\frac{59.5-50}{6.7082} < Z) = P(1.4162, < Z) = normalcdf(1.4162, 5) = .0784$

Exercise 5.3.5. The probability that a computer chip produced in a factory is defective is .2. If you have a sample of 60 chips, what is the probability that the number of defective chips will be less than 20?

Solution.
Here $p = .2$ and $n = 60$.
First step is to compute the mean μ and the standard deviation σ :
$\mu = np = 60 * .2 = 12$ and $\sigma = \sqrt{np(1-p)} = \sqrt{60 * .2 * (1 - .2)} = 3.0984$
Let X = number of defective chips.
Then $X \sim B(60, .2)$ random variable and we will use $N(60, 3.0984)$ to approximate.

Now, "X is less than 20" means "$X < 20$", which is "$X \leq 19$".
$P(X \leq 19) = P(X < 19.5)$ (This is continuity correction.)
$= P(\frac{X-\mu}{\sigma} < \frac{19.5-\mu}{\sigma}) \approx P(Z < \frac{19.5-\mu}{\sigma})$
$= P(Z < \frac{19.5-12}{3.0984}) = P(Z < 2.4206) = normalcdf(-5, 2.4206) = .9923$

Exercise 5.3.6. The probability that a light bulb produced by a machine is defective is p = 0.2. Suppose a quality control inspector takes a sample of 120 bulbs. What is the probability that more than 30 bulbs will be defective?

Solution.
Here $p = .2$ and $n = 120$.
First step is to compute the mean μ and the standard deviation σ :
$\mu = np = 120 * .2 = 24$ and $\sigma = \sqrt{np(1-p)} = \sqrt{120 * .2 * (1 - .8)} = 4.3818$
Let X = number of defective bulbs.
Then $X \sim B(120, .2)$ random variable and we will use $N(24, 4.3818)$ to approximate.

Now "X is more than 30" means "$X > 30$", which is "$X \geq 31$".
$P(31 \leq X) = P(30.5 < X)$ (This is continuity correction.)
$= P(\frac{30.5-\mu}{\sigma} < \frac{X-\mu}{\sigma}) \approx P(\frac{30.5-\mu}{\sigma} < Z)$
$= P(\frac{30.5-24}{4.3818} < Z) = P(1.4834, < Z) = normalcdf(1.4834, 5) = .0694$

Exercise 5.3.7. Suppose the probability that a student has access to the Internet is $p = 0.8$. Suppose you interview 160 students. What is the probability that less than 120 students will have access to the Internet?

Exercise 5.3.8. Suppose that the probability that a person favors medical use of marijuana is $p = 0.6$. If 780 individuals are interviewed, what is the probability that less than 450 will be in favor?

Exercise 5.3.9. Suppose that the probability that a middle-income family invests in the stock market is $p = 0.8$. If we interview 880 middle-income families, what is the probability that more than 700 have invested in the stock market?

Exercise 5.3.10. Suppose that an insurance company knows from experience that the probability that a life-insurance policyholder will survive another 10 years is $p = 0.9$. The company has 2280 policyholders. What is the probability that more than 2025 will survive another 10 years.

Chapter 6

Elements of Sampling Distribution

6.1 Sampling Distribution

The goal of this course has been to develop methods and to use sample **statistics** t (or T) to estimate the population **parameters** θ. For example, to estimate the mean weight μ (the parameter) of the fish population in the nearest lake, you may catch a sample of fish and compute the mean weight $\overline{x}$ (the statistic) of this sample and declare it as an estimate for population mean μ.

Since t would only be an estimate of θ,

1. there would be an error $\varepsilon = |t - \theta|$. We would like this error ε to be small or within our tolerable (specified) **error limit** ρ (say).

2. Further, we would like this error ε to be within our tolerable limit ρ, **more often than not**. For example, we may require that the error ε should be within our tolerance at least 90 percent times (among all our trials).

In fact, our estimate t is a variable number. It varies each time we take a sample. We denote this variable by T. Whenever we have a sample, T has a value $T = t$. Indeed, T is a **random variable** on the sample space of all the possible samples. Therefore, T has a probability distribution. Since T depends on samples, its probability distribution is called a sampling distribution. When we say that "that error ε should be within our tolerance ρ at least 90 percent times", we mean that

$$P(|T - \theta| \leq \rho) = .90 \qquad \text{or higher.}$$

In particular, the sample means $\overline{x}$ of numerical data that we computed in chapter 2 would be the observed values of a random variable $\overline{X}$, corresponding to the sample data we had. Similarly, the sample variances s^2 that we computed in chapter 2 would be the observed values of a random variable S^2. Each time you collect a sample (or data), the

computed sample mean $\overline{x}$ (respectively, the variance s^2, standard deviation s) would be the value of the random variable $\overline{X}$ (respectively, S^2, S) for that sample.

Example. Suppose we want to study the height distribution of the U.S. population. Let X represent the height of the whole US population. We collect sample of size n. The sample would be n numbers

$$x_1, x_2, \ldots, x_n$$

representing the height of n individuals. We shall consider the height x_i of the i^{th} individual as the observed value of a random variable X_i. Here X_i is the notation for the height of the i^{th} member of the sample, which could be the height of anybody from population. Therefore, these n measurements

$$x_1, x_2, \ldots, x_n$$

are, respectively, the observed values of n random variables

$$X_1, X_2, \ldots, X_n.$$

We (re)define the sample mean $\overline{X}$ as a **the random variable**

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

We (re)define sample variance S^2 as a **random variable**

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \tag{6.1}$$

(*See the remark, below* (7.1), *for justification, why the denominator is* $n-1$, *instead of* n.)

We (re)define sample standard deviation S as a **random variable**

$$S = \sqrt{Sample\ Var(X)} = \sqrt{S^2}$$

So, the sample means $\overline{x}$, sample variance s^2 and sample st. deviation s that we computed in chapter 2, would be observed values of the radom variables $\overline{X}$, S^2 and S, respectively.

6.1.1 Sampling Types

There are many ways to do sampling. Most commonly discussed among them are

1. Sampling without replacement,
2. Sampling with replacement.

The **Sampling without replacement** is the type of sampling where, whenever a sample member is selected, the member is excluded from the subsequent selections. It is analogous to selecting n balls from a box of N balls. Balls are selected one by one, without replacing

them back in the box before subsequent selections. This type of sampling is meant to rule out the possibility of selecting a member more than once. For small populations, possibility of selecting a member twice may be significant. For such small populations, sampling without replacement would be appropriate.

The **Sampling with replacement** is the type of sampling where each selection is done without any regard to previous selections. In other words, each time a sample member is drawn, it is placed back to the whole population before the next selection is made. This way, each selection is done from the same whole population. A member could, therefore, be selected more than once. This may seem unnatural. But when working with large populations this is not likely to happen and is **most natural from the statistical point of view**. (How often one receives calls twice for the same poll?) **We will only consider sampling with replacement**.

6.1.2 Properties

Let X (like height) be a random variable with mean μ and variance σ^2. Let $X_1, X_2, \ldots, , X_n$ be a sample from the X-population. We assume that the sampling was done with replacement. Such a sample has the following properties.

1. X would be called the parent population or the population random variable. Also μ and σ^2 are called the population mean and variance.
2. Each of the sample member X_i has the same distribution as X. So, mean of X_i is μ and variance of X_i is σ^2.
3. The sample members $X_1, X_2, \ldots, , X_n$ are all mutually independent. (In fact, one had to ensure that they are drawn independently.)
4. The distribution of the sample mean $\overline{X}$ is called the sampling distribution of $\overline{X}$.

Theorem. Let $X_1, X_2, \ldots, X_n$ be a sample from the X population as above. Then,

1. The mean of the sample mean $\overline{X}$ is the population mean μ, that is
$$E(\overline{X}) = E(X) = \mu$$
(*For this reason, in future, we would say that* $\overline{X}$ *is an unbiased estimator of* μ. *See section 7.1.1*)
2. The variance of the sample mean $\overline{X}$ is given by
$$Var(\overline{X}) = \frac{\sigma^2}{n}$$
3. So, the standard deviation of $\overline{X}$, denoted by $\sigma_{\overline{X}}$, is given by
$$St.\ Dev(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{6.2}$$
Definition. The standard deviation $\sigma_{\overline{X}}$ is also called **standard error**.

6.2 Central Limit Theorem

The following theorem describes the sampling distribution of the Sample Mean. It is called the **Central Limit Theorem** (CLT). It is truly central in terms of applicability of statistics.

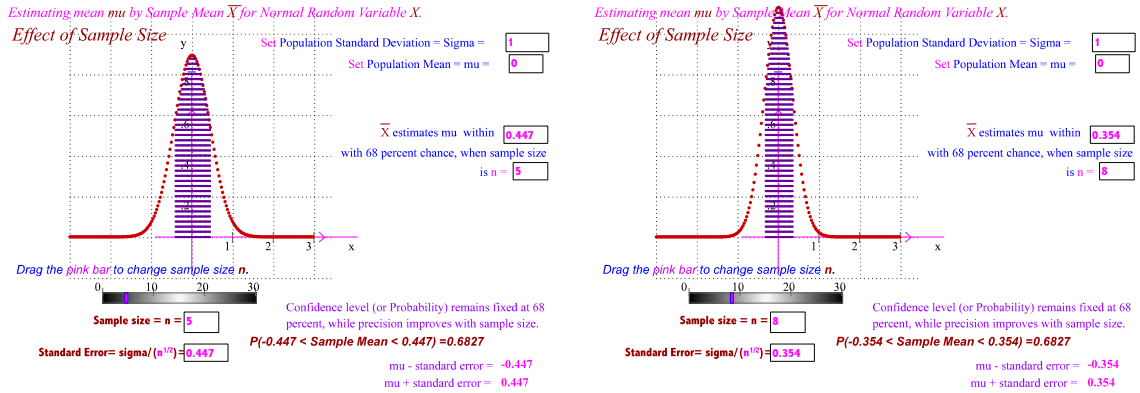
Theorem (CLT). Suppose $X_1, X_2, \cdots, X_n$ is a sample from a population X, with mean μ and variance σ^2. Assume n is large. Then the sample mean $\overline{X}$ is, approximately, distributed as

$$\overline{X} \sim N\left(\mu, \sigma_{\overline{X}}\right) \qquad \text{where} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

So, as in the case of Binomial approximation (5.8),

$$\begin{cases} P(a \le \overline{X} \le b) = P\left(\frac{a-\mu}{\sigma_{\overline{X}}} \le \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \le \frac{b-\mu}{\sigma_{\overline{X}}}\right) \\ \approx P\left(\frac{a-\mu}{\sigma_{\overline{X}}} \le Z \le \frac{b-\mu}{\sigma_{\overline{X}}}\right) = P\left(L \le Z \le U\right) \end{cases} \quad \text{with} \quad \begin{cases} \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \\ L = \frac{a-\mu}{\sigma_{\overline{X}}} \\ U = \frac{b-\mu}{\sigma_{\overline{X}}} \end{cases} \tag{6.3}$$

If $X \sim N(\mu, \sigma)$, then this is exact probability, not just approximate probability.



These two diagrams show that, the probability mass $p = .6827$ is concentrated more closely to the mean (zero), when the sample size n increases, from $n = 5$ to $n = 8$.

6.2.1 Problems: on Central Limit Theorem

Exercise 6.2.1. It is known that the tuition paid per semester by students in a univer-

sity has a distribution with mean \$2,050 and standard deviation \$310. If 64 students are interviewed, what is the approximate probability that the sample mean tuition paid will be above \$2,060?

Solution.

Here the population mean $\mu = 2050$ and the population standard deviation $\sigma = 310$.
The sample size $n = 64$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 2050$
and the standard deviation (error) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{310}{\sqrt{64}} = 38.75$
Let $X =$ Tuition paid by the students.
Then, the distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(2050, 38.75)$

Now " $\overline{X}$ will be above 2060" means " $X > 2060$".
$P(2060 < \overline{X}) = P\left(\frac{2060-\mu}{\sigma_{\overline{X}}} < \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}\right)$
$\approx P\left(\frac{2060-\mu}{\sigma_{\overline{X}}} < Z\right) = P\left(\frac{2060-2050}{38.75} < Z\right) = P(.2580, < Z) = normalcdf(.2580, 5) = .3982$

Exercise 6.2.2. The monthly water consumption X per household in a subdivision in Kansas City has normal distribution with mean 15000 gallons and standard deviation 3000 gallons. What is the probability that the mean consumption of the 44 households in the subdivision will exceed 16000 gallons?

Solution.

Here the population mean $\mu = 15000$ and the population standard deviation $\sigma = 3000$.
The sample size $n = 44$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 15000$.
and the standard deviation (error) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3000}{\sqrt{44}} = 452.2670$
Let $X =$ monthly water consumption by the households.
Then, the distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(15000, 452.2670)$

Now " $\overline{X}$ will exceed 16000" means " $\overline{X} > 16000$".
$P(16000 < \overline{X}) = P\left(\frac{16000-\mu}{\sigma_{\overline{X}}} < \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}\right) \approx P\left(\frac{16000-\mu}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{16000-15000}{452.2670} < Z) = P(2.2111 < Z) = normalcdf(2.2111, 5) = .0135$

Exercise 6.2.3. In a class of more than thousand students, the instructor announced after a test that the mean score was $\mu = 77$ points and standard deviation $\sigma = 24$ points. You took a sample of 81 students. What would be the approximate probability that the sample mean would be less than 80?

Solution.

Here the population mean $\mu = 77$ and the population standard deviation $\sigma = 24$.
The sample size $n = 81$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 77$,
and the standard deviation (error) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{81}} = 2.6667$
Let $X =$ Points scored by students.
the distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(77, 2.6667)$.

Now "the sample mean would be less than 80" means "$\overline{X} < 80$".
$P(\overline{X} < 80) = P\left(\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} < \frac{80-\mu}{\sigma_{\overline{X}}}\right) \approx P\left(Z < \frac{80-\mu}{\sigma_{\overline{X}}}\right)$
$= P(Z < \frac{80-77}{2.6667}) = P(Z < 1.1250) = normalcdf(-5, 1.1250) = .8697$

Exercise 6.2.4. The mean salary X of the university professors in a state is $\mu = \$65,000$ and standard deviation $\sigma = \$14,000$. You collect a sample of 75 professors. What is the probability that sample mean salary of these 75 professors would be above \$60,000.

Solution.

Here the population mean $\mu = 65000$ and the population standard deviation $\sigma = 14000$.
The sample size $n = 75$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 65000$, athe standard deviation (error)
$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{14000}{\sqrt{75}} = 1616.5808$
Let $X =$ monthly water consumption by the households.
he distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(65000, 1616.5808)$.

Now " $\overline{X}$ would be above 60,000" means " $\overline{X} > 60000$".
$P(16000 < \overline{X}) = P\left(\frac{16000-\mu}{\sigma_{\overline{X}}} < \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}\right) = P\left(\frac{16000-\mu}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{60000-65000}{1616.5808} < Z) = P(-3.0929, < Z) = normalcdf(-3.0929, 5) = .9990$

Exercise 6.2.5. The time X that a child spends watching TV on week- ends has a normal distribution with mean $\mu = 330$ minutes and standard deviation $\sigma = 95$ minutes. You sample 50 kids in a school. What is the probability that the sample mean time $\overline{X}$ that these kids watch TV on a weekend will be less than 300 minutes.

The Following Problems are Posed in terms of the Total

Exercise 6.2.6. The weight X of fish in a lake has mean $\mu = 12$ pounds and standard deviation $\sigma = 4.5$ pounds. Suppose you catch 150 fish. What is the probability that total

weight of fish will be less than 1900 pounds?

Solution.

Here the population mean $\mu = 12$ and the population standard deviation $\sigma = 4.5$.
The sample size $n = 150$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 12$,
and the standard deviation (error) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4.5}{\sqrt{150}} = .3674$
Let X = weight of fish.
Then, the distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(12, .3674)$.

The problem is posed in terms of Total weight of all the fish.
The sample mean $\overline{X} = \frac{Total}{n}$.

Now " Total weight will be less than 1900 pounds" means that "$Total < 1900$".
This means " $\overline{X} = \frac{Total}{n} < \frac{1900}{n} = \frac{1900}{150} = 12.6667$".

$P(\overline{X} < 12.6667) = P\left(\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} < \frac{12.6667-\mu}{\sigma_{\overline{X}}}\right) \approx P\left(Z < \frac{12.6667-\mu}{\sigma_{\overline{X}}}\right)$
$= P(Z < \frac{12.6667-12}{.3674}) = P(Z < 1.8146) = normalcdf(-5, 1.8146) = .9652$
(*Now, you are fairly sure (96 percent sure) that you did not catch 1900 pounds.*)

Exercise 6.2.7. The waiting time for the campus bus has a mean $\mu = 7$ minutes and the standard deviation $\sigma = 2$ minutes. A student used the bus 120 times in a month. What is the probability that the student would have waited more than 900 minutes during the whole month?

Solution.

Here the population mean $\mu = 7$ and the population standard deviation $\sigma = 2$.
The sample size $n = 120$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 7$,
and the standard deviation (error) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{120}} = .1826$
Let X = waiting time for the bus.
Then, the distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(7, .1826)$
The problem is posed in terms of Total weight of a ll the fish.
The sample mean $\overline{X} = \frac{Total}{n}$.
Now " total of more than 900 minutes will be spent" means that "$Total > 900$".
This means"$\overline{X} = \frac{Total}{n} > \frac{900}{120} = 7.5$".

$P(7.5 < \overline{X}) = P\left(\frac{7.5-\mu}{\sigma_{\overline{X}}} < \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}\right) \approx= P\left(\frac{7.5-\mu}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{7.5-7}{.1826} < Z) = P(2.7382, < Z) = normalcdf(2.7382, 5) = .0031.$
(*Now you know, the chances are fairly low that you will spend that kind of time waiting for the bus.*)

Exercise 6.2.8. According to some data, the annual Kansas wheat export X has a mean 733 million dollars and standard deviation 163 million dollars. What is the probability that over the next 10 years Kansas wheat exports will exceed 8040 million dollars?

Solution.

Here the population mean $\mu = 733$ and the population standard deviation $\sigma = 163$.
The sample size $n = 10$.
First step is to compute the mean
$\mu_{\overline{X}} = \mu = 733,$
and the standard deviation (error) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{163}{\sqrt{10}} = 51.5451$
Let $X =$ Kansas wheat export annually.
Then, the distribution of $\overline{X}$ is, approximately, $\overline{X} \sim N(733, 51.5451)$

The problem is posed in terms of Total export in 10 years.
The sample mean $\overline{X} = \frac{Total}{n}$.
Now " Total export will exceed 8040" means that "$Total > 8040$".
This means "$\overline{X} = \frac{Total}{n} > \frac{8040}{n} = \frac{8040}{10} = 804$".
$P(804 < X) = P\left(\frac{804-\mu}{\sigma_{\overline{X}}} < \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}\right) \approx P\left(\frac{804-\mu}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{804-733}{51.5451} < Z) = P(1.3774, < Z) = normalcdf(1.3774, 5) = .0842$

6.3 Sampling Distribution of the Sample Proportion

Suppose you are a statistical quality control (SQC) officer in a lamp factory. Your job would include estimating proportion p of the defective lamps. When you test a lamp, it is a $Bernoulli(p)$-trial. Correspondingly, a $Bernoulli(p)$ random variable X is define as follows:

$$X = \begin{cases} 1 & if \textbf{ success } (i.e.\ defective) \\ 0 & if \textbf{ failure } (i.e.\ not\ defective) \end{cases}$$

From chapter 4, we know that

$$\begin{cases} \text{mean of } X & \mu = p \\ \text{st. dev. of } X & \sigma = \sqrt{p(1-p)} \end{cases}$$

In subsequent chapters, we would try to estimate p. As usual, we will use a sample mean to estimate the mean $\mu = p$. So, we take a sample $X_1, X_2, \ldots, X_n$ of size n from this $Bernoulli(p)$ population and X_i represents the outcome of testing the i lamp as follows:

$$X_i = \begin{cases} 1 & if\ i^{th}\ lamp\ is\ a\ \textbf{success}\ (i.e.\ defective) \\ 0 & if\ i^{th}\ lamp\ is\ a\ \textbf{failure}\ (i.e.\ not\ defective) \end{cases}$$

An estimator of $\mu = p$ would be the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{T}{n} \quad \text{where} \quad T = X_1 + X_2 + \cdots + X_n$$

Since X_i is 1 or 0 according as the i^{th} trial is success or failure (i.e. the ith sample lamp is defective or not), we have the following:

$$\begin{cases} T = & \textbf{the total Number of Success}\ \text{ in these } n \text{ trials} \\ \overline{X} = \frac{T}{n} = & \textbf{Sample proportion success} \end{cases}$$

Theorem. Let p be the proportion of a population with a certain attribute. Out of a sample of size n, suppose T have the attribute (or is the number of success). Let $\overline{X} = \frac{T}{n}$ be the **proportion of success**. If n is large and p is not too close to 0 or 1, then by CLT, when n is large, the distribution of $\overline{X}$, is approximately,

$$\overline{X} \sim N(p, \sigma_{\overline{X}}) \qquad \text{where} \qquad \sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} \tag{6.4}$$

There is obviously nothing special about the testing lamps and estimating proportion p of defective lamps produced in the factory. The above applies to any situation of $Bernoulli(p)$-trials. Other examples would be estimating

1. proportion p of the voter population who favors a particular candidate,
2. proportion p of the population who suffers from asthma
3. proportion p of the grass seeds of a variety that germinates in a particular situation
4. proportion p of the population who benefit from a particular vaccine
5. proportion p of the population who live beyond 70.

Theorem. With notations, as above, the distribution of $\overline{X} \sim N(p, \sigma_{\overline{X}})$ is normal, as in (6.4). As in (6.3, 5.8), for $0 \le a \le b \le 1$, we have

$$\begin{cases} P(a \le \overline{X} \le b) = P\left(\frac{a-p}{\sigma_{\overline{X}}} \le \frac{\overline{X}-p}{\sigma_{\overline{X}}} \le \frac{b-p}{\sigma_{\overline{X}}}\right) \\ \approx P\left(\frac{a-p}{\sigma_{\overline{X}}} \le Z \le \frac{b-p}{\sigma_{\overline{X}}}\right) = P\left(L \le Z \le U\right) \end{cases} \quad \text{with} \quad \begin{cases} \sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} \\ L = \frac{a-\mu}{\sigma_{\overline{X}}} \\ U = \frac{b-\mu}{\sigma_{\overline{X}}} \end{cases} \tag{6.5}$$

6.3.1 Problems: on Sample Proportion

Exercise 6.3.1. According to a report entitled "Pediatric Nutrition Surveillance" published by Centers for Disease Control (CDC) 18 percent of the children younger than two had anemia in 1997. On a particular day in that year, a pediatrician examined 180 children. What is the probability that the proportion will exceed 0.20? (Equivalently, *this would be the probability that the number T of children with anemia would exceed* $180 * .2 = 36$.)

Solution.

Here the population mean $p = .18$ and the sample size $n = 180$.

Here $\overline{X}$ = the sample proportion of patients with anemia.

First step is to compute the mean

$\mu_{\overline{X}} = p = .18$

and the standard deviation of $\overline{X}$

$\sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.18(1-.18)}{180}} = .028636$

(We take unto six decimal points, because we are already working with small numbers.)

Approximately, $\overline{X} \sim N(.18, .028636)$ has normal distribution.

Now "$\overline{X}$ will exceed 0.20" means $\overline{X} > .20$".

$P(.20 < \overline{X}) = P\left(\frac{.20-p}{\sigma_{\overline{X}}} < \frac{\overline{X}-p}{\sigma_{\overline{X}}}\right) \approx P\left(\frac{.20-p}{\sigma_{\overline{X}}} < Z\right)$

$= P(\frac{.20-.18}{.028636} < Z) = P(.6984 < Z) = normalcdf(.6984, 5) = .2625$

Exercise 6.3.2. In 1996, the House of Representatives impeached President Clinton. As a part of the political discourse, numerous polls were conducted and reported. One poll claimed that 75 percent of eligible voters think the President should not be impeached. Suppose 700 voters were interviewed. Assuming the claim, what would be the probability that less than 72 percent (in this sample of 700) would have thought the President should not be impeached. (Equivalently, *this would be the probability that less than* $.72 * 700 = 504$ *voters would have thought the President should not be impeached.*)

Solution. Here the population mean $p = .75$ and the sample size $n = 700$.

Denote $\overline{X}$ = the sample proportion of voters who thought that the President should not be impeached.

First step is to compute the mean

$\mu_{\overline{X}} = p = .18,$

and the standard deviation of $\overline{X}$

$\sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.75(1-.75)}{700}} = .016366.$

(We take up to six decimal points, because we are already working with small numbers.)

Approximately, $\overline{X} \sim N(.75, .016366)$ has normal distribution.

Now " less than 72 percent would have thought the President should not be impeached" means that "$\overline{X} < .72$".
$P(X < .72) = P\left(\frac{\overline{X}-p}{\sigma_{\overline{X}}} < \frac{.72-p}{\sigma_{\overline{X}}}\right) \approx= P\left(Z < \frac{.72-p}{\sigma_{\overline{X}}}\right)$
$= P\left(Z < \frac{.72-.75}{.016366}\right) = P(Z < -1.8331) = normalcdf(-5, -1.8331) = .0334$

Exercise 6.3.3. It is believed proportion of voters (in a county) who vote by absentee ballot is p=.22. You sample 725 voters. Compute an approximate the probability the sample proportion of absentee votes will exceed 25 percent. (Equivalently, *this would be the probability that the number of absentee votes will exceed* $725 * .22 = 159.5$.)

Solution.
Here the population mean $p = .22$ and the sample size $n = 725$.
Denote $\overline{X}$ = the sample proportion of absentee votes.
First step is to compute the mean
$\mu_{\overline{X}} = p = .22$,
and the standard deviation of $\overline{X}$
$\sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.22(1-.22)}{725}} = .015385$
(We take unto six decimal points, because we are already working with small numbers.)
Approximately, $\overline{X} \sim N(.22, .015385)$ has normal distribution.

Now "sample proportion of absentee votes will exceed 25 percent"
means $\overline{X}$ will exceed 0.25". That means $\overline{X} > .25$".
$P(.25 < \overline{X}) = P\left(\frac{.25-p}{\sigma_{\overline{X}}} < \frac{\overline{X}-p}{\sigma_{\overline{X}}}\right) \approx= P\left(\frac{.25-p}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{.25-.22}{.015385} < Z) = P(1.9500, < Z) = normalcdf(1.9500, 5) = .0256$

Exercise 6.3.4. It is believed that 35 percent of the population in a county shop in health food market. If you sample 800 individuals, what would be an approximate the probability the sample proportion of those who shop in health food market exceed 40 percent. (Equivalently, *this would be the probability that the number T of those who shop in health food market would exceed* $800 * .40 = 320$.)

Solution.
Here the population mean $p = .35$ and the sample size $n = 800$. Denote, $\overline{X}$ = the sample proportion those who shop in health food market.
First step is to compute the mean
$\mu_{\overline{X}} = p = .35$,

and the standard deviation of $\overline{X}$
$\sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.35(1-.35)}{800}} = .016863$
(We take unto six decimal points, because we are already working with small numbers.)
Approximately, $\overline{X} \sim N(.35, .016863)$has normal distribution.

Now "sample proportion of those who shop in health food market will exceed 40 percent"
means $\overline{X}$ will exceed 0.40". That means $\overline{X} > .40$".
$P(.40 < \overline{X}) = P\left(\frac{.40-p}{\sigma_{\overline{X}}} < \frac{\overline{X}-p}{\sigma_{\overline{X}}}\right) \approx P\left(\frac{.40-p}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{.40-.35}{.016863} < Z) = P(2.9651, < Z) = normalcdf(2.9651, 5) = .0015$

Exercise 6.3.5. It is known that a vaccine may cause fever as side effect, after one takes the shot. The producer of the vaccine claims that only 17 percent of those who take the shot experience such side effects. You sample 978 individuals who took the shot. What would be an approximate probability that more than 15 percent would experience side effect? (Equivalently, *this would be the probability that more than* $.15 * 978 = 146.7$ *would experience side effect.*)

Solution.
Here the population mean $p = .11$ and the sample size $n = 978$.
Denote $\overline{X}$ = the sample proportion those who experienced sided effect.
First step is to compute the mean
$\mu_{\overline{X}} = p = .17$,
and the standard deviation of $\overline{X}$
$\sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.17(1-.17)}{978}} = .012011$
(We take up to six decimal points, because we are already working with small numbers.)
Approximately, $\overline{X} \sim N(.17, .012011)$ has normal distribution.

Now "more than 15 percent would experience side effect "
means $\overline{X}$ will be more than .15". That means $\overline{X} > .15$".
$P(.15 < \overline{X}) = P\left(\frac{.15-p}{\sigma_{\overline{X}}} < \frac{\overline{X}-p}{\sigma_{\overline{X}}}\right) \approx= P\left(\frac{.15-p}{\sigma_{\overline{X}}} < Z\right)$
$= P(\frac{.15-.17}{.012011} < Z) = P(2.9651, < Z) = normalcdf(-1.6651, 5) = .9521$

Exercise 6.3.6. About 27 percent of the population take flu shots. You are in a class of 750 students. Compute an approximate the probability the sample proportion of those who took the shot would be less than 25 percent. (Equivalently, *this would be the probability that the number T of those who took the shot would be less than* $.25 * 750 = 167.5$.)

Exercise 6.3.7. It is known that 78 percent of the microwave ovens last more than five years. A SQC inspector sampled 600 microwaves. What would be the approximate probability that more than 78 percent of this sample would last more than five years? (Equivalently, *this would be the probability that more than* $.78 * 600 = 468$ *of this sample would last more than five years.*)

Chapter 7

Estimation

The **objective of this course** has been to develop methods to use **sample statistics** T (e.g. sample mean $\overline{X}$, sample standard deviation S, sample proportion of success $\overline{X}$) to estimate **population parameters** θ (e.g. *mean* μ*, standard deviation* σ*, population proportion* p). Finally, We are ready to do the so, in this chapter. We use the sampling distributions, from chapter 6, of sample mean $\overline{X} \sim N(\mu, \sigma_{\overline{X}})$, and of sample proportion of success $\overline{X} \sim N(p, \sigma_{\overline{X}})$, (*and the sampling distributions of other such statistics*) to develop methods to estimate the corresponding population parameters.

We consider two methods of estimating parameters.

1. The first one is called **point estimation**. In point estimation, a number t is given as an estimate for the parameter θ. Most people are used to the concept of point estimation. For example, to estimate the mean annual income μ of the US population, one is used to the idea of taking a sample of a certain size, compute the sample mean annual income $\overline{x}$, and call it an estimate for μ.

2. The second one is called **interval estimation**. In interval estimation, an interval (L, U) is given as a range where the parameter θ is estimated to be within. Most of the interval estimations we consider would consist of

   (a) a point estimate t fot θ and

   (b) an estimate e for the error (**precision**) of this estimation.

   Correspondingly, θ would be estimated to be within the interval $(t - e, t + e)$. For example, when estimating the mean annual income μ of the US population, a statistician may take a sample, compute the sample mean $\overline{x}$ and say that the population mean μ is estimated to be within the $(\overline{x} - 1000, \overline{x} + 1000)$. Obviously, smaller error (i.e. **higher precision**) would always be more desirable. Statistical estimation would always come with information about how often the actual error $\varepsilon = |\theta - T|$ would remain within a specified error limit e. This is given in terms of probability $\mathbf{P(\varepsilon \leq e)}$. This probability $P(\varepsilon \leq e)$ would be called the **level of confidence**, in

the later sections. Ideally, we would like to estimate θ by T, with high precision (i.e, low error) with high level of confidence.

7.1 Point and Interval Estimation

In statistical point estimation, a statistic T is used to estimate a parameter θ. Corresponding to a sample, the value t of T would be called a **point estimate** of θ. The statistic T would be called an **estimator** of θ.

For example, the sample mean $\overline{X}$ would be an estimator of the population mean μ and the computed value $\overline{X} = \overline{x}$, corresponding to a sample, would be a point estimate of μ. Similarly, the sample variance S^2 is an estimator of the population variance σ^2 and the computed value $S^2 = s^2$, corresponding to a sample, is a point estimate of σ^2. It would be a common instinct to accept that the sample mean $\overline{X}$ would be a natural estimator for the population mean μ and the sample variance S^2 would be a natural estimator for the population variance σ^2.

7.1.1 Criterion for Good Estimators (read only)

There are some established mathematical criteria, to justify and to decide, why or when an estimator T would be a good candidate as an estimator for a population parameter θ. The most basic such criterion, is called the unbiasedness as defined below.

Definition. A statistic T is said to be an **unbiased estimator** of a population parameter θ, if the mean of T is equal to θ, in other words, the expected value $E(T) = \theta$. In fact, it can be proved that

$$E(\overline{X}) = \mu, \qquad \text{and} \qquad E(S^2) = \sigma^2 \tag{7.1}$$

Therefore, the sample mean $\overline{X}$ is an unbiased estimator of the population mean μ, and the sample variance S^2 is an unbiased estimator of the population variance σ^2.

Remark. Recall the definition of the sample variance, in (2.2) of s^2, and (6.1) of S^2. In these definitions, justification for having the denominator $n-1$, instead of n, was to ensure $E(S^2) = \sigma^2$, so that S^2 is an unbiased estimator of σ^2.

Remark. Other than unbiasedness, there are other criterion, for good estimators.We outline one of them. Recall (6.2)

$$St.\ Dev(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{decreases to 0 as the sample size } n \text{ increases.}$$

Because of this, as two diagrams in chapter 6 indicates, the values of $\overline{X}$ would be close to the mean μ, more frequently (higher probability), if and when the sample size n is larger. This is another justification why the sample mean $\overline{X}$ would be a good estimator

for mean μ. They say $\overline{X}$ is a **consistent estimator** of μ, for this reason. It means, the estimator improves, when sample size n is larger. In the next subsection, this property is use (implicitly), when we consider **confidence level**.

7.1.2 Interval Estimation

Almost never, a point estimate $T = t$ of a parameter θ would be exactly equal to the actual value of θ. This is why, it would be more reasonable to provide an interval (L, U) where the parameter θ would estimated to be within. Here L, U would be statistics. The values of $L = l$, $U = u$ would depend on the samples. While θ would be estimated to be within (l, u), for some samples, the true value of θ would sometimes be left outside the interval (l, u). **We would be happy as long as the true value of θ falls within the interval (l, u) most often (or often enough), allowing the possibility of being "wrong" a few times**.

The next question would be, how often is often enough? The probability $P(L \leq \theta \leq U)$ is exactly the proportion of times when θ would fall within the interval (l, u). The interval estimation of θ, formally defined below, provides an interval (L, U) together with the probability $\mathbf{P(L \leq \theta \leq U)}$.

Definition. Let θ be a population parameter. An interval estimate for θ provides the following:

1. It gives an interval (L, U) as an estimate for θ. Here L, U are statistics or $L = -\infty$ or $U = \infty$. It also gives the probability $\mathbf{P(L \leq \theta \leq U)}$, to be called the **level of confidence**.

   In summary, the confidence interval, with level of significance $1 - \alpha$ is defined by the equation:

   $$P(L \leq \theta \leq U) = 1 - \alpha \tag{7.2}$$

2. The interval, (L, U) is said to be a $(1 - \alpha)100$ percent confidence interval of θ.

3. In practice, α will be a small number, like, 0.1, 0.01, 0.05. That means, 90, 95 or 99 percent confidence intervals of θ are commonly considered.

4. Three different types of confidence intervals are commonly considered.

5. When both L and U are statistics, the confidence interval (L, U) would be called a **two sided confidence interval** of θ.

6. When $L = -\infty$ and U is a statistic, the confidence interval $(L, U) = (-\infty, U)$ would be called a **one sided left confidence interval** of θ.

7. When L is a statistic and $U = \infty$, the confidence interval $(L, U) = (L, \infty)$ would be called a **one sided right confidence interval** of θ.

8. In this course, we only consider two sided confidence intervals.

7.1.3 Procedure to Construct Confidence Interval

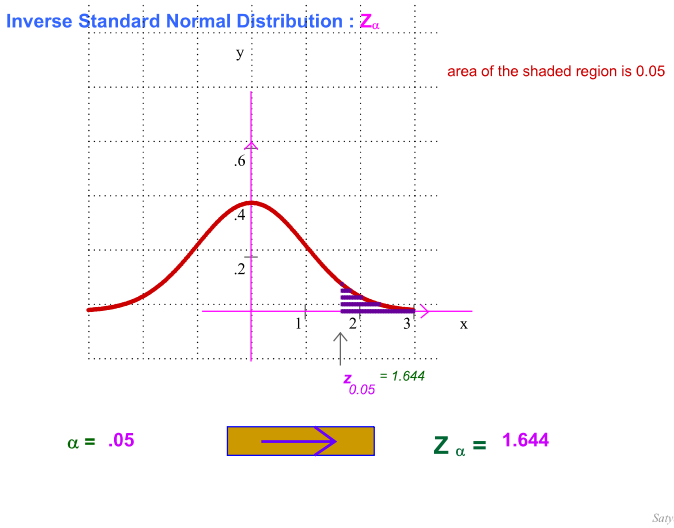
In this subsection, we will construct a two sided confidence interval of the population mean $\mathfrak{m}$. As you saw, till chapter 6, we (mostly) computed probability, while the values of the parameters like mean μ were given.

The situation reverses now. As was the goal of this course, at this stage, the value of the parameters like mean μ would have to be estimated and the probability will be given (or specified) in terms of the level of confidence. In section 5.2.2, we discussed inverse probability (or cut-off), which becomes useful now. To construct confidence intervals, the following cut-off numbers for standard normal random variables $Z \sim N(0,1)$ would be useful.

Definition. As usual $Z \sim N(0,1)$ would denote the standard normal random variable. Given a number $0 < \alpha < 1$, the number z_α is defined by the formula

$$P(z_\alpha \leq Z) = \alpha. \quad \text{Equivalently,} \quad \begin{cases} P(Z \leq z_\alpha) = 1 - \alpha, & \text{Or,} \\ \mathbf{z_\alpha = invNorm(1-\alpha)} \end{cases} \tag{7.3}$$

These numbers z_α are also called **Critical values**. There are many internet sites that would give these numbers.

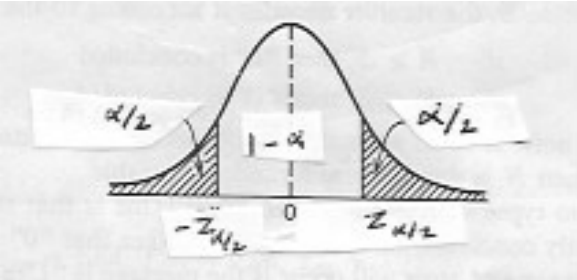


To explain the procedure to develop a confidence interval (of any parameter θ), we will construct a $(1-\alpha)100$ percent confidence interval for the mean μ when the standard deviation σ is known, as follows.

Suppose X is a random variable with mean μ and known standard deviation σ. Let $X_1, X_2, \ldots, X_n$ be a sample from X-population. By CLT, approximately,

$$Z = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0,1) \quad \text{has a st. normal distribution.}$$

So, $\quad P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1-\alpha \quad OR, \qquad P\left(-z_{\alpha/2} \leq \dfrac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \leq z_{\alpha/2}\right) = 1-\alpha$



Simplifying,

$$P\left(\overline{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = 1-\alpha \tag{7.4}$$

By Equation (7.2), this is exactly what is needed for a confidence interval, with $(1-\alpha)100$ percent confidence interval. We state the same, as in the theorem below.

Theorem. We use the notations as above. Assume that σ is known. A $(1-\alpha)100$ (approximate) percent confidence interval, for μ, is given by

$$\overline{X} - E \leq \mu \leq \overline{X} + E \qquad \text{where} \quad E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

If X is normal, then it is an exact confidence interval for μ (not just an approximation). We add few more definitions, to summarize this:

$$\begin{cases} \mathbf{E} = \frac{\mathbf{z_{\alpha/2}\sigma}}{\sqrt{\mathbf{n}}} & \text{is called the Margin of Error (MOE)} \\ LEP = \overline{X} - E & \text{is called the Left end point} \\ REP = \overline{X} + E & \text{is called the Right end point} \\ l = 2E & \text{is the total length of the confidence interval.} \end{cases}$$

Informally, this confidence interval is also called a Z-interval.

As was defined above, this will be a two-sided confidence interval of μ. A $(1-\alpha)100$ percent confidence intervals is interpreted as, if one computes $(1-\alpha)100$ percent confidence intervals on a regular basis, the true value of μ will be within the confidence interval $(1-\alpha)100$ percent times and it outside the interval $\alpha 100$ percent times.

The Required Sample Size: If we want to increase the level of confidence, the Margin of error(MOE) E would increase and conversely. Both can be controlled by increasing the

sample size n. The required sample size n can be determined by solving the formula for the MOE above. Given a specified MOE E and a level of confidence $(1-\alpha)$, sample size n needed is given by

$$\mathbf{n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2} \qquad \text{always round upward.} \tag{7.5}$$

Note that the formula depends on σ, but on on μ or the sample.

7.1.4 Problems: on Z-intervals for μ

Exercise 7.1.1. Suppose X is a normal population with mean μ and standard deviatio $\sigma = 15$. A sample of size 25 and the sample mean $\overline{X}$ was found to be 81. Compute a 99 percent confidence interval fo μ nd give the MOE, LEP, REP.

Solution.
Here the population standard deviation $\sigma = 15$,
the sample size $n = 25$, the sample mean $\overline{X} = 81$,
Level of confidence = 99 percent. So
$1-\alpha = .99$, $\alpha = .01$ and $\frac{\alpha}{2} = .005$.
Therefore, $z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.5758*15}{\sqrt{25}} = 7.7274$
$LEP = \overline{X} - E = 81 - 7.7274 = 73.2726$
$REP = \overline{X} + E = 81 + 7.7274 = 88.7274$

Exercise 7.1.2. The weight distribution X of a group of students is normal with standard deviation $\sigma = 9.8$. We estimate the mean weight μ, in three situations.

1. A sample of size 14 was collected and the sample mean $\overline{X}$ was found to be 151.1. Find a 99 percent confidence interval for μ and give the margin of error (MOE), LEP, REP.

2. (Change the Confidence level.) A sample of size 14 was collected and the sample mean $\overline{X}$ was found to be 151.1. Find a 90 percent confidence interval for μ and give the margin of error (MOE), LEP, REP.

3. (Change the sample size.) A sample of size 78 was collected and the sample mean $\overline{X}$ was found to be 151.1. Find a 90 percent confidence interval for μ and give the margin of error (MOE), LEP, REP.

Solution.

1. We solve (1).
   Here the population standard deviation $\sigma = 9.8$,
   the sample size $n = 14$,
   the sample mean $\overline{X} = 151.1$,
   Level of confidence = 99 percent. So
   $1 - \alpha = .99, \quad \alpha = .01, \quad \frac{\alpha}{2} = .005.$
   Therefore, $z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

   $MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.5758*9.8}{\sqrt{14}} = 6.7464$
   $LEP = \overline{X} - E = 151.1 - 6.7464 = 144.3536$
   $REP = \overline{X} + E = 151.1 + 6.7464 = 157.8464$

2. We solve (2).
   Here the population standard deviation $\sigma = 9.8$,
   the sample size $n = 14$,
   the sample mean $\overline{X} = 151.1$,
   Level of confidence = 90 percent. So
   $1 - \alpha = .90, \quad \alpha = .10, \quad \frac{\alpha}{2} = .05.$
   Therefore, $z_{\alpha/2} = z_{.05} = invNorm(.95) = 1.6449$

   $MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.6449*9.8}{\sqrt{14}} = 4.3083$
   $LEP = \overline{X} - E = 151.1 - 4.3083 = 146.7917$
   $REP = \overline{X} + E = 151.1 + 4.3083 = 155.4083$
   Note that the length and/or MOE has decreased, due to the decline in confidence level.

3. We solve (3).
   Here the population standard deviation $\sigma = 9.8$,
   the sample size $n = 78$,
   the sample mean $\overline{X} = 151.1$,
   Level of confidence = 90 percent. So
   $1 - \alpha = .90, \quad \alpha = .10, \quad \frac{\alpha}{2} = .05.$
   Therefore, $z_{\alpha/2} = z_{.05} = invNorm(.95) = 1.6449$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.6449*9.8}{\sqrt{78}} = 1.8252$
$LEP = \overline{X} - E = 151.1 - 1.8252 = 149.2748$
$REP = \overline{X} + E = 151.1 + 1.8252 = 152.9252$
Note that the length and/or MOE has decreased, due to the increase in sample size.

Exercise 7.1.3. The time taken by an athlete to run an event is normally distributed with mean μ and known standard deviation $\sigma = 3.5$ seconds. To estimate the mean μ, he ran 16 times and the sample mean was found to be $\overline{X} = 33$ seconds. Find a 95 percent confidence interval for true mean time μ and give the margin of error (MOE).

Solution.
Here the population standard deviation $\sigma = 3.5$,
the sample size $n = 16$,
the sample mean $\overline{X} = 33$,
Level of confidence = 95 percent. So
$1 - \alpha = .95, \quad \alpha = .05, \quad \frac{\alpha}{2} = .025.$
Therefore, $z_{\alpha/2} = z_{.05} = invNorm(1 - .025) = 1.9600$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.9600*3.5}{\sqrt{16}} = 1.715$
$LEP = \overline{X} - E = 33 - 1.715 = 31.285$
$REP = \overline{X} + E = 33 + 1.715 = 34.715$

Exercise 7.1.4. Let X represent the monthly consumption of electricity (in KWH , in a winter month) by the households in a county. From past experience, it is known that the standard deviation is $\sigma == 150$ KWH. To estimate the mean monthly consumption μ, a sample of size 115 was collected and the sample mean $\overline{X}$ of the monthly consumption was found to be 874. Find a 98 percent confidence interval for μ and give the margin of error (MOE).

Solution.
Here the population standard deviation $\sigma = 3.5$,
the sample size $n = 115$,
the sample mean $\overline{X} = 874$,
Level of confidence = 98 percent. So
$1 - \alpha = .98, \quad \alpha = .02, \quad \text{and} \quad \frac{\alpha}{2} = .01.$
Therefore, $z_{\alpha/2} = z_{.01} = invNorm(1 - .01) = 2.3263$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.3263*150}{\sqrt{115}} = 32.5393$

$LEP = \overline{X} - E = 874 - 32.5393 = 841.4607$
$REP = \overline{X} + E = 874 + 32.5393 = 906.5393$

Exercise 7.1.5. The monthly gas consumption by an individual in a city is normally distributed with mean μ and known standard deviation $\sigma = 7.5$ gallons. To estimate the mean μ, a sample of 33 individual collated and their mean consumption $\overline{X}$ was found to be 56 gallons. Find a 96 percent confidence interval for true mean consumption μ and give the margin of error (MOE).

Solution.
Here the population standard deviation $\sigma = 7.5$,
the sample size $n = 33$,
the sample mean $\overline{X} = 56$,
Level of confidence = 96 percent. So $1 - \alpha = .96, \quad \alpha = .04, \quad$ and $\quad \frac{\alpha}{2} = .02$.
Therefore, $z_{\alpha/2} = z_{.02} = invNormal(1 - .02) = 2.0537$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.0537*7.5}{\sqrt{33}} = 2.6813$
$LEP = \overline{X} - E = 56 - 2.6813 = 53.3187$
$REP = \overline{X} + E = 56 + 2.6813 = 58.6813$

Exercise 7.1.6. The monthly cell phone minutes used by an individual in a city has a mean μ and known standard deviation $\sigma = 350$ minutes. To estimate the mean μ, a sample of 780 users was collected. The sample mean $\overline{X}$ was found to be 2222. Find a 97 percent confidence interval for true mean cellphone minutes μ and give the margin of error (MOE).

Solution.
Here the population standard deviation $\sigma = 350$,
the sample size $n = 780$,
the sample mean $\overline{X} = 2222$ minutes,
Level of confidence = 97 percent. So
$1 - \alpha = .97, \quad \alpha = .03, \quad$ and $\quad \frac{\alpha}{2} = .015$.
Therefore, $z_{\alpha/2} = z_{.015} = invNorm(1 - .015) = 2.1701$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{2.1701*350}{\sqrt{780}} = 27.1957$
$LEP = \overline{X} - E = 2222 - 27.1957 = 2194.8043$
$REP = \overline{X} + E = 2222 + 27.1957 = 2249.1957$

Exercise 7.1.7. The diameter of pumpkins in a farm has a mean μ and known standard deviation $\sigma = 13$ inches. To estimate the mean μ, a sample of 460 pumpkins was collected. The sample mean $\overline{X}$ was found to be 29 inches. Find a 80 percent confidence interval for true mean diameter μ and give the margin of error (MOE).

Solution.

Here the population standard deviation $\sigma = 13$,
the sample size $n = 460$,
the sample mean $\overline{X} = 29$
Level of confidence = 80 percent. So,
$1 - \alpha = .80, \quad \alpha = .20, \quad$ and $\quad \frac{\alpha}{2} = .10$
Therefore, $z_{\alpha/2} = z_{.10} = invNorm(1 - .1) = 1.2816$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.2816*13}{\sqrt{460}} = .7768$
$LEP = \overline{X} - E = 29 - .7768 = 28.2232$
$REP = \overline{X} + E = 29 + .7768 = 29.7768$

Exercise 7.1.8. The amount of time a student spends doing homework for a three credit math course has a mean μ and known standard deviation $\sigma = 44$ minutes. To estimate the mean μ, a sample of 178 student homework time was collected. The sample mean $\overline{X}$ was found to be 145 minutes. Find a 92 percent confidence interval for true mean time μ and give the margin of error (MOE).

Solution.

Here the population standard deviation $\sigma = 44$,
the sample size $n = 178$,
the sample mean $\overline{X} = 145$ minutes,
Level of confidence = 92 percent. So,
$1 - \alpha = .92, \quad \alpha = .08, \quad$ and $\quad \frac{\alpha}{2} = .04$
Therefore, $z_{\alpha/2} = z_{.04} = invNorm(1 - .04) = 1.7507$

$MOE = E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.7507*44}{\sqrt{178}} = 5.7737$
$LEP = \overline{X} - E = 139.2263$
$REP = \overline{X} + E = 145 + 5.7737 = 150.7737$

Problems: On Required sample size for Z-intervals

Exercise 7.1.9. The hourly wages in an industry has a standard deviation σ = \$17. The mean hourly wages μ has to be estimated within \$3 from the true value of μ, with 90 percent confidence. What would be the sample size needed?

Solution.
Here the population standard deviation σ = 17, and $E = 3$
Level of confidence = 90 percent. So,
$1 - \alpha = .90, \quad \alpha = .10, \quad$ and $\quad \frac{\alpha}{2} = .05.$
Therefore, $z_{\alpha/2} = z_{.05} = invNorm(1 - .05) = 1.6449$

The sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.6449*17}{3}\right)^2 = 86.8829.$
We would round it upward. So, answer = 87

Exercise 7.1.10. The tuition X paid by a student per semester in a university has a distribution with mean μ and σ = \$416. How large a sample should you draw so that you are 95 percent sure that the true value of μ will be within \$10 of the sample mean $\overline{x}$?

Solution.
Here the population standard deviation σ = 416, and $E = 10$.
Level of confidence = 95 percent. So,
$1 - \alpha = .95, \quad \alpha = .05, \quad$ and $\quad \frac{\alpha}{2} = .025.$
Therefore, $z_{\alpha/2} = z_{.025} = invNorm(1 - .025) = 1.9600.$

The sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.9600*416}{10}\right)^2 = 6648.1189.$
We round it upward. So, answer = 6649 *Note that the required sample size is really high, because the required precision within \$10 was too low compared to the standard deviation* $\sigma = 416$.

Exercise 7.1.11. Let X represent the monthly gas consumption (in CCF, in a winter month), by the households in a county. From past experience, it is known that the standard deviation is σ = 55 CCF. The county wants to estimate the mean monthly gas consumption μ within 15 CCF, with 98 percent confidence level. What would be the sample size required?

Solution.
Here the population standard deviation $\sigma = 55$, and $E = 15$.
Level of confidence = 98 percent. So,
$1 - \alpha = .98, \quad \alpha = .02, \quad$ and $\quad \frac{\alpha}{2} = .01.$
Therefore, $z_{\alpha/2} = z_{.01} = invNorm(1 - .01) = 2.3263$

The sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{2.3263*55}{15}\right)^2 = 72.7569$.
We round it upward. So, answer = 73

Exercise 7.1.12. The telephone company's data shows that length X of their international calls has a standard deviation 14.5 minutes. The company wants to estimate the mean length μ of these calls within 2.5 minutes of the mean, with a 96 percent level of confidence. What would be the sample size required?

Solution.
Here the population standard deviation $\sigma = 14.5$, and the MOE $E = 2.5$.
Level of confidence = 96 percent. So,
$1 - \alpha = .96, \quad \alpha = .04, \quad$ and $\quad \frac{\alpha}{2} = .02$.
Therefore, $z_{\alpha/2} = z_{.02} = invNorm(1 - .02) = 2.0537$.

The sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{2.0537*14.5}{2.5}\right)^2 = 141.8829$.
We round it upward. So, answer = 142

Exercise 7.1.13. The mean birth weight μ of babies have to be estimated within 5 oz from the actual mean μ, with a 99 percent confidence level. The standard deviation of the birth weight is known to be $\sigma = 25$ oz. What would be the sample size required?

Solution.
Here the population standard deviation $\sigma = 25$, and the MOE $E = 5$.
Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad$ and $\quad \frac{\alpha}{2} = .005$.
Therefore, $z_{\alpha/2} = z_{.005} = invNorm(1 - .005) = 2.5758$.

The sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{2.5758*25}{5}\right)^2 = 165.8686$.
We round it upward. So, answer = 166

Exercise 7.1.14. The mean monthly water consumption μ by the households in a subdivision has to be estimated within 500 gallons, with 96 percent confidence level. The standard deviation of the water consumption is known to be $\sigma = 3000$ gallons. What would be the sample size required?

Solution.
Here the population standard deviation $\sigma = 3000$, and the MOE $E = 500$.
Level of confidence = 96 percent. So,
$1 - \alpha = .96, \quad \alpha = .04, \quad$ and $\quad \frac{\alpha}{2} = .02$.
Therefore, $z_{\alpha/2} = z_{.02} = invNormal(1 - .02) = 2.0537$

The sample size $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{2.0537*3000}{500}\right)^2 = 151.8366$
We round it upward. So, answer = 152

7.2 Confidence interval for mean μ when σ is Unknown

As above, let X be a random variable with mean μ and standard deviation σ. As in the last section, we want to estimate the mean μ by confidence intervals. In the last section, we assume that the standard deviation σ was known, which would not be the case with populations without much familiarity or experience.

In this section, we will deal with such populations where the standard deviation σ is not known. In the last section, the main tool (or fact) that we used was that, by CLT, the sampling distribution of the statistic $Z = \frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$ was standard normal. In this section, we use the distribution of a similar statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{(\overline{X} - \mu)\sqrt{n}}{S} \tag{7.6}$$

In fact, T has a t-distribution, with degrees of freedom $n-1$, as described below.

7.2.1 Student's t distribution

We briefly mentioned T-random variable, in section 5.1.2. We further elaborate properties of T-random variables.

Given a positive integer m, there is a random variable, denoted by t_m, that is said to have t-distribution with **degrees of freedom** m. The distribution is also known as Student's t-distribution, because it was discovered by a chemist W. S. Gosset, under the pseudonym "Student". The important properties of t distributions are listed below. A t random variable with **degrees of freedom** m would be denoted by T. We also say that the random variable T has a t_m-distribution, or just write $T \sim t_m$. Properties of $T \sim t_m$ random variables are given as follows:

1. The equation $y = f(x)$ of the pdf of a $T \sim t_m$ random variable was given in section 5.1.2, Equation 5.4. The the graph of the pdf $y = f(x)$ of a $T \sim t_m$ was also given in section 5.1.2, which we reproduce below.

   The graph looks very similar to the graph of the pdf of the standard normal random variable $Z \sim N(0,1)$. Like normal, the graph is also has a bell shape. For an ordinary reader, it would not be easy to distinguish the graph from that of the standard normal random variable.

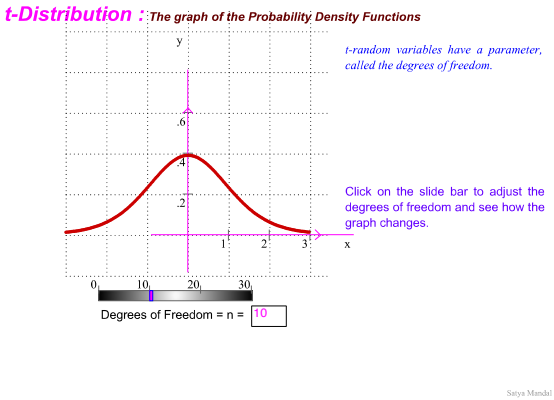
2. (**Normal Approximation:**) When m is large (say 30 or above), a $T \sim t_m$ random variable can be approximated by the standard normal random variable $Z \sim N(0,1)$.

3. It was stated in section 5.1.2, for $T \sim t_m$, the mean and standard deviation of T is given by

$$\begin{cases} \text{mean of } X & \mu = E(X) = 0 \\ \text{St.Dev. of } X & \sigma = \sqrt{\frac{m}{m-2}} \quad m \geq 3 \end{cases}$$

4. The distribution T is symmetric around the y-axis (or around the mean zero).

5. When the degrees of freedom m is small, the graph is flatter. For higher degrees of freedom m, the graph is stiffer.

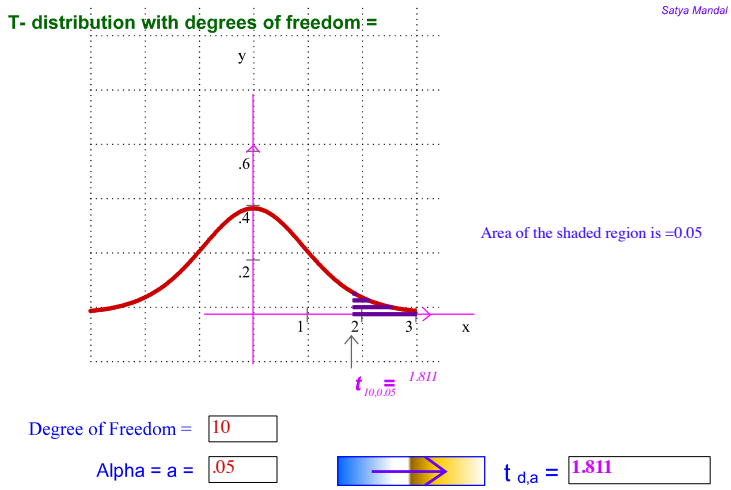


Inverse Probability for Student's t Inverse probability for $T \sim t_m$ random variables is defined as was done for normal random variables $Z \sim N(0,1)$. In particular, like z_α defines above (7.3), we define $t_{m,\alpha}$ as follows:

Definition. Let $T \sim t_m$). Given a number $0 < \alpha < 1$, the number $t_{m,\alpha}$ is defined by the formula

$$P(t_{m,\alpha} \leq T) = \alpha. \quad \text{Equivalently,} \quad \begin{cases} P(T \leq t_{m,\alpha}) = 1 - \alpha, & \text{Or,} \\ \mathbf{t_{m,\alpha} = invT(1 - \alpha, m)} \end{cases} \tag{7.7}$$

where **invT**$(-,-)$ is a function in TI-84, under Distr-key. These numbers $t_{m,\alpha}$ are also called **Critical values**. There are many internet sites that would give these numbers.



7.2.2 The T-Interval for μ

First, we give the sampling distribution of the statistic $T = \frac{(\overline{X}-\mu)\sqrt{n}}{S}$, mentioned above (7.6). This can be viewed as a counter part of the CLT, for T.

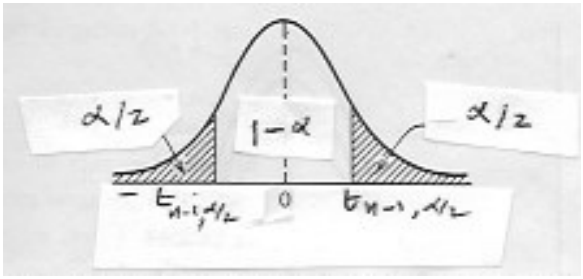
Theorem. Let $X \sim N(\mu, \sigma)$ be a normal random variable with mean μ and standard deviation σ. Let $X_1, X_2, \ldots, X_n$ be a sample of size n, from the X-population. Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{(\overline{X} - \mu)\sqrt{n}}{S} \sim t_{n-1} \tag{7.8}$$

has a T-distribution, with degrees of freedom $n - 1$. ■

As in the computations for Z-interval,

$$\begin{cases} P\left(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}\right) = 1 - \alpha \quad OR, \\ P\left(-t_{n-1,\alpha/2} \le \frac{(\overline{X}-\mu)\sqrt{n}}{S} \le t_{n-1,\alpha/2}\right) = 1 - \alpha \end{cases}$$



Simplifying,

$$P\left(\overline{X} - \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{t_{n-1,\alpha/2}S}{\sqrt{n}}\right) = 1 - \alpha \tag{7.9}$$

By Equation (7.2), this is exactly what is needed for a confidence interval, with $(1-\alpha)100$ percent confidence interval. We state the same, as in the theorem below.

Theorem. We use the notations as above. Assume that σ is unknown. A $(1-\alpha)100$ (approximate) percent confidence interval, for μ, is given by

$$\overline{X} - E \leq \mu \leq \overline{X} + E \qquad \text{where} \quad E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}}$$

We add few more definitions, to summarize this:

$$\begin{cases} \mathbf{E} = \frac{\mathbf{t_{n-1,\alpha/2}S}}{\sqrt{\mathbf{n}}} & \text{is called the Margin of Error (MOE)} \\ LEP = \overline{X} - E & \text{is called the Left end point} \\ REP = \overline{X} + E & \text{is called the Right end point} \end{cases}$$

Informally, this confidence interval is also called a T-interval.

As was defined above, this will called be a two-sided confidence interval of μ. We could compute one sided confidence intervals, which we avoid.

Remarks.:

1. A common question would be, to estimate μ, when do we use the Z-Interval and when do we use the T-Interval? Use the T-Interval only when σ is not known. Not using σ, when it is known would amount to disregarding some available information, which would not be a good philosophy in estimation.

2. Recall, for large n (say 30 or above), t_{n-1} would be approximately equal to the standard normal $Z \sim N(0,1)$. That is why, for large sample size n, using $z_{\alpha/2}$, instead of $t_{n-1,\alpha/2}$ in the formula for E would not make any noticeable difference.

3. In this course, we avoid probability problems for T. The **tcdf** function of TI-84, makes it easy to compute probability for T. We would not assign such problems. But we would workout one example.

Example. The mean salary X of the engineers in a state is $\mu = \$65,000$, and $X \sim N(\mu, \sigma)$. You collect a sample of 12 engineers. What is the probability that the T statistics (7.8) of the sample would be between will be 1.1 and 10.
Solution. Here $n = 12$ and$T$ has degrees of freedom $df = n - 1 = 12 - 1 = 11$.
So, $P(1.1 \leq T < 10) = tcdf(1.1, 10, 11) = 0.1474$.

7.2.3 Problems: On T-intervals for mean

Exercise 7.2.1. The mean weight μ of a campus population is to be estimated. It would be reasonable to assume that weight X has a normal distribution. A sample of size $n = 18$ was collected. The sample mean was found to be $\overline{x} = 170.5$ and standard deviation was $s = 13.3$. Compute a 99 percent a confidence interval for μ. State the degrees of freedom df, compute the margin of error, LEP and REP.

Solution.
Since the population standard deviation σ is unknown, we will compute a T-interval.

Here the sample size $n = 18$, the sample mean $\overline{X} = 170.5$,
Sample standard deviation $s = 13.3$.
The degrees of freedom $df = n - 1 = 18 - 1 = 17$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \frac{\alpha}{2} = .005$. Therefore,
$t_{n-1,\alpha/2} = t_{17,.005} = invT(1 - .005, 17) = 2.8982$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{2.8982*13.3}{\sqrt{18}} = 9.0854$
$LEP = \overline{X} - E = 170.5 - 9.0854 = 161.4146$
$REP = \overline{X} + E = 170.5 + 9.0854 = 179.5854$

Exercise 7.2.2. The time taken to complete a problem in a Math 365 test is normally distributed with mean μ and standard deviation σ. A sample of size 23 was taken, and sample mean and standard deviation were found to be $\overline{x} = 4.7$ and $s = .47$. Estimate the mean time μ taken to complete a problem using a 98 percent confidence interval.

Solution.
Since the population standard deviation σ is unknown, we will compute a T-interval.

Here the sample size $n = 23$,
the sample mean $\overline{X} = 4.7$,
Sample standard deviation $s = .47$

The degrees of freedom $df = n - 1 = 23 - 1 = 22$

Level of confidence = 98 percent.
$1 - \alpha = .98, \quad \alpha = .02, \quad and \quad \frac{\alpha}{2} = .01.$ Therefore,
$t_{n-1,\alpha/2} = t_{22,.01} = invT(1 - .01, 22) = 2.5083$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{2.5083*.47}{\sqrt{23}} = .2458$
$LEP = \overline{X} - E = 4.7 - .2458 = 4.4542$
$REP = \overline{X} + E = 4.7 + .2458 = 4.9458$

Exercise 7.2.3. To estimate the mean length μ of babies at birth a sample of size $n = 16$ was taken. The sample mean was found to be $\overline{x} = 18.6$ and standard deviation was $s = 9.486$. Construct a 96 percent confidence interval for mean μ. State the degrees of freedom df, compute the margin of error, LEP and REP.

Solution.
It is reasonable to assume that the length X is normal. Since the population standard deviation σ is unknown, we will compute a T-interval.

Here the sample size $n = 16$,
the sample mean $\overline{X} = 18.6$,
Sample standard deviation $s = 9.486$

The degrees of freedom $df = n - 1 = 16 - 1 = 15$

Level of confidence = 96 percent. So,
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \frac{\alpha}{2} = .02.$ Therefore,
$t_{n-1,\alpha/2} = t_{15,.02} = invT(1 - .02, 15) = 2.2485$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{2.2485*9.486}{\sqrt{16}} = 5.3323$
$LEP = \overline{X} - E = 18.6 - 5.3323 = 13.2677$
$REP = \overline{X} + E = 18.6 + 5.3323 = 23.9323$

Exercise 7.2.4. The time taken by an athlete to run an event is normally distributed with mean μ and unknown standard deviation σ. To estimate the mean μ he ran 9 times and the sample mean was found to be $\overline{X} = 33$ seconds and the sample standard deviation $s = 3.5$ seconds. Construct a 95 percent confidence interval for mean μ. State the degrees of freedom df, compute the margin of error, LEP and REP.

Solution.
It is reasonable to assume that the time X is normal. Since the population standard deviation σ is unknown, we will compute a T-interval.

Here the sample size $n = 9$,
the sample mean $\overline{X} = 33$,
Sample standard deviation $s = 3.5$

The degrees of freedom $df = n - 1 = 9 - 1 = 8$

Level of confidence = 95 percent.
$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \frac{\alpha}{2} = .025.$ Therefore,
$t_{n-1,\alpha/2} = t_{8,.025} = invT(1 - .025, 8) = 2.3060$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{2.3060*3.5}{\sqrt{9}} = 2.6903$
$LEP = \overline{X} - E = 33 - 2.6903 = 30.3097$
$REP = \overline{X} + E = 33 + 2.6903 = 35.6903$

Exercise 7.2.5. The mean length μ of telephone calls would have to be estimated. In sample of 14 calls had a sample mean $\overline{X} = 13$ minutes and the sample standard deviation was $s = 5.5$ minutes. Construct a 90 percent confidence interval for mean μ. State the degrees of freedom df compute the margin of error, LEP and REP.

Solution.
It is reasonable to assume that the time length X is normal. Since the population standard deviation σ is unknown, we will compute a T-interval.

Here the sample size $n = 14$,
the sample mean $\overline{X} = 13$,

Sample standard deviation $s = 5.5$

The degrees of freedom $df = n - 1 = 14 - 1 = 13$

Level of confidence = 90 percent.
$1 - \alpha = .90, \quad \alpha = .10, \quad and \quad \frac{\alpha}{2} = .05.$ Therefore,
$t_{n-1,\alpha/2} = t_{13,.05} = invT(1 - .05, 13) = 1.7709$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{1.7709*5.5}{\sqrt{14}} = 2.6031$
$LEP = \overline{X} - E = 13 - 2.6031 = 10.3969$
$REP = \overline{X} + E = 13 + 2.6031 = 15.6031$

Exercise 7.2.6. The mean weight μ of a variety of bananas in a grocery store has to be estimated. A sample of 20 bananas was collected. The mean weight was found to be $\overline{X} = 180$ grams and the sample standard deviation was $s = 27$ grams. Construct a 85 percent confidence interval for mean μ. State the degrees of freedom df compute the margin of error, LEP and REP.

Solution.
It is reasonable to assume that the weight X is normal. Since the population standard deviation σ is unknown, we will compute a T-interval.

Here the sample size $n = 20$,
the sample mean $\overline{X} = 180$,
Sample standard deviation $s = 27$

The degrees of freedom $df = n - 1 = 20 - 1 = 19$

Level of confidence = 85 percent.
$1 - \alpha = .85, \quad \alpha = .15, \quad and \quad \frac{\alpha}{2} = .075.$ Therefore,
$t_{n-1,\alpha/2} = t_{19,.075} = invT(1 - .075, 19) = 1.5002$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{1.5002*27}{\sqrt{20}} = 9.0573$
$LEP = \overline{X} - E = 180 - 9.0573 = 170.9427$
$REP = \overline{X} + E = 180 + 9.0573 = 189.0573$

Exercise 7.2.7. It is assumed that the lifetime (in hours) of lightbulbs produced in a

factory is normally distributed with mean μ and standard deviation σ. To estimate μ the following data was collected on the lifetime of bulbs.

5110	4671	6441	3331	5055	5270	5335	4973	1837
7783	4560	6074	4777	4707	5263	4978	5418	5123

Construct a 92 percent confidence interval for mean μ. State the degrees of freedom df, compute the margin of error, LEP and REP.

Solution.

It is reasonable to assume that the time X is normal. Since the population standard deviation σ is unknown, we will compute a T-interval.

As in chapter 2, enter data in a LIST of your TI-84, and
compute mean $\overline{X}$ and sample standard deviation S:
Here the sample size $n = 18$,
the sample mean $\overline{X} = 5039.2222$,
Sample standard deviation $s = 1203.8025$

The degrees of freedom $df = n - 1 = 18 - 1 = 17$

Level of confidence = 92 percent. So,
$1 - \alpha = .92, \quad \alpha = .08, \quad and \quad \frac{\alpha}{2} = .04.$ Therefore,
$t_{n-1,\alpha/2} = t_{17,.04} = invT(1 - .04, 17) = 1.8619$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{1.8619*1203.8025}{\sqrt{18}} = 528.2936$
$LEP = \overline{X} - E = 5039.2222 - 528.2936 = 4510.9286$
$REP = \overline{X} + E = 5039.2222 + 528.2936 = 5567.5158$

Exercise 7.2.8. To estimate the mean weight (in pounds) of salmon in a river the following sample was collected:

34.7	33.8	38.2	20.3	27.8	45.3	43.1	37.3	32.5	32.3
31.8	41.5	44.5	29.2	25.3	29.6	39.5	29.1	37.3	

Construct a 97 percent confidence interval for mean μ. State the degrees of freedom df, compute the margin of error, LEP and REP.

Solution.

It is reasonable to assume that the weight X is normal. Since the population standard

deviation σ is unknown, we will compute a T-interval.
As in chapter 2, enter data in a LIST of your TI-84, and
compute mean $\overline{X}$ and sample standard deviation S:

Here the sample size $n = 19$,
the sample mean $\overline{X} = 34.3737$,
Sample standard deviation $s = 6.7608$

The degrees of freedom $df = n - 1 = 19 - 1 = 18$

Level of confidence = 92 percent.So,
$1 - \alpha = .97, \quad \alpha = .03, \quad and \quad \frac{\alpha}{2} = .015.$ Therefore,
$t_{n-1,\alpha/2} = t_{18,.04} = invT(1 - .015, 18) = 2.3562$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{2.3562*6.7608}{\sqrt{19}} = 3.6545$
$LEP = \overline{X} - E = 34.3737 - 3.6545 = 30.7192$
$REP = \overline{X} + E = 34.3737 + 3.6545 = 38.0282$

Exercise 7.2.9. The following data represents the time (in minutes) taken by students to drive to campus.

5.7	22.4	14.7	16.7	10.6	26.9	17.9	32.5	20.9	9.2
3.4	20.4	17.3	31.9	4.4	23.7				

Construct a 98 percent confidence interval for mean μ. State the degrees of freedom df compute the margin of error, LEP and REP.

Solution.
It is reasonable to assume that the driving time X is normal. Since the population standard deviation σ is unknown, we will compute a T-interval. As in chapter 2, enter data in a LIST of your TI-84, and
compute mean $\overline{X}$ and sample standard deviation S:

Here the sample size $n = 16$,
the sample mean $\overline{X} = 17.4125$,
Sample standard deviation $s = 9.0843$

The degrees of freedom $df = n - 1 = 16 - 1 = 15$

Level of confidence = 98 percent. So,
$1 - \alpha = .98, \quad \alpha = .02, \quad and \quad \frac{\alpha}{2} = .01.$ Therefore,
$t_{n-1,\alpha/2} = t_{15,.01} = invT(1 - .01, 15) = 2.6025$

$MOE = E = \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} = \frac{2.6025*9.0843}{\sqrt{16}} = 5.9104$
$LEP = \overline{X} - E = 17.4125 - 5.9104 = 11.5021$
$REP = \overline{X} + E = 17.4125 + 5.9104 = 23.3229$

7.3 Confidence Interval for p

In this section, we would estimate the population proportion p of an attribute by a confidence interval. Examples would be

1. proportion p of defective lamps produced in a factory,
2. proportion p of a variety of seeds that will germinate in a particular condition,
3. proportion p of voters in favor of a candidate and others.

In fact, this section is about polls and surveys, that you hear about everyday in new media.

To estimate such a proportion p, we would use the sampling distribution of the sample proportion of success $\overline{X}$. The distribution of the sample proportion of success $\overline{X}$ was discussed in section 6.3. As in section 6.3, let X represent the corresponding $Bernoulli(p)$ random variable. That means

$$X = \begin{cases} 1 & if \textbf{ success} \\ 0 & if \textbf{ failure} \end{cases}$$

To estimate p, we take a sample $X_1, X_2, \ldots, X_n$ of size n from this $Bernoulli(p)$-population and X_i represents the outcome of testing the i^{th}-trial as follows:

$$X_i = \begin{cases} 1 & if\ i^{th}\ trial\ is\ a \textbf{ success} \\ 0 & if\ i^{th}\ trial\ is\ a \textbf{ failure} \end{cases}$$

Then, $T = X_1 + X_2 + \cdots + X_n$ is the total number of success, and

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{T}{n} = \textbf{the sample proportion of success}$$

Also, recall from section 6.3, sample size n is large and p is not very close to 0 or 1, then

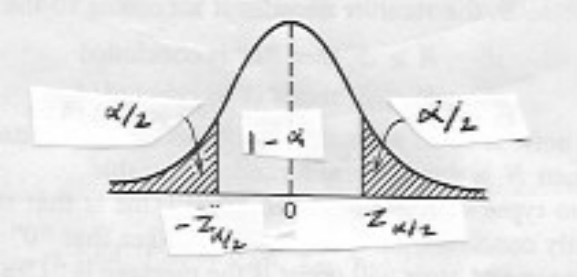
$$\overline{X} \sim N(p, \sigma_{\overline{X}}) \qquad \text{where} \qquad \sigma_{\overline{X}} = \sqrt{\frac{p(1-p)}{n}}$$

So,

$$Z = \frac{\overline{X} - p}{\sigma_{\overline{X}}} = \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \qquad \text{is st. normal.}$$

We proceed, as we did in the case of the Z-intervals. We have

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha \quad OR, \qquad P\left(-z_{\alpha/2} \le \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{\alpha/2}\right) = 1 - \alpha$$



Simplifying,

$$P\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \overline{X} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha \tag{7.10}$$

By Equation (7.2), roughly, this is exactly what is needed for a confidence interval, with $(1-\alpha)100$ percent confidence interval, for p. We state the same, as in the theorem below. However, two sides still depend on p, which is the unknown parameter, we are trying to estimate. We use the sample proportion $\overline{X}$ of success as a point estimate of p to get approximate equalities

$$P\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \le p \le \overline{X} + z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right) = 1 - \alpha \tag{7.11}$$

Now, by Equation (7.2), this gives a $(1-\alpha)100$ percent, approximate confidential interval for p. We state the same, as in the theorem below.

Theorem. We use the notations as above. A (approximate) $(1-\alpha)100$ percent confidence interval, for p, is given by

$$\overline{X} - e \le \mu \le \overline{X} + e \qquad \text{where} \quad e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

We add few more definitions, to summarize this:

$$\begin{cases} e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} & \text{is called the Margin of Error (MOE)} \\ E = z_{\alpha/2}\sqrt{\frac{1}{4n}} & \text{is called the Conservative Margin of Error (MOE)} \\ LEP = \overline{X} - e & \text{is called the Left end point} \\ REP = \overline{X} + e & \text{is called the Right end point} \end{cases}$$

In this list, we introduced the **Conservative MOE** $E = z_{\alpha/2}\sqrt{\frac{1}{4n}}$, which only depends on the sample size. Two inequalities, to be noted:

$$\begin{cases} \text{The error} \quad e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \leq E \\ \text{the actual error, as in (7.10)} \quad = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq E \end{cases}$$

The required samples size for estimating p: In the case of Z-intervals, we computed the required sample size (7.5), when MOE was given. Likewise, given a specified **conservative MOE**, E, for a $(1-\alpha)100$ percent confidence interval of p, the sample size n required is given by

$$n = \left(\frac{z_{\alpha/2}}{2E}\right)^2 \qquad \text{rounded to the higher integer.} \tag{7.12}$$

Informally, this confidence interval is also called a One proportion Z-interval.

An Example to interpret Conservative MOE: President Clinton was impeached in 1998-99 by the House of Representatives and acquitted by the Senate. During the impeachment trial, a typical news item would read as follows:

President Clinton has **64 percent** *approval rating.*
The poll has a margin of error **plus or minus 3.1 percentage** *points.*
The poll surveyed 972 people.

This meant the following, with p as the proportion of the population who "approved" President Clinton:

1. (A point estimate): that the sample proportion $\overline{x}$ of people who "approve" President Clinton was 0.64. So, a point estimate of p is .64.
2. Such polls usually use 95 percent confidence level.
3. Assuming that they are using a 95 percent confidence interval, they mean that

$$E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = z_{.025}\sqrt{\frac{1}{4n}} = 1.96\sqrt{\frac{1}{4 * 972}} = 0.031.$$

Interestingly, they give both the Conservative MOE and sample size, while one can be computed from the other.

7.3.1 Problems: One proportion Z-interval

Exercise 7.3.1. In a sample of 197 apples from a lot, 19 were found to be sour. Set a 99 percent confidence interval for the proportion p of sour apples in the lot.

Solution.

Here the sample size $n = 197$,
The number of success $T = 19$
So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{19}{197} = .0964$.
Level of confidence = 99 percent. So
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$ Therefore,
$z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

$MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 2.5758 * \sqrt{\frac{.0964(1-.0964)}{197}} = .054163$
In this section, for error terms, we retain at least 6 decimal points.

$LEP = X - e = .0964 - .054163 = .042237$
$REP = X + e = .0964 + .054163 = .150563$
Conservative MOE $E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 2.5758\sqrt{\frac{1}{4*197}} = .091759$
In this section, for error terms, we retain at least 6 decimal points.

Exercise 7.3.2. A new vaccine was tried on 147 randomly selected individuals, and it was determined that 97 of them developed immunity. Find a 95 percent confidence interval for the proportion p of individuals in the population for whom the vaccine would help. Give the MOE, LEP, REP and the conservative MOE.

Solution.

Here the sample size $n = 147$,
The number of success $T = 97$
So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{97}{147} = .6599$

Level of confidence = 95 percent. So
$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$ Therefore,
$z_{\alpha/2} = z_{.025} = invNorm(.975) = 1.9600$

$MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 1.9600 * \sqrt{\frac{.6599(1-.6599)}{147}} = .076584$
In this section, for error terms, we retain at least 6 decimal points.

$LEP = \overline{X} - e = .6599 - .076584 = .583316$
$REP = \overline{X} + e = .6599 + .076584 = .736484.$
Conservative MOE $E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 1.9600\sqrt{\frac{1}{4*147}} = .080829$
In this section, for error terms, we retain at least 6 decimal points.

Exercise 7.3.3. Before a congressional election, a poll was conducted. Out of 887 randomly selected voters interviewed, 389 said that they would vote for Candidate A, and 359 said that they would vote for Candidate B.

1. Construct a 98 percent confidence interval for the proportion p of voters who would vote for A.

2. Construct a 98 percent confidence interval for the proportion p of voters who would vote for B.

3. What is the conservative margin of error for both?

Solution.
Here the sample size = 887,
Level of confidence = 98 percent. So
$1 - \alpha = .98, \quad \alpha = .02, \quad and \quad \alpha/2 = .01.$ Therefore,
$z_{\alpha/2} = z_{.01} = invNorm(.99) = 2.3263$

1. **Candidate A:**
   The number of success $T = 389$
   So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{389}{887} = .4386$

   $MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 2.3263 * \sqrt{\frac{.4386(1-.4386)}{887}} = .038759$
   In this section, for error terms, we retain at least 6 decimal points.

   $LEP = \overline{X} - e = .4386 - .038759 = .399841$
   $REP = \overline{X} + e = .4386 + .038759 = .477359$
   Conservative MOE $E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 2.3263\sqrt{\frac{1}{4*887}} = .039055$
   In this section, for error terms, we retain at least 6 decimal points.

2. **Candidate B:**
   The number of success $T = 359$

So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{359}{887} = .4047$

$MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 2.3263 * \sqrt{\frac{.4047(1-.4047)}{887}} = .038339$
In this section, for error terms, we retain at least 6 decimal points.

$LEP = \overline{X} - e = .4047 - .038339 = .399841$
$REP = \overline{X} + e = .4047 + .038339 = .477359$
Conservative MOE = same, because the formula only depends on the sample size and level of confidence

Exercise 7.3.4. A researcher wants to estimate the proportion p of the children younger than two, in a community, who has anemia. He/She examined 180 children and found 42 children with anemia? Compute a 97 percent confidence interval for p. Give the MOE, LEP, REP and the conservative MOE.

Solution.
Here the sample size $n = 180$,
The number of success $T = 42$
So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{42}{180} = .2333$

Level of confidence = 97 percent. So
$1 - \alpha = .97, \quad \alpha = .03, \quad and \quad \alpha/2 = .015.$ Therefore,
$z_{\alpha/2} = z_{.015} = invNorm(.985) = 2.1701$

$MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 2.1701 * \sqrt{\frac{.2333(1-.2333)}{180}} = .068409$
In this section, for error terms, we retain at least 6 decimal points.

$LEP = \overline{X} - e = .2333 - .068409 = .164891$
$REP = \overline{X} + e = .2333 + .068409 = .301709$
Conservative MOE $E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 2.1701\sqrt{\frac{1}{4*180}} = .080875$
In this section, for error terms, we retain at least 6 decimal points.

Exercise 7.3.5. The proportion p of voters (in a county) who vote by absentee ballot is to be estimated. You sample 683 voters and found that 165 would have voted by absentee ballots. Compute 96 percent confidence interval for p. Give the MOE, LEP, REP and the conservative MOE.

Solution.
Here the sample size $n = 683$,

The number of success $T = 165$
So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{165}{683} = .2416$

Level of confidence = 96 percent. So
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \alpha/2 = .02.$ Therefore,
$z_{\alpha/2} = z_{02} = invNorm(.98) = 2.0537$

$MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 2.0537 * \sqrt{\frac{.2416(1-.2416)}{683}} = .033638$
In this section, for error terms, we retain at least 6 decimal points.

$LEP = \overline{X} - e = .2416 - .033638 = .207962$
$REP = \overline{X} + e = .2416 + .033638 = .275238$
Conservative MOE $E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 2.0537\sqrt{\frac{1}{4*683}} = .039291$
In this section, for error terms, we retain at least 6 decimal points.

Exercise 7.3.6. It is known that a vaccine may cause fever as side effect, after one takes the shot. The proportion p of those who suffer side effect is to be estimated. A sample of 913 individuals were examined and it was found the 153 suffered the side effect. Compute 94 percent confidence interval for p. Give the MOE, LEP, REP and the conservative MOE.

Solution.
Here the sample size $n = 913$,
The number of success $T = 153$
So, the sample proportion of success $\overline{X} = \frac{T}{n} = \frac{153}{913} = .1676$
Level of confidence = 94 percent. So
$1 - \alpha = .94, \quad \alpha = .06, \quad and \quad \alpha/2 = .03.$ Therefore,
$z_{\alpha/2} = z_{.03} = invNorm(.97) = 1.8808$

$MOE = e = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = 1.8808 * \sqrt{\frac{.1676(1-.1676)}{913}} = .023249$
In this section, for error terms, we retain at least 6 decimal points.

$LEP = \overline{X} - e = .1676 - .023249 = .144351$
$REP = \overline{X} + e = .1676 + .023249 = .190849$
Conservative MOE $E = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 1.8808\sqrt{\frac{1}{4*913}} = .031123$
In this section, for error terms, we retain at least 6 decimal points.

Exercise 7.3.7. The proportion p of those on campus who took a flu shot is to be estimated. A sample of 733 individuals were examined and it was found that 220 among them

took the shot. Compute 93 percent confidence interval for p. Give the MOE, LEP, REP and the conservative MOE.

Required sample size for 1-Proportion Z-intervals:
We use formula 7.12.

Exercise 7.3.8. A pollster wants to estimate the proportion p of the US population who would not support Government shutdown due to budget dispute. What would be the sample size required to estimate p within .02 of the actual value of p, with 99 percent confidence?

Solution.
Here given precision $E = .02$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$ Therefore,
$z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

The sample size $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2 = \left(\frac{2.5758}{2*.02}\right)^2 = 4146.7160.$
We round it upward. So, answer = 4147

Exercise 7.3.9. The proportion p of defective lightbulbs produced by a machine needs to be estimated within .01 to determine whether the machine needs to be replaced. How large a sample should we take to do this with 90 percent confidence?

Solution.
Here given precision $E = .01$

Level of confidence = 90 percent. So,
$1 - \alpha = .90, \quad \alpha = .10, \quad and \quad \alpha/2 = .05.$ Therefore,
$z_{\alpha/2} = z_{.05} = invNorm(1 - .05) = 1.6449$

The sample size $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2 = \left(\frac{1.6449}{2*.01}\right)^2 = 6764.2400.$
We would round it upward. So, answer = 6765

Exercise 7.3.10. The proportion p of those who earn more than \$50 K, in an industry, is to be estimated within .07 from the actual value of p, with 90 percent confidence. What would be the sample size needed?

Solution.

Here given precision $E = .07$

Level of confidence = 90 percent. So,
$1 - \alpha = .90, \quad \alpha = .10, \quad and \quad \alpha/2 = .05.$ Therefore,
$z_{\alpha/2} = z_{.05} = invNorm(1 - .05) = 1.6449$

The sample size $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2 = \left(\frac{1.6449}{2*.07}\right)^2 = 138.0457.$
We would round it upward. So, answer = 139

Exercise 7.3.11. The proportion p of those who paid more than \$15 K tuition, in a university, is to be estimated within .03 from the actual value of p, with 95 percent confidence. What would be the sample size needed?

Solution.

Here given precision $E = .03$

Level of confidence = 95 percent. So
$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$ Therefore,
$z_{\alpha/2} = z_{.025} = invNorm(.975) = 1.9600$

The sample size $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2 = \left(\frac{1.9600}{2*.03}\right)^2 = 1067.1111.$
We would round it upward. So, answer = 1068

Exercise 7.3.12. The proportion p of those who consumed more than 200 CCF in Gas in January, in a county, would have to be estimated within .05 from the actual value of p, with 98 percent confidence. What would be the sample size needed?

Solution.

Here given precision $E = .05$

Level of confidence = 98 percent. So
$1 - \alpha = .98, \quad \alpha = .02, \quad and \quad \alpha/2 = .01.$ Therefore,
$z_{\alpha/2} = z_{.01} = invNorm(.99) = 2.3263$

The sample size $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2 = \left(\frac{2.3263}{2*.05}\right)^2 = 541.1672.$
We would round it upward. So, answer = 542

Exercise 7.3.13. A telephone company wants to estimate the proportion p of call that are longer than 20 minutes. They want to estimated p within .03 from the actual value of p, with 94 percent confidence. What would be the sample size needed?

Exercise 7.3.14. The proportion p of babies born with weight more than 8 lbs, in a hospital, is to be estimated. It has to be estimated within .08 from the actual value of p, with 99 percent confidence. What would be the sample size required?

Exercise 7.3.15. he proportion p of households, in a county that use more than 5000 gallons of water in June, is to be estimated. It has to be estimated within .025 from the actual value of p, with 96 percent confidence. What would be the sample size required?

Interpretation:*In a poll released on October 28,1998, it was revealed that 60 percent of the US population wanted President Clinton be rebuked but not impeached. The poll was conducted among 1,013 adults, and it had a margin of error of 3 percentage points.*
Can you relate the last two numbers?
What is the level of confidence used here?

Solution:
News media polls use **95 percent** confidence intervals. When they say "margin of error," they mean "conservative margin of error."
Here $n = 1013$
Conservative MOE $E = .03$

Level of confidence = 95 percent. So
$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$ Therefore,
$z_{\alpha/2} = z_{.025} = invNorm(.975) = 1.9600$
We have

$$E = z_{\alpha/2}\sqrt{\frac{1}{4n}} \qquad \text{we check} \quad E = .03 = z_{\alpha/2}\sqrt{\frac{1}{4n}} = 1.96\sqrt{\frac{1}{4 * 1013}}$$

7.4 Confidence Interval of the Variance σ^2

Let $X \sim N(\mu, \sigma^2)$ be a normal random variable. In this section, we estimate the variance σ^2, by confidence intervals.

As usual, to estimate σ^2, take a sample $X_1, X_2, \cdots, X_n$ of size n from the X-population. As always, S^2 would denote the sample variance. To compute confidence intervals of σ^2, the distribution of

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \overline{X})^2}{\sigma^2} \qquad \text{will be used.} \tag{7.13}$$

In fact, U has a χ^2-distribution, which we describe next.

7.4.1 χ^2-random variable

We briefly mentioned χ^2-random variable, in section 5.1.2. Given a positive integer n, there is a random variable Y, that is said to have χ^2-distribution with **degrees of freedom** n, or we simply say $Y \sim \chi^2_n$. The important properties of the distributions, of such a random variable $Y \sim \chi^2_n$ are listed below. Let $Y \sim \chi^2_n$ be such a random variable.

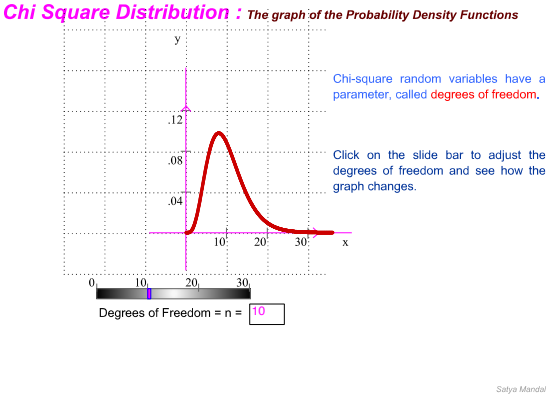
1. The equation $y = f(x)$ of the pdf of a $Y \sim \chi^2_n$ random variable was given in section 5.1.2, Equation 5.5.

2. A $Y \sim \chi^2_n$ is always non-negative.

3. As stated, $Y \sim \chi^2_n$ random variables comes, with degrees of freedom $df = n$.

4. The mean (or expected value) and variance of $Y \sim \chi^2_n$, with degrees of freedom $df = n$, is given by:
$$\begin{cases} E(Y) = n \\ Var(Y) = 2n \end{cases}$$

5. The the graph of the pdf $y = f(x)$ of a $Y \sim \chi^2_n$ was also given in section 5.1.2, which we reproduce below. The graph is somewhat like a bell. However, unlike normal or t-distribution, the graph does not maintain any symmetry around a vertical line.



As the degrees of freedom n increases, the graph starts looking more like that of a normal random variable. So, for large enough n, $Y \sim \chi^2_n$ can be approximated by a normal random variable $N(n, \sqrt{2n})$.

6. (**Probability computation:**) We use the $\chi^2 cdf$ function, under the key "Distr" in TI-84, to compute probability. For example suppose $Y \in \chi^2_{22}$-random variable. Then

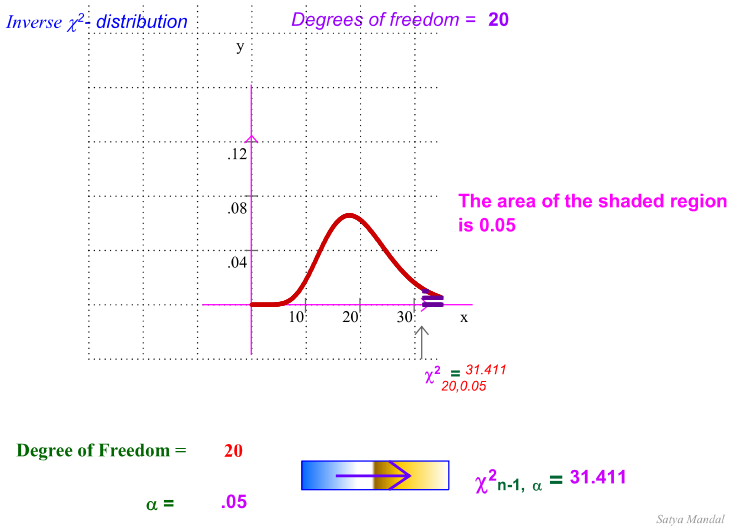
$$P(0 \leq Y \leq 47.5) = \chi^2 cdf(0, 47.5, 22) = .9987$$

Inverse Probability for Student's t Inverse probability for $T \sim t_m$ random variables is defined as was done for normal random variables $Z \sim N(0,1)$. In particular, like z_α defines above (7.3), we define $t_{m,\alpha}$ as follows:

Definition. Let $Y \sim \chi^2_n$. Given a number $0 < \alpha < 1$, the number $\chi^2_{n,\alpha}$ is defined by the formula

$$P(\chi^2_{n,\alpha} \leq Y) = \alpha. \quad \text{Equivalently,} \quad \begin{cases} P(Y \leq \chi^2_{n,\alpha}) = 1 - \alpha, & \text{Or,} \\ \chi^2\mathbf{cdf}(\mathbf{0}, \chi^2_{\mathbf{n},\alpha}, \mathbf{n}) = 1 - \alpha \end{cases} \tag{7.14}$$

These numbers $\chi^2_{n,\alpha}$ are also called **Critical values**. There are many internet sites that would give these numbers.



Unfortunately (Inconveniently), TI-84 does not have a function $inv\chi^2$, similar to **invNorm**$(-)$ or **invT**$(-)$ (see 7.7). This makes this section, less comfortable than the above sections, in this chapter, on Z-intervals, T-intervals and 1-Prop$Z$-intervals. The above animation

(in Flash) was created by this author, which is not working any more, because of discontinuance of Flash by the browsers. However, there are other such internet sites that would compute such inverse probabilities. We will make use of such websites.

7.4.2 The χ^2-Interval for σ^2

First, we give the sampling distribution of the statistic $U = \frac{(n-1)S^2}{\sigma^2}$, mentioned above (7.13).

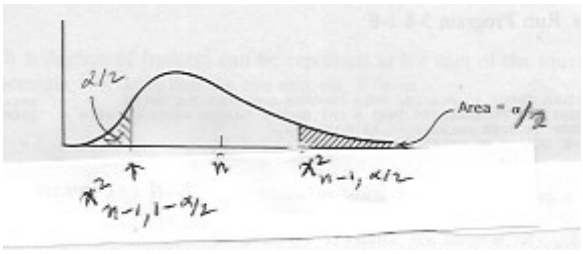
Theorem. Let $X \sim N(\mu, \sigma)$ be a normal random variable with mean μ and standard deviation σ. Let $X_1, X_2, \ldots, X_n$ be a sample of size n, from the X-population. Then

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1} \tag{7.15}$$

has a χ^2-distribution, with degrees of freedom $n-1$. ■

Since the pdf of $U \sim \chi^2_{n-1}$ is not symmetric, line Z or T, we have approach a little more carefully.

$$\begin{cases} P\left(\chi^2_{n-1,1-\alpha/2} \le U \le \chi^2_{n-1,\alpha/2}\right) = 1-\alpha \quad OR, \\ P\left(\chi^2_{n-1,1-\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{n-1,\alpha/2}\right) = 1-\alpha \end{cases}$$



Simplifying,

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha \tag{7.16}$$

By Equation (7.2), this is exactly what is needed for a confidence interval, with $(1-\alpha)100$ percent confidence interval. We state the same, as in the theorem below.

Theorem. We use the notations as above. In particular $X \sim N(\mu, \sigma^2)$. A $(1-\alpha)100$ (approximate) percent confidence interval, for σ^2, is given by

$$\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}$$

For consistency and convenience, we add the notations:

$$\begin{cases} LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} & \text{is called the Left end point} \\ REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} & \text{is called the Right end point} \end{cases}$$

Note, we did not introduce any Margin of error, which makes more sense when there is a symmetry. This is sometimes informally known as χ^2-interval.

7.4.3 Problems: On χ^2-Interval of σ^2

Before we proceed, we need figure out a way, to deal with the inconvenient situation mentioned above, that TI-84 does not have $inv\chi^2$-function. So, we need to have an alternate way to compute these numbers $\chi^2_{n-1,\alpha/2}$, and $\chi^2_{n-1,1-\alpha/2}$. If you ever use statistics in your work or higher study, you would have a statistical software, better than TI-84, and this would not be an issue. As mentioned above, certain websites would compute these numbers $\chi^2_{n-1,\alpha/2}$, and $\chi^2_{n-1,1-\alpha/2}$. Old fashioned way is to use certain tables. We would be using the websites, in conjunction with such tables. We would (optionally) double check with the formula 7.14. The tables are usually limited to $\alpha = .01, .05, .1$ and few other similar values. But the internet websites, and any software, would be able to deal with any value α.

Exercise 7.4.1. The birth weight of babies has a normal distribution, with variance σ^2. Because of the economic and social diversity of the community, there are concerns about variability of the birth weight. A sample of 26 birth-weight was collected and the sample variance was found to be $s^2 = 26.7$. Compute a 95 percent confidence interval for the variance σ^2 of weight.

Solution.

Here sample size $n = 26$ and

the sample variance $s^2 = 26.7$

Degrees of freedom $df = n - 1 = 25$

Level of confidence = 95 percent. So

$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$

From the websites (or, in this case, directly from the tables):

$$\begin{cases} \chi^2_{n-1,\alpha/2} = \chi^2_{25,.025} = 40.646 \\ \chi^2_{n-1,1-\alpha/2} = \chi^2_{25,.975} = 13.120 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 40.646, 25) = .975 \\ \chi^2\mathbf{cdf}(0, 13.120, 25) = .025 \end{cases}$$

Therefore, $LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(26-1)*26.7}{40.646} = 16.4223$

$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(26-1)*26.7}{13.120} = 50.8765$

Exercise 7.4.2. To estimate the variance length σ^2 of babies at birth a sample of size n = 16 was taken. The sample variance was found to be $s^2 = 90$ square-inches. Construct a 96 percent confidence interval for σ^2. State the degrees of freedom df, LEP and REP.

Solution.

It is reasonable to assume that the length X is normal.
Here the sample size $n = 16$,
Sample variance $s^2 = 90$

The degrees of freedom $df = n - 1 = 16 - 1 = 15$

Level of confidence = 96 percent. So
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \alpha/2 = .02.$
We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{15,.02}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{15,.98}$

These numbers are not available in the tables. So, we use the websites:

$$\begin{cases} \chi^2_{15,.02} = 28.2594 \\ \chi^2_{15,.98} = 5.9849 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 28.2594, 15) = .98 \\ \chi^2\mathbf{cdf}(0, 5.9849, 15) = .02 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(16-1)*90}{28.2594} = 47.7717$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(16-1)*90}{5.9849} = 225.5262$

Exercise 7.4.3. The variability of the length of the telephone calls would have to be estimated. A sample of 14 calls had a sample variance $s^2 = 30$ minutes-square. Construct a 90 percent confidence interval for σ^2 . State the degrees of freedom df, LEP and REP.

Solution.

It is reasonable to assume that the length X is normal.
Here the sample size $n = 14$,
Sample variance $s^2 = 30$

The degrees of freedom $df = n - 1 = 14 - 1 = 13$

Level of confidence = 90 percent. So,
$1 - \alpha = .90, \quad \alpha = .10, \quad and \quad \alpha/2 = .05.$
We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{13,.05}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{13,.95}$

From the websites (or, in this case, directly from the tables):

$$\begin{cases} \chi^2_{n-1,\alpha/2} = \chi^2_{13,.05} = 22.362 \\ \chi^2_{n-1,1-\alpha/2} = \chi^2_{13,.95} = 5.892 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 22.362, 13) = .95 \\ \chi^2\mathbf{cdf}(0, 5.892, 13) = .05 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(14-1)*30}{22.362} = 17.4403$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(14-1)*30}{5.892} = 66.1802$

Exercise 7.4.4. The variability of weight of a variety of bananas in a grocery store has to be estimated. A sample of 20 bananas was collected. The sample variance was $s^2 = 729$ grams-square. Construct a 96 percent confidence interval for σ^2. State the degrees of freedom df, LEP and REP.

Solution.
It is reasonable to assume that the weight $X \sim N(\mu, \sigma)$ is normal. Here the sample size $n = 20$,
Sample variance $s^2 = 729$

The degrees of freedom $df = n - 1 = 20 - 1 = 19$.

Level of confidence = 96 percent. So
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \alpha/2 = .02.$
We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{19,.02}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{19,.98}$

These numbers are not available in the tables. So, we use the above websites:

$$\begin{cases} \chi^2_{19,.02} = 33.6874 \\ \chi^2_{19,.98} = 8.5670 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 33.6874, 19) = .98 \\ \chi^2\mathbf{cdf}(0, 8.5670, 19) = .02 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(20-1)*729}{33.6874} = 411.1626$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(20-1)*729}{8.56702} = 1618.0697$

Exercise 7.4.5. The variability of monthly consumption of electricity in a subdivision has to be estimated. A sample of 9 household had a sample variance was $s^2 = 22500$ KWH-square. Construct a 94 percent confidence interval for σ^2. State the degrees of freedom df, LEP and REP.

Solution.

It is reasonable to assume that the monthly consumption $X \sim N(\mu, \sigma)$ is normal.
Here the sample size $n = 9$,
Sample variance $s^2 = 22500$

The degrees of freedom $df = n - 1 = 9 - 1 = 8$

Level of confidence = 94 percent. So
$1 - \alpha = .94, \quad \alpha = .06, \quad and \quad \alpha/2 = .03$. We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{8,.03}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{8,.97}$

These numbers are not available in the tables. So, we use the websites:

$$\begin{cases} \chi^2_{8,.03} = 17.0105 \\ \chi^2_{8,.97} = 2.3101 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 17.0105, 8) = .97 \\ \chi^2\mathbf{cdf}(0, 2.3101, 8) = .03 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(9-1)*22500}{17.0105} = 10581.6995$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(9-1)*22500}{2.3101} = 77918.7048$

Exercise 7.4.6. The lifetime (in hours) of lightbulbs produced in a factory is normally distributed. To estimate the variability, the following data was collected on the lifetime of bulbs.

5110	4671	6441	3331	5055	5270	5335	4973	1837
7783	4560	6074	4777	4707	5263	4978	5418	5123

Construct a 92 percent confidence interval for σ^2. State the degrees of freedom df, LEP and REP.

Solution.

It is reasonable to assume that the weight $X \sim N(\mu, \sigma)$ is normal.
As in chapter 2, enter data in the TI-84 and compute the variance.
Here the sample $sizen = 18$,
Sample variance $s^2 = (1203.8025)^2 = 1449140.459$
The degrees of freedom $df = n - 1 = 18 - 1 = 17$

Level of confidence = 92 percent. So,
$1 - \alpha = .92, \quad \alpha = .08, \quad and \quad \frac{\alpha}{2} = .04.$

We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{17,.04}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{17,.96}$
These numbers are not available in the tables. So, we use the websites:

$$\begin{cases} \chi^2_{17,.04} = 28.4450 \\ \chi^2_{17,.96} = 8.2878 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 28.4450, 17) = .96 \\ \chi^2\mathbf{cdf}(0, 8.2878, 17) = .04 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(18-1)*1449140.459}{28.4450} = 866070.937$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(18-1)*1449140.459}{8.2878} = 2972488.183$

Exercise 7.4.7. The following data represents the time (in minutes) taken by students to drive to campus. The variance of time σ^2 is to be estimated. The following date on time was collected.

5.7	22.4	14.7	16.7	10.6	26.9	17.9	32.5	20.9	9.2
3.4	20.4	17.3	31.9	4.4	23.7				

Construct a 99 percent confidence interval for σ^2. State the degrees of freedom df, LEP and REP.

Solution.
It is reasonable to assume that the weight $X \sim N(\mu, \sigma)$ is normal.
As in chapter 2, enter data in the TI-84 and compute the variance.

Here the sample size $n = 16$,
Sample variance $s^2 = (9.0843)^2 = 82.5245$
The degrees of freedom $df = n - 1 = 16 - 1 = 15$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$
We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{15,.005}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{15,.995}$
These numbers are available in the tables. So, from the table or the websites:

$$\begin{cases} \chi^2_{15,.005} = 32.8013 \\ \chi^2_{15,.995} = 4.6009 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 32.8013, 15) = .995 \\ \chi^2\mathbf{cdf}(0, 4.6009, 15) = .005 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(16-1)*82.5245}{32.8013} = 37.7384$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(16-1)*82.5245}{4.6009} = 269.0490$

Exercise 7.4.8. The following is sample data on the amount (in 1000 bushels) of wheat harvested by Kansas farmers in 2002.

206	300	200	385	280
600	225	933	320	260

Compute a 99 percent confidence interval for the variance σ^2 of harvest. State the degrees of freedom df, LEP and REP.

Solution.

It is reasonable to assume that the wheat production $X \sim N(\mu, \sigma)$ is normal. As in chapter 2, enter data in the TI-84 and compute the variance.

Here the sample size $n = 10$,
Sample variance $s^2 = (229.6149)^2 = 52723.0023$
he degrees of freedom $df = n - 1 = 10 - 1 = 9$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$
We need to compute, $\chi^2_{n-1,\alpha/2} = \chi^2_{9,.005}$ and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{9,.995}$
These numbers are available in the tables. So, from the table or the websites:

$$\begin{cases} \chi^2_{9,.005} = 23.5894 \\ \chi^2_{9,.995} = 1.7349 \end{cases} \quad \text{Double check by (7.14)} \quad \begin{cases} \chi^2\mathbf{cdf}(0, 23.5894, 9) = .995 \\ \chi^2\mathbf{cdf}(0, 1.7349, 9) = .005 \end{cases}$$

Therefore,
$LEP = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} = \frac{(10-1)*52723.0023}{23.5894} = 20115.2645$
$REP = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} = \frac{(10-1)*52723.0023}{1.7349} = 273506.8423$

Chapter 8

Comparing Populations

In this lesson, two populations will be compared by interval estimation. The following will be considered:

1. Compute confidence intervals of the difference $\mu_1 - \mu_2$ of the means of two populations. For example, difference $\mu_1 - \mu_2$ between the mean annual income of the male population μ_1 and the mean annual income of the female population μ_2 could of some interest.

2. Compute a confidence interval of the difference $p_1 - p_2$ of the proportions of an attribute present (or proportions of "success") in two populations. For example, there may be some interest in the difference $p_1 - p_2$ between of the proportion p_1 of the defective items produced by the new machine and the proportion p_2 of the defective items produced by the old machine.

8.1 Confidence Interval of $\mu_1 - \mu_2$

Suppose X, Y are two similar random variables. Let mean and st. deviations of X and Y be denoted, as follows:

	Population X	**Population** Y
mean	$E(X) = \mu_1$	$E(Y) = \mu_2$
St. Dev.	σ_1	σ_2

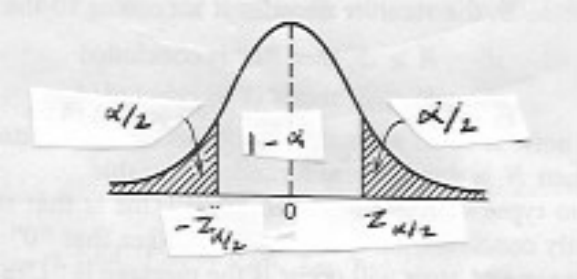
We proceed as follows.

1. A sample $X_1, X_2, \ldots, X_m$, of size m, is drawn from the X-population and a sample $Y_1, Y_2, \ldots, Y_n$, of size n, is drawn from the Y population. Let

$$\begin{cases} \overline{X} = \frac{X_1+X_2+\cdots+X_m}{m} \\ \overline{Y} = \frac{Y_1+Y_2+\cdots+Y_m}{n} \end{cases} \quad \text{be the sample means}$$

2. BY CLT, we have, approximately,
$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{m}}\right), \qquad \overline{Y} \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n}}\right) \qquad \text{distributions.}$$
3. Assume that the X-samples and Y-samples are drawn independently. In that case, it follows that, approximately,
$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \sigma\right) \quad \text{where} \quad \begin{cases} mean(\overline{X} - \overline{Y}) = E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \\ st.\ dev(\overline{X} - \overline{Y}) = \sigma = \sqrt{\frac{\sigma_1^2}{\mathbf{m}} + \frac{\sigma_2^2}{\mathbf{n}}} \end{cases}$$
If X and Y are normal, then this is exact (not just approximate) distribution.

It follows,
$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma} \sim N(0,1) \quad \text{has a st. normal distribution. Hence,}$$
$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha \quad OR, \quad P\left(-z_{\alpha/2} \le \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \le z_{\alpha/2}\right) = 1 - \alpha$$



Simplifying,
$$P\left((\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \le \mu_1 - \mu_2 \le (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha \qquad (8.1)$$
By Equation (7.2), this is exactly what is needed for a confidence interval, with $(1-\alpha)100$ percent confidence interval, for $\mu_1 - \mu_2$. We state the same, as in the theorem below.

Theorem. We use the notations as above. Assume that σ_1 and σ_2 are known. A $(1-\alpha)100$ (approximate) percent confidence interval, for $\mu_1 - \mu_2$, is given by
$$\overline{X} - E \le \mu \le \overline{X} + E \qquad \text{where} \quad E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$
If X, Y are normal, then it is an exact confidence interval for $\mu_1 - \mu_2$ (not just an approximation). We add few more definitions, to summarize this:
$$\begin{cases} \mathbf{E} = \mathbf{z}_{\alpha/2}\sqrt{\frac{\sigma_1^2}{\mathbf{m}} + \frac{\sigma_2^2}{\mathbf{n}}} & \text{is called the Margin of Error (MOE)} \\ LEP = (\overline{X} - \overline{Y}) - E & \text{is called the Left end point} \\ REP = (\overline{X} - \overline{Y}) + E & \text{is called the Right end point} \end{cases}$$
In formally, this is called the Two sample Z-Interval, for $\mu_1 - \mu_2$.

8.1.1 Problems: on Two sample Z-Interval for $\mu_1 - \mu_2$

Exercise 8.1.1. Suppose we have two normal populations $X \sim M(\mu_1, \sigma_1)$ and $X \sim M(\mu_2, \sigma_2)$. It is known that $\sigma_1 = 8.1$ and $\sigma_2 = 11.3$. A sample of size $m = 64$ was collected from the first population, and the sample mean was found to be $\overline{X} = 3.7$. A sample of size $n = 99$ was collected from the second population, and the sample mean was found to be $\overline{Y} = 4.1$. Compute a 99 percent confidence interval for the difference of mean $\mu_1 - \mu_2$.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 8.1$	$\sigma_2 = 11.3$
samplemean	$\overline{X} = 3.7$	$\overline{Y} = 4.1$
sample size	$m = 64$	$n = 99$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$ Therefore,
$z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

$$MOE = E = z_{\alpha/2}\sigma = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = 2.5758\sqrt{\frac{8.1^2}{64} + \frac{11.3^2}{99}} = 3.9191$$
$$LEP = (\overline{X} - \overline{Y}) - E = (3.7 - 4.1) - 3.9191 = -4.3191$$
$$REP = (\overline{X} - \overline{Y}) + E = (3.7 - 4.1) + 3.9191 = 3.5191$$

Exercise 8.1.2. The birth weight of babies in developed and developing countries are normally distributed with mean μ_1, μ_2 and standard deviation σ_1, σ_2, respectively. (My data is not real.) Given $\sigma_1 = 2.3$ pounds and $\sigma_2 = 2.9$ pounds. A sample of size $m = 35$ babies from the developed nations were collected and the sample mean birth weight was found to be $\overline{X} = 8.9$ pounds. A sample of size $n = 48$ babies from the developing nations was collected and the sample mean birth weight was found to be $\overline{Y} = 7.1$ pounds.

Determine the margin of error of the difference $\mu_1 - \mu_2$ and a confidence interval at the 95 percent level of confidence.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 2.3$	$\sigma_2 = 2.9$
sample mean	$\overline{X} = 8.9$	$\overline{Y} = 7.1$
sample size	$m = 35$	$n = 48$

Level of confidence = 95 percent. So
$1-\alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$ Therefore,
$z_{\alpha/2} = z_{.025} = invNorm(.975) = 1.9600$

$$MOE = E = z_{\alpha/2}\sigma = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = 1.9600\sqrt{\frac{2.3^2}{35} + \frac{2.9^2}{48}} = 1.1197$$
$$LEP = (\overline{X} - \overline{Y}) - E = (8.9 - 7.1) - 1.1197 = .6803$$
$$REP = (\overline{X} - \overline{Y}) + E = (8.9 - 7.1) + 1.1197 = 2.9197$$

Exercise 8.1.3. African elephants and Indian elephants are different in height, weight, and length of ear and tusk. It is natural to assume that all these are normally distributed. The mean height and standard deviation of African elephants are μ_1, $\sigma_1 = 1.2$ feet, respectively. The mean height and standard deviation of Indian elephants are μ_2, $\sigma_2 = 1.1$ feet, respectively. A sample of size 25 African elephants were collected and the sample mean height was found to be $\overline{X} = 10.9$ feet. A sample of size 28 Indian elephants was collected and the sample mean height was found to be $\overline{Y} = 9.1$ feet.

Determine the margin of error and a confidence interval of the difference $\mu_1 - \mu_2$ at the 98 percent level of confidence.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 1.2$	$\sigma_2 = 1.1$
sample mean	$\overline{X} = 10.9$	$\overline{Y} = 9.1$
sample size	$m = 25$	$n = 28$

Level of confidence = 98 percent. So
$1-\alpha = .98, \quad \alpha = .02, \quad and \quad \alpha/2 = .01.$ Therefore,
$z_{\alpha/2} = z_{.01} = invNorm(.99) = 2.3263$

$$MOE = E = z_{\alpha/2}\sigma = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = 2.3263\sqrt{\frac{1.2^2}{25} + \frac{1.1^2}{28}} = .7386$$
$$LEP = (\overline{X} - \overline{Y}) - E = (10.9 - 9.1) - .7386 = 1.0614$$
$$REP = (\overline{X} - \overline{Y}) + E = (10.9 - 9.1) + .7386 = 2.5386$$

Exercise 8.1.4. The mean weight of King salmon in Kenai and Anchor River would have to be compared. The mean weight of King in Kenai is μ_1 and the standard deviation $\sigma_1 = 7.7$ pounds. The mean weight of King in Anchor is μ_2 and the standard deviation $\sigma_2 = 9.1$ pounds. A sample of 51 King from Kenai had a mean $\overline{X} = 31$ pounds. A sample of 63 King from Anchor had a mean $\overline{Y} = 33$ pounds.

Determine the margin of error and a confidence interval of the difference $\mu_1 - \mu_2$ at the 97 percent level of confidence.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 7.7$	$\sigma_2 = 9.1$
sample mean	$\overline{X} = 31$	$\overline{Y} = 33$
sample size	$m = 51$	$n = 63$

Level of confidence = 97 percent. So
$1 - \alpha = .97, \quad \alpha = .03, \quad and \quad \alpha/2 = .015.$ Therefore,
$z_{\alpha/2} = z_{.015} = invNorm(.985) = 2.1701$

$MOE = E = z_{\alpha/2}\sigma = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = 2.1701\sqrt{\frac{7.7^2}{51} + \frac{9.1^2}{63}} = 3.4154$
$LEP = (\overline{X} - \overline{Y}) - E = (31 - 33) - 3.4154 = -5.4154$
$REP = (\overline{X} - \overline{Y}) + E = (31 - 33) + 3.4154 = 1.4154$

Exercise 8.1.5. There is a difference between fall semester grades and spring semester grades. The mean percentage score in fall is μ_1 and the standard deviation $\sigma_1 = 27$ percent. The mean percentage score in spring is μ_2 and the standard deviation $\sigma_2 = 23$ percent. A sample of 87 students in fall had a sample mean score $\overline{X} = 73$ percent. A sample of 77 students in spring had a sample mean score $\overline{Y} = 69$ percent.

Determine the margin of error and a confidence interval of the difference $\mu_1 - \mu_2$ at the 96 percent level of confidence.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 27$	$\sigma_2 = 23$
sample mean	$\overline{X} = 73$	$\overline{Y} = 69$
sample size	$m = 87$	$n = 77$

Level of confidence = 96 percent. So
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \alpha/2 = .02.$ Therefore,
$z_{\alpha/2} = z_{02} = invNorm(.98) = 2.0537$

$MOE = E = z_{\alpha/2}\sigma = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = 2.0537\sqrt{\frac{27^2}{87} + \frac{23^2}{77}} = 8.0198$
$LEP = (\overline{X} - \overline{Y}) - E = (73 - 69) - 8.0198 = -4.0198$
$REP = (\overline{X} - \overline{Y}) + E = (73 - 69) + 8.0198 = 12.0198$

Exercise 8.1.6. The difference in mean annual salary of the professors in two state universities have to be estimated. The mean annual salary in the University -I is μ_1 and the standard deviation $\sigma_1 = \$16,000$. The mean annual salary in the University -II is μ_2 and the standard deviation $\sigma_1 = \$11,500$. A sample of 47 professors in University-I had a mean salary $\overline{X} = \$79,000$. A sample of 58 professors in University-II had a mean salary $\overline{Y} = \$71,500$.

Determine the margin of error and a confidence interval of the difference $\mu_1 - \mu_2$ at the 94 percent level of confidence.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 11500$	$\sigma_2 = 16000$
sample mean	$\overline{X} = 79000$	$\overline{Y} = 71500$
sample size	$m = 47$	$n = 58$

Level of confidence = 94 percent. So
$1 - \alpha = .94, \quad \alpha = .06, \quad and \quad \alpha/2 = .03.$ Therefore,
$z_{\alpha/2} = z_{.03} = invNorm(.97) = 1.8808$

$$MOE = E = z_{\alpha/2}\sigma = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} = 1.8808\sqrt{\frac{11500^2}{47} + \frac{16000^2}{58}} = 85228.1439$$
$$LEP = (\overline{X} - \overline{Y}) - E = (79000 - 71500) - 5228.1439 = 2271.8561$$
$$REP = (\overline{X} - \overline{Y}) + E = (79000 - 71500) + 5228.1439 = 12728.1439$$

8.2 Two sample T-interval: Unknown σ_1, σ_2

As in the last section, X, Y represent two similar populations. In this section also, we will construct confidence intervals for the difference of means $\mu_1 - \mu_2$, when the standard deviations σ_1, σ_2 are unknown.

As a price, we would assume that $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ have normal distributions, and $\sigma_1 = \sigma_2 = \sigma$.

We proceed as follows:

1. As indicated above, $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ have normal distributions, and assume
$$\sigma_1 = \sigma_2 = \sigma$$

2. A sample $X_1, X_2, \ldots, X_m$, of size m, is drawn from the X-population and a sample $Y_1, Y_2, \ldots, Y_n$, of size n, is drawn from the Y population. Use the notations for sample mean and sample standard deviations (variance), as follows:

$$\begin{cases} \overline{X} = \frac{X_1+X_2+\cdots+X_m}{m} \\ \overline{Y} = \frac{Y_1+Y_2+\cdots+Y_m}{n} \end{cases} \qquad \begin{cases} S_1^2 = S_X^2 = \frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_m-\overline{X})^2}{m-1} \\ S_2^2 = S_Y^2 = \frac{(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2+\cdots+(Y_n-\overline{Y})^2}{n-1} \end{cases}$$

3. Since $\sigma^2 = \sigma_1^2 = \sigma_2^2$, we combined S_X^2 and S_Y^2, to obtain a **pooled estimate**, S_p^2 for σ^2, as follows:

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

This happens to be the weighted mean of S_X^2 with weight $(m-1)$, and S_Y^2 with weight $(n-1)$. So, **pooled estimate** S_p for σ is

$$S_p = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}$$

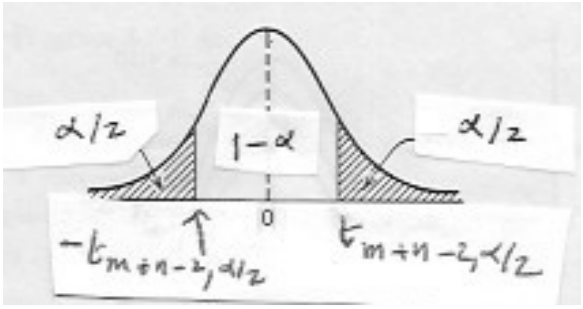
4. As in section 7.2,

$$T = \frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t_{m+n-2}$$

has a T-distribution, with degrees of freedom $df = m+n-2$.

As in the computations for Z-interval,

$$\begin{cases} P\left(-t_{m+n-2,\alpha/2} \le T \le t_{m+n-2,\alpha/2}\right) = 1-\alpha \quad OR, \\ P\left(-t_{m+n-2,\alpha/2} \le \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \le t_{m+n-2,\alpha/2}\right) = 1-\alpha \end{cases}$$



Simplifying,

$$P\left((\overline{X}-\overline{Y}) - t_{m+n-2,\alpha/2}S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \le \mu_1-\mu_2 \le (\overline{X}-\overline{Y}) + t_{m+n-2,\alpha/2}S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\right) = 1-\alpha \tag{8.2}$$

By Equation (7.2), this is exactly what is needed for a confidence interval, with $(1-\alpha)100$ percent confidence interval, for $\mu_1 - \mu_2$. We state the same, as in the theorem below.

Theorem. We use the notations as above. Assume that σ is unknown. A $(1-\alpha)100$ (approximate) percent confidence interval, for μ, is given by

$$\overline{X} - E \leq \mu_1 - \mu_2 \leq \overline{X} + E \qquad \text{where} \quad E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

We add few more definitions, to summarize this:

$$\begin{cases} \mathbf{E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} & \text{is called the Margin of Error (MOE)} \\ LEP = (\overline{X} - \overline{Y}) - E & \text{is called the Left end point} \\ REP = (\overline{X} - \overline{Y}) + E & \text{is called the Right end point} \end{cases}$$

This is referred to as the Two sample T-interval.

8.2.1 Problems: Two sample T-interval

Exercise 8.2.1. Suppose that two "similar" normal populations have means μ_1, μ_2 respectively and same standard deviations σ. A sample of size $m = 11$ from the first population the sample mean was found to be $\overline{X} = 13.2$ and the sample standard deviation $S_1 = 2.33$. A sample of size $n = 13$ was collected from the second population that had a sample mean $\overline{Y} = 11.5$ and sample standard deviation $S_2 = 2.73$.

Compute the pooled estimate S_p of σ and a confidence interval for $\mu_1 - \mu_2$ at the 96 percent level of significance.

Solution. We will use Two Sample T-interval, because $\sigma = \sigma_1 = \sigma_2$ are unknown.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 11$	$n = 13$
sample mean	$\overline{X} = 13.2$	$\overline{Y} = 11.5$
sample St. Dev.	$S_1 = 2.33$	$S_2 = 2.73$

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} = \sqrt{\frac{(11-1)2.33^2 + (13-1)2.73^2}{11+13-2}} = 2.5560$$

The degrees for freedom $df = m + n - 2 = 11 + 13 - 2 = 22$

Level of confidence = 96 percent. So
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \alpha/2 = .02.$ Therefore,

$t_{m+n-2,\alpha/2} = invT(.98, 22) = 2.1829$

$MOE = E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 2.1829 * 2.5560 \sqrt{\frac{1}{11} + \frac{1}{13}} = 2.2858$
$LEP = (\overline{X} - \overline{Y}) - E = (13.2 - 11.5) - 2.2858 = -.5858$
$REP = (\overline{X} - \overline{Y}) + E = (13.2 - 11.5) + 2.2858 = 3.9858$

Exercise 8.2.2. Suppose we have two normal populations with means μ_1, μ_2 and equal standard deviation σ. A sample of size $m = 64$ was collected from the first population and the sample mean and standard deviation were found to be $\overline{X} = 3.7$, $S_1 = 9.2$. A sample of size $n = 99$ was collected from the second population and the sample mean and standard deviation were $\overline{Y} = 4.1$, $S_2 = 8.7$.

Compute the pooled estimate S_p of σ and a confidence interval for $\mu_1 - \mu_2$ at the 95 percent level of significance.

Solution. We will use Two Sample T-interval, because $\sigma = \sigma_1 = \sigma_2$ are unknown.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 64$	$n = 99$
sample mean	$\overline{X} = 3.7$	$\overline{Y} = 9.2$
sample St. Dev.	$S_1 = 4.1$	$S_2 = 8.7$

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} = \sqrt{\frac{(64-1)4.1^2 + (99-1)8.7^2}{64+99-2}} = 8.8990$$

The degrees for freedom $df = m + n - 2 = 64 + 99 - 2 = 161$

Level of confidence = 95 percent. So
$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$ Therefore,
$t_{m+n-2,\alpha/2} = t_{161,.025} = invT(.975, 161) = 1.9748$

$MOE = E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 1.9748 * 8.8990 \sqrt{\frac{1}{64} + \frac{1}{99}} = 2.8187$
$LEP = (\overline{X} - \overline{Y}) - E = (3.7 - 4.1) - 2.8187 = -3.2187$
$REP = (\overline{X} - \overline{Y}) + E = (3.7 - 4.1) + 2.8187 = 2.4187$

Exercise 8.2.3. The difference in mean monthly water consumption in two adjacent towns has to be estimated estimated. A sample 37 household in the Town-I had a sample mean

6300 gallons and standard deviation 450 gallons. A sample 49 household in the Town-II had a sample mean 6800 gallons and standard deviation 650 gallons. Compute a 94 percent confidence interval for the difference $\mu_1 - \mu_2$.

Solution. We will use Two Sample T-interval, because $\sigma = \sigma_1 = \sigma_2$ are unknown.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 37$	$n = 49$
sample mean	$\overline{X} = 6300$	$\overline{Y} = 6800$
sample St. Dev.	$S_1 = 450$	$S_2 = 650$

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} = \sqrt{\frac{(37-1)450^2 + (49-1)650^2}{37+49-2}} = 572.8990$$

The degrees for freedom $df = m + n - 2 = 37 + 49 - 2 = 84$

Level of confidence = 94 percent. So
$1 - \alpha = .94, \quad \alpha = .06, \quad and \quad \alpha/2 = .03.$ Therefore,
$t_{m+n-2,\alpha/2} = t_{84,.03} = invT(.97, 84) = 1.9065$

$MOE = E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 1.9065 * 572.8990 \sqrt{\frac{1}{37} + \frac{1}{49}} = 237.8840$
$LEP = (\overline{X} - \overline{Y}) - E = (6300 - 6800) - 237.8840 = -737.884$
$REP = (\overline{X} - \overline{Y}) + E = (6300 - 6800) + 237.8840 = -262.116$

Exercise 8.2.4. The birth weight of the babies in developed and developing countries are normally distributed with mean μ_1, μ_2 and equal standard deviation σ. (My data is not real.) Suppose the following data about the birth weight from developed and developing nations were collected.

Developed				
8.8	8.1	6.3	9.7	6.3
7.1	5.3	7.7	9.1	8.1
8.2	7.9	8.3	8.9	9.0
10.1	9.9	8.8	7.8	5.2
7.2				

Developing				
6.3	5.2	8.3	5.9	5.5
7.1	8.1	7.9	6.3	6.9
9.1	8.1	7.0	4.9	5.3
6.3	7.1	6.3	6.1	5.8
5.7	6.8	8.3	7.7	

Compute the pooled estimate S_p of σ and a confidence interval for $\mu_1 - \mu_2$ at the 97 percent level of significance.

Solution. We will use Two Sample T-interval, because $\sigma = \sigma_1 = \sigma_2$ are unknown. As in chapter 2, enter data in a LIST of your TI-84, and summarize data, as follows:

	Population I X	**Population II** Y
sample size	$m = 21$	$n = 24$
sample mean	$\overline{X} = 7.9905$	$\overline{Y} = 6.75$
sample St. Dev.	$S_1 = 1.3758$	$S_2 = 1.1417$

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} = \sqrt{\frac{(21-1)1.3758^2 + (24-1)1.1417^2}{21+24-2}} = 1.2560$$

The degrees for freedom $df = m + n - 2 = 21 + 24 - 2 = 43$

Level of confidence = 97 percent. So
$1 - \alpha = .97, \quad \alpha = .03, \quad and \quad \alpha/2 = .015.$ Therefore,
$t_{m+n-2,\alpha/2} = t_{43,.015} = invT(.985, 43) = 2.2445$

$MOE = E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 2.2445 * 1.2560 \sqrt{\frac{1}{21} + \frac{1}{24}} = .8424$
$LEP = (\overline{X} - \overline{Y}) - E = (7.9905 - 6.75) - .8424 = 0.3981$
$REP = (\overline{X} - \overline{Y}) + E = (7.9905 - 6.75) + .8424 = 2.0829$

Exercise 8.2.5. African elephants and Indian elephants are different in height, weight, and length of ear and tusk. It is natural to assume that all these are normally distributed. Assume that the height of African and Indian elephants have an equal mean σ. The mean heights of African elephants and Indian elephants are μ_1, μ_2, respectively. Suppose the following data were collected on the height of elephants from the two continents (these are not real data).

African				
10.9	11.7	9.3	9.9	11.5
8.8	12.9	11.7	9.1	11.1
9.1	8.7	10.5	11.3	12.3
13.1	12.9	9.5	10.7	11.3

Indian				
7.1	8.3	8.2	9.1	10.3
9.3	9.7	8.9	8.8	9.1
7.9	9.9	9.2	8.8	8.1
8.7	8.8	9.3	10.1	9.9
9.9				

Compute the pooled estimate S_p of σ and a confidence interval for $\mu_1 - \mu_2$ at the 99 percent level of significance.

Solution. We will use Two Sample T-interval, because $\sigma = \sigma_1 = \sigma_2$ are unknown.
As in chapter 2, enter data in a LIST of your TI-84, and
summarize data, as follows:

	Population I X	**Population II** Y
sample size	$m = 20$	$n = 21$
sample mean	$\overline{X} = 10.815$	$\overline{Y} = 9.0190$
sample St. Dev.	$S_1 = 1.4162$	$S_2 = 0.8072$

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} = \sqrt{\frac{(20-1)1.4162^2 + (21-1)0.8072^2}{20+21-2}} = 1.1451$$

The degrees for freedom $df = m + n - 2 = 20 + 21 - 2 = 39$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$ Therefore,
$t_{m+n-2,\alpha/2} = t_{39,.005} = invT(.995, 39) = 2.7079$

$MOE = E = t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}} = 2.7079 * 1.1451 \sqrt{\frac{1}{20} + \frac{1}{21}} = 0.9688$
$LEP = (\overline{X} - \overline{Y}) - E = (10.815 - 9.0190) - 0.9688 = 0.8272$
$REP = (\overline{X} - \overline{Y}) + E = (10.815 - 9.0190) + 0.9688 = 2.7648$

8.3 Comparing Two Population Proportions

In this section, estimation of the difference $p_1 - p_2$ of the proportions of an attribute in two populations by confidence interval will be considered. Examples of such differences of proportions would include

1. the difference between the proportions $p_1 - p_2$ of the male and female populations who earn more than fifty thousand dollars annually;

2. the difference $p_1 - p_2$ of the proportions of defective items produced by the old machine and the new machine in a factory.

Consider proportions of an attribute A in two populations. Let p_1 and p_2 represent the proportions of the attribute A, in Population I and Population II, respectively. A confidence interval of $p_1 - p_2$ will be constructed. We proceed as follows.

1. A sample of size m from Population I is collected. Let X be the number of sample members that have the attribute A and $\overline{X} = \frac{X}{m}$ be the sample proportion that has the attribute A.

2. We take a sample from Population 2 of size n. Let Y be the number of sample members that has attribute A and $\overline{Y} = \frac{Y}{n}$ be the sample proportion that has the attribute A.

   (In other words, X is the number of "success" and $\overline{X} = \frac{X}{m}$ is the proportion of success in Population I sample. Similarly, Y and $\overline{Y} = \frac{Y}{n}$ for the Population II-sample.) Importantly, **these two samples are collected independently**.

3. BY CLT, we have, approximately,

$$\overline{X} \sim N(p_1, \sigma_1), \qquad \overline{Y} \sim N(p_2, \sigma_2) \qquad \text{where} \quad \begin{cases} \sigma_1 = \sqrt{\frac{p_1(1-p_1)}{m}} \\ \sigma_2 = \sqrt{\frac{p_2(1-p_2)}{n}} \end{cases}$$

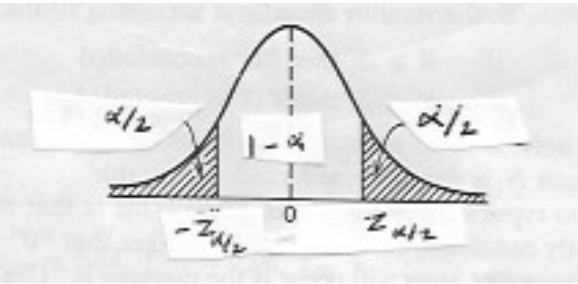
4. We assumed that the X-samples and Y-samples were drawn independently. In that case, it follows that, approximately,

$$\overline{X} - \overline{Y} \sim N(p_1 - p_2, \sigma) \quad \text{where} \quad \begin{cases} mean = E(\overline{X} - \overline{Y}) = p_1 - p_2 \\ st.\ dev = \sigma = \sqrt{\frac{\mathbf{p_1(1-p_1)}}{\mathbf{m}} + \frac{\mathbf{p_2(1-p_2)}}{\mathbf{n}}} \end{cases}$$

It follows,

$$Z = \frac{(\overline{X} - \overline{Y}) - (p_1 - p_2)}{\sigma} \sim N(0,1) \quad \text{has a st. normal distribution. Hence,}$$

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha \quad OR, \quad P\left(-z_{\alpha/2} \le \frac{(\overline{X} - \overline{Y}) - (p_1 - p_2)}{\sigma} \le z_{\alpha/2}\right) = 1 - \alpha$$



Simplifying,

$$P\left((\overline{X} - \overline{Y}) - z_{\alpha/2}\sigma \le p_1 - p_2 \le (\overline{X} - \overline{Y}) + z_{\alpha/2}\sigma\right) = 1 - \alpha \tag{8.3}$$

By Equation (7.2), this **seems what is needed** for a confidence interval, with $(1-\alpha)100$ percent confidence interval, for $p_1 - p_2$. However, $\sigma = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$ still depends on unknown parameter p_1, $\mathfrak{p}_2$. We use approximate

$$\begin{cases} p_1 \approx \overline{X} \\ p_2 \approx \overline{Y} \end{cases} \quad \text{and} \quad \sigma \approx \sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = \mathfrak{s}$$

(*We use the notation* $\approx$ *to abbreviate "approximately equal to".*)

So, as in the case of One proportion Z-interval (section 7.3), Equation 8.3, leads to an approximate confidence interval, as follows:

Theorem. We use the notations as above. A $(1-\alpha)100$ (approximate) percent confidence interval, for $p_1 - p_2$, is given by

$$\overline{X} - E \le p_1 - p_2 \le \overline{X} + E \qquad \text{where} \quad E = z_{\alpha/2}\mathfrak{s} = \sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}}$$

$$\begin{cases} \mathbf{E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}}} & \text{is called the Margin of Error (MOE)} \\ LEP = (\overline{X} - \overline{Y}) - E & \text{is called the Left end point} \\ REP = (\overline{X} - \overline{Y}) + E & \text{is called the Right end point} \end{cases}$$

In formally, this is called the Two proportion Z-Interval, for $p_1 - p_2$.

Example: Compare proportion of the male and female members in a community who earn more than \$50,000 annually. A sample of m male members are interviewed and X would be the number of those who make more than fifty thousand annually and $\overline{X} = \frac{X}{m}$ would be the sample proportion of those who make more than fifty thousand annually. Similarly, interview n female members and $\overline{Y} = \frac{Y}{n}$ would be the sample proportion of female members who make more than fifty thousand.

8.3.1 Problems: Two proportion Z-interval, for $p_1 - p_2$

Exercise 8.3.1. Two independent samples were collected from two populations to compare the proportions p_1, p_2 of an attribute A present, respectively, in these two populations. Use 95 percent confidence interval to estimate $p_1 - p_2$. It is given that $X = 55$ had the attribute A in a sample of size $m = 117$ from the first population and $Y = 37$ had the attribute A in a sample of size $n = 79$ from the second sample.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
Number of Success	$X = 55$	$Y = 37$
Sample size	$m = 117$	$n = 79$
Sample Proportion of success	$\overline{X} = \frac{55}{117} = 0.4701$	$\overline{Y} = \frac{37}{79} = 0.4684$

Level of confidence = 95 percent. So

$1 - \alpha = .95, \quad \alpha = .05, \quad and \quad \alpha/2 = .025.$ Therefore,

$z_{\alpha/2} = z_{.025} = invNorm(.975) = 1.9600$

$MOE = E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = 1.9600\sqrt{\frac{0.4701(1-0.4701)}{117} + \frac{0.4684(1-0.4684)}{79}} = .142435$
In this section, for the error term, we retain at least 6 decimal points.
$LEP = (\overline{X} - \overline{Y}) - E = (.4701 - .4684) - .142435 = .140735$
$REP = (\overline{X} - \overline{Y}) + E = (.4701 - .4684) + .142435 = .114135$

Exercise 8.3.2. To compare the proportions p_1, p_2 of defective lamps produced by new production center and old the production center, respectively, samples were collected. In a sample of 157 lamps from the new center, 26 were found to be defective; and in a sample of 141 lamps from the old center, 32 were defective. Compute a 99 percent confidence interval for $p_1 - p_2$.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
Number of Success	$X = 26$	$Y = 32$
Sample size	$m = 157$	$n = 141$
Sample Proportion of success	$\overline{X} = \frac{26}{157} = .1656$	$\overline{Y} = \frac{32}{141} = .2270$

Level of confidence = 99 percent. So,
$1 - \alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$ Therefore,
$z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

$MOE = E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = 2.5758\sqrt{\frac{.1656(1-.1656)}{157} + \frac{.2270(1-.2270)}{141}} = .118281$
In this section, for the error term, we retain at least 6 decimal points.
$LEP = (\overline{X} - \overline{Y}) - E = (.1656 - .2270) - .118281 = .179681$
$REP = (\overline{X} - \overline{Y}) + E = (.1656 - .2270) + .118281 = .056881$

Exercise 8.3.3. To compare the proportionsp_1, p_2 of men and women, respectively, who watch football, data was collected. In a sample of 199 men, 83 said that they watch football; and in a sample of 161 women, 51 said they watch football. (These are not real data.) Construct a 99 percent confidence interval for $p_1 - p_2$.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
Number of Success	$X = 83$	$Y = 51$
Sample size	$m = 199$	$n = 161$
Sample Proportion of success	$\overline{X} = \frac{83}{199} = .4171$	$\overline{Y} = \frac{51}{161} = .3168$

Level of confidence = 99 percent. So,
$1-\alpha = .99, \quad \alpha = .01, \quad and \quad \alpha/2 = .005.$ Therefore,
$z_{\alpha/2} = z_{.005} = invNorm(.995) = 2.5758$

$MOE = E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = 2.5758\sqrt{\frac{.4171(1-.4171)}{199} + \frac{.3168(1-.3168)}{161}} = .130804$
In this section, for the error term, we retain at least 6 decimal points.
$LEP = (\overline{X} - \overline{Y}) - E = (.4171 - .3168) - .130804 = -.030504$
$REP = (\overline{X} - \overline{Y}) + E = (.4171 - .3168) + .130804 = .231104$

Exercise 8.3.4. Two varieties of grapes are compared. To compare the proportions p_1, p_2 of acceptable grapes in these two varieties, respectively, samples were drawn. In a sample of 131 grapes from the variety I, 107 were acceptable. In a sample of 143 grapes from the variety II, 113 were acceptable. Construct a 97 percent confidence interval for the difference $p_1 - p_2$.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
Number of Success	$X = 107$	$Y = 113$
Sample size	$m = 131$	$n = 143$
Sample Proportion of success	$\overline{X} = \frac{107}{131} = .8168$	$\overline{Y} = \frac{113}{143} = .7902$

Level of confidence = 97 percent. So
$1-\alpha = .97, \quad \alpha = .03, \quad and \quad \alpha/2 = .015.$ Therefore,
$z_{\alpha/2} = z_{.015} = invNorm(.985) = 2.1701$

$MOE = E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = 2.1701\sqrt{\frac{.8168(1-.8168)}{131} + \frac{.7902(1-.7902)}{143}} = .104111$
In this section, for the error term, we retain at least 6 decimal points.
$LEP = (\overline{X} - \overline{Y}) - E = (.8168 - .7902) - .104111 = -.077511$
$REP = (\overline{X} - \overline{Y}) + E = (.8168 - .7902) + .104111 = .130711$

Exercise 8.3.5. To compare the proportions p_1, p_2 of students, respectively, in two state universities who pay more than \$15 K tuition per year, samples were collected. In a sample of 217 students in the university I, 129 paid more than \$15 K. In a sample of 313 students in the university II, 158 paid more than \$15 K. Construct a 98 percent confidence interval for the difference $p_1 - p_2$.

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
Number of Success	$X = 129$	$Y = 158$
Sample size	$m = 217$	$n = 313$
Sample Proportion of success	$\overline{X} = \frac{129}{217} = .5945$	$\overline{Y} = \frac{158}{313} = .5048$

Level of confidence = 98 percent. So
$1 - \alpha = .98, \quad \alpha = .02, \quad and \quad \alpha/2 = .01.$ Therefore,
$z_{\alpha/2} = z_{.01} = invNorm(.99) = 2.3263$

$MOE = E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = 2.3263\sqrt{\frac{.5945(1-.5945)}{217} + \frac{.5048(1-.5048)}{313}} = .101656$
In this section, for the error term, we retain at least 6 decimal points.
$LEP = (\overline{X} - \overline{Y}) - E = (.5945 - .5048) - .101656 = -.011956$
$REP = (\overline{X} - \overline{Y}) + E = (.5945 - .5048) + .101656 = .191356$

Exercise 8.3.6. To compare the proportions p_1, p_2 of college graduates who earn more than 50 K, in two states, data was collected. In a sample of 444 college graduates in the state I, 334 earn more than 50 K. In a sample of 546 college graduates in the state II, 414 earn more than 50 K. Construct a 96 percent confidence interval for the difference $p_1 - p_2$.

:

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
Number of Success	$X = 334$	$Y = 414$
Sample size	$m = 444$	$n = 546$
Sample Proportion of success	$\overline{X} = \frac{334}{444} = .7523$	$\overline{Y} = \frac{414}{546} = .7582$

Level of confidence = 96 percent. So
$1 - \alpha = .96, \quad \alpha = .04, \quad and \quad \alpha/2 = .02.$ Therefore,
$z_{\alpha/2} = z_{02} = invNorm(.98) = 2.0537$

$MOE = E = z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{m} + \frac{\overline{Y}(1-\overline{Y})}{n}} = 2.0537\sqrt{\frac{.7523(1-.7523)}{444} + \frac{.7582(1-.7582)}{546}} = .056448$
In this section, for the error term, we retain at least 6 decimal points.
$LEP = (\overline{X} - \overline{Y}) - E = (.7523 - .7582) - .056448 = -.062343$
$REP = (\overline{X} - \overline{Y}) + E = (.7523 - .7582) + .056448 = .050548$

Exercise 8.3.7. It is believed that women are safer drivers than men. Let p_1, p_2 denote the proportions of women and men drivers, respectively, who were involved in an auto

accident in a year period. In a sample of a size 739 women drivers 39 were involved in auto accident during this period. During the same period, in a sample of size 1215 men 79 were involved in auto accident in a year. Construct a 95 percent confidence interval for the difference $p_1 - p_2$.

Chapter 9

Significance Test

9.1 Introduction and Jargon

The Testing of hypotheses is another approach to estimation of parameters. These are also called Significance Tests. A hypothesis H_0, called the **Null hypothesis**, is tested against another hypothesis H_A, called the **alternative hypothesis**. Only one of these two hypotheses is true. Based on the collected sample and established testing criterion, one of them is accepted and the other one rejected. The following two examples would provide further insight.

Example 1. An assertion is made that the disparity between the wages (annual income) of working men and women does not exist any more. To test this assertion, the mean annual incomes μ_1, μ_2, respectively, of the working male and female populations were compared. Our Null hypothesis H_0 would be that the mean annual income μ_1 of the working male population would be higher than the mean annual income μ_2 of the working female population. The Alternative Hypothesis H_A would be, as the assertion suggests, that these two means would be equal. We write them formally as:

$$\begin{cases} H_0: & \mu_1 - \mu_2 > 0 \\ H_A: & \mu_1 - \mu_2 = 0 \end{cases}$$

Example 2. A TV commentator mentioned that, during the last decade, the life expectancy of human being has increased substantially from 75 years. To test this assertion, the mean life expectancy μ was compared with 75. The Null hypothesis H_0 would be that the mean life expectancy μ remains equal to 75, as it was before. The Alternative Hypothesis H_A would be that, as the assertion suggests, the mean μ rose above 75 year by now. We write them formally as:

$$\begin{cases} H_0: & \mu = 75 \\ H_A: & \mu > 75 \end{cases}$$

Following are some definitions and terminologies.

1. A **statistical hypothesis** is defined to be a statement, claim, or proposition regarding a population. Usually, it would be about the values of the population parameters. The hypotheses H_0 and H_A in the above two examples would examples of statistical hypotheses.

2. It would be important to distinguish which one would be the Null hypothesis and which one would be the alternative hypothesis in a given context. One of them would, essentially, be the negation of the other.

3. The **Null hypothesis** H_0 represents the status quo or the norm. It would be the conventional wisdom. It represents something that was accepted for a long time, or some assumption or method that has been working reliably for a long time. Null hypothesis would remain as the default, unless the collected data provides **very strong evidence** against it, in favor of the alternative. There is a clear **bias in favor** of the Null Hypothesis.

   The **alternative hypothesis** represents a **new claim** or something **out of the ordinary**. It could be a researcher's new technology or some sales person's claim. The bar for acceptance of the Alternative Hypothesis is very high. The burden of proof of its validity belongs to those who asserts the same. There may even be resistance or skepticism about its validity. It would be accepted only if there is **very strong evidence**, in the collected data, in its support.

   There are reasons for such favoritism in favor of Null Hypothesis. This is because an incorrect decision to reject the Null may have more serious consequences than rejecting the Alternative incorrectly. For example, in any medical test, erroneously concluding that the patient does not have an ailment would have more grievous consequences than erroneously concluding that the patient has the same ailment. In particular,

   (a) For a COVID-19 test, priority would be to minimize the chances (probability) of erroneously concluding that patient does not have the infection, while he/she has the infection (**false negative**). While it is less serious, that the test result erroneously concludes that the patient has been infected, while he/she did not (**false positive**).

   (b) For a pregnancy test, the priority would be to minimize the chances (probability) of erroneously concluding that one is not pregnant when one is indeed pregnant (**false negative**), than the converse (**false positive**). Common sense dictates that such a test could only allow a maximum of five percent of such erroneous conclusions (**false negative**).

4. Given a Null hypothesis H_0 and an alternative hypothesis H_A, a **test of hypothesis** is a rule or a procedure to decide, based on the collected sample, whether to accept H_0 or H_A. The test will be based on the value of a test statistic. The rule is also called the **decision rule**. A test of hypothesis is also known as a **Significance Test**.

5. **Two Types of errors. In** such testing of hypotheses, two types of mistaken conclusions (errors) are possible as follows.

   (a) Rejecting the Null H_0 when it is in fact true would be called the **type one error**. The analogy would be a false negative.

   (b) Accepting the Null H_0 when it is in fact false would be called the **type two error**. The corresponding analogy would be a false positive.

   (c) The probability of type one error would be called the **level of significance**. It would be denoted by α. Since the priority would be to minimize the frequency of false negative, α would be a small number. Most often, $\alpha = .1, .05, .01$ or a small number.

The rest of this chapter would be analogous to chapter 7, 8. Corresponding to each interval estimation we considered, there would consider one Significance Test.

9.1.1 Design Decision rules: for μ, when σ is known

Let X be a random variable with mean μ and standard deviation σ. Some of our hypotheses testing would look like the following.

Two Tail Test	Left Tail Test	Right Tail Test
$\begin{cases} H_0: & \mu = 75 \\ H_A: & \mu \neq 75 \end{cases}$	$\begin{cases} H_0: & \mu = 75 \\ H_A: & \mu < 75 \end{cases}$	$\begin{cases} H_0: & \mu = 75 \\ H_A: & \mu > 75 \end{cases}$

More generally, they would look like one of the following.

Two Tail Test	Left Tail Test	Right Tail Test
$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu < \mu_0 \end{cases}$	$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu > \mu_0 \end{cases}$

(9.1)

Analogous to Z-Interval, we develop a significance test, of μ, when σ is known, for all the cases (9.1). We proceed as follows:

1. Draw a sample $X_1, X_2, \ldots, X_n$ from the X-population, of size n. Let

   $$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad \text{denote the sample mean.}$$

   By CLT

   $$\overline{X} \sim N\left(\mu, \sigma_{\overline{X}}\right) \quad \text{where} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

2. By increasing the sample size m, both type one and type two errors can be controlled. Once the sample size is fixed, it is not possible to control both simultaneously. As one of them is minimized the other one goes up. As indicated above, **priority would be to control the probability** $P(\textit{type one error}) = \alpha$. Therefore, we proceed to develop a Test of Significance, at the level of significance α.

3. Assume $H_0 : \mu = \mu_0$ is true. Then

$$Z = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} \sim N(0,1) \quad \text{has st. normal distribution, where } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{9.2}$$

Expression Z above would be called a **test statistic**. We use lower case, to denote

$$z = \frac{\overline{x} - \mu_0}{\sigma_{\overline{X}}} = \frac{(\overline{x} - \mu_0)\sqrt{n}}{\sigma} \quad \text{the observed value of} \quad Z. \tag{9.3}$$

For three cases of alternate hypotheses, as in (9.1), we formulate the decision rules:

(a) **Two Tail Test:** When H_0 is true, by (9.2),

$$\begin{cases} P\left(|Z| \le z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha, & OR \\ P\left(|Z| > z_{\alpha/2}\right) = \alpha \end{cases} \tag{9.4}$$

In this case $H_A : \mu \neq \mu_0$. So, it would be reasonable to reject H_0, when $\left|\overline{X} - \mu_0\right|$ is large.

i. **Reject** H_0, if the observed (absolute) value $|Z| = |z|$ is large.

ii. Accept H_0, if the observed (absolute) value $|Z| = |z|$ is small and

Before we state the decision rules, we define p-value, for this test, as

$$\mathbf{p} = P\left(|Z| > |z|\right) = 1 - normalcdf\left(-|z|, |z|\right) \tag{9.5}$$

Because of equation 9.4, **at level of significance** α, we set our decision rule, as follows:

$$\begin{cases} \textbf{Reject } H_0 & if \ |z| > z_{\alpha/2} \\ \textbf{Accept } H_0 & if \ |z| \le z_{\alpha/2} \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if \ |z| > z_{\alpha/2} \\ \textbf{Reject } H_A & if \ |z| \le z_{\alpha/2} \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \ge \alpha \end{cases} \tag{9.6}$$

(b) **Left Tail Test:** When H_0 is true, by (9.2),

$$P\left(Z < -z_\alpha\right) = \alpha \tag{9.7}$$

In this case $H_A : \mu < \mu_0$. So, it would be reasonable to reject H_0, when $\overline{X}$ is much smaller than μ_0. So,

i. **Reject** H_0, if the observed (absolute) value $Z = z$ is small (negative) enough.

ii. Accept H_0, otherwise.

we define p-value, for this test, as

$$\mathbf{p} = P(Z < z) = normalcdf(-5, z). \tag{9.8}$$

Because of equation 9.7, **at level of significance** α, we set our decision rule, as follows:

$$\begin{cases} \textbf{Reject } H_0 & if\ z < -z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\ z < -z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.9}$$

(c) **Right Tail Test:** When H_0 is true, by (9.2),

$$P\left(z_\alpha < Z\right) = \alpha \tag{9.10}$$

In this case $H_A : \mu > \mu_0$. So, it would be reasonable to reject H_0, when $\overline{X}$ is much higher than μ_0. So,

i. **Reject** H_0, if the observed (absolute) value $Z = z$ is large (positive) enough.

ii. Accept H_0, otherwise.

we define p-value, for this test, as

$$\mathbf{p} = P(Z > z) = normalcdf(z, 5). \tag{9.11}$$

Because of equation 9.7, **at level of significance** α, we set our decision rule, as follows:

$$\begin{cases} \textbf{Reject } H_0 & if\ z > z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if z > z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.12}$$

Universal Decision rule. Consider three significance tests (9.1). In all three cases, the **p-value based decision rules** (9.6, 9.9, 9.12) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.13}$$

However, p-values $\mathbf{p}$ are defined differently (9.5, 9.35, 9.11), are specific to the respective tests. These three significance tests, for μ, when the value of σ is known, is also called **Z-test**.

9.1.2 Problems: Z-Test

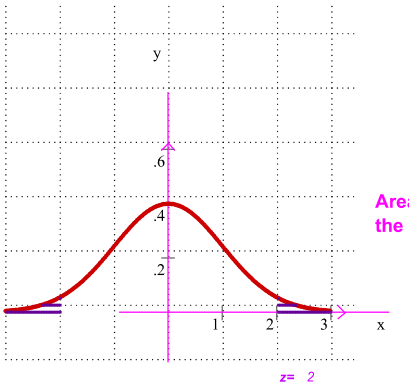
Exercise 9.1.1. It is speculated that the mean life expectancy in a certain community, is not equal to 75 years. It is known that the standard deviation of life expectancy of a population is $\sigma = 15$ years. A a sample of size 25 had mean life expectancy $\overline{X} = 81$ years. Perform a significance test for the null and alternative hypothesis, regarding the mean life expectancy μ:

$$\begin{cases} H_0: & \mu = 75 \\ H_A: & \mu \neq 75 \end{cases}$$

1. Compute the value of the test statistic Z.
2. Compute the p-value.
3. At the 5 percent level of significance will you reject or accept the null hypothesis?
4. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.
Here the population standard deviation $\sigma = 15$,
the sample size $n = 25$,
the sample mean $\overline{X} = 81$,
Also, $\mu_0 = 75$.

1. The test statistic $Z = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma} = \frac{(81-75)\sqrt{25}}{15} = 2$.
2. This a a **two tail** test. So, by (9.5), the p-value
   $\mathbf{p} = P(|Z| > |z|) = 1 - normalcdf(-|z|, |z|)$
   $= 1 - normalcdf(-2, 2) = 1 - .9545 = .0455$.
3. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .0455 < \alpha = .05$,
   we REJECT the null hypothesis at 5 percent level of significance. That means,
   at five percent level of significance, we accept that the mean life expectancy $\mu \neq 75$.
4. Since $\mathbf{p} = .0455$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent; 5 percent ($\alpha = .05$) would be the lowest level at which we would reject the null hypothesis.



Area of the shaded region is the p-value

z = 2 p-value = 0.0454

z α/2 = 1.96 α = 0.05

We change Ex. 9.1.1, and do a right tail test.

Exercise 9.1.2. It is speculated that the mean life expectancy in a certain community, is significantly higher than 75 years. It is known that the standard deviation of life expectancy of a population is $\sigma = 15$ years. A a sample of size 25 had mean life expectancy $\overline{X} = 81$ years. Accordingly, perform a significance test for the null and alternative hypothesis, regarding the mean life expectancy μ:

$$\begin{cases} H_0: & \mu = 75 \\ H_A: & \mu > 75 \end{cases}$$

1. Compute the value of the test statistic Z.
2. Compute the p-value.
3. At the 5 percent level of significance will you reject or accept the null hypothesis?

4. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.

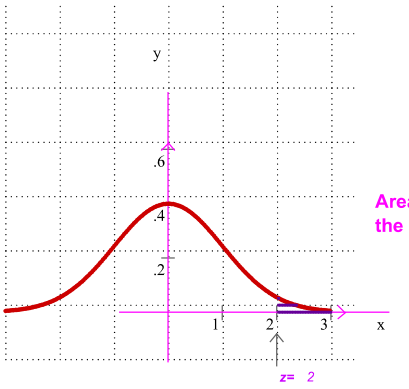
Here the population standard deviation $\sigma = 15$,
the sample size $n = 25$,
the sample mean $\overline{X} = 81$,
Also, $\mu_0 = 75$.

1. The test statistic $Z = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma} = \frac{(81-75)\sqrt{25}}{15} = 2$.

2. This a a **right tail** test. So, by (9.11), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5) = .02275$

3. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .02275 < \alpha = .05$,
   we REJECT the null hypothesis at 5 percent level of significance. That means,
   at five percent level of significance, we accept that the mean life expectancy is higher:
   $\mu > 75$.

4. Since $\mathbf{p} = .02275$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent; 3 percent ($\alpha = .03$) would be the lowest level at which we would reject the null hypothesis (i.e. accept the alternative $\mu > 75$).



Area of the shaded region is the p-value

z = 2 p-value = 0.0226

z_α = 1.65 α = 0.05

Exercise 9.1.3. The time taken by an athlete to run an event is normally distributed with mean μ and known standard deviation $\sigma = 3.5$ seconds. The coach believes that his/her mean time μ has improved from last year's mean 34 seconds. To test, the athlete ran 16 times and the sample mean was found to be $\overline{X} = 31$ seconds.

1. Formulate the null and alternative hypotheses to perform a significance test for coach's belief.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 5 percent level of significance will you reject or accept the null hypothesis (or that his/her time has improved or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.
Here the population standard deviation $\sigma = 3.5$,
the sample size $n = 16$,
the sample mean $\overline{X} = 31$,

1. The alternative hypothesis is the coach's belief: $H_A : \mu < 34$. The null and alternative hypotheses are:
$$\begin{cases} H_0 : & \mu = 34 \\ H_A : & \mu < 34 \end{cases} \qquad \text{so,} \qquad \mu_0 = 34$$

2. The test statistic $Z = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma} = \frac{(31-34)\sqrt{16}}{3.5} = -3.4286$.

3. This a a **left tail** test. So, by (9.35), the p-value
$\mathbf{p} = P(Z < z) = normalcdf(-5, z) = normalcdf(-5, -3.4286) = 3.0311 * 10^{-4}$

4. Five percent level of significance means $\alpha = .05$. Since,
p-value $\mathbf{p} = 3.0311 * 10^{-4} < \alpha = .05$,
we REJECT the null hypothesis at 5 percent level of significance. That means, at five percent level of significance, we accept that the mean time is: $\mu < 34$.

5. Since $\mathbf{p} = 3.0311 * 10^{-4}$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent;
.1 percent ($\alpha = .001$) would be the lowest level at which we would reject the null hypothesis (i.e. accept the alternative $\mu < 34$).

Exercise 9.1.4. The effectiveness of a weight loss program is to be tested on a group of 83 participants. At the beginning of the program, the mean weight of group is 210 pounds. At the end of the program the mean weight of the group is 199 pounds. The standard deviation of weight is known to be $\sigma = 53.1$ pounds. In terms of mean weight μ, perform a significance test that the program is effective.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 2 percent level of significance will you reject or accept the null hypothesis (or that the program is effective or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.
Here the population standard deviation $\sigma = 53.1$,
the sample size $n = 83$,
the sample mean $\overline{X} = 199$,

1. he alternative hypothesis is program was effective to reduce the mean weight from 210 pounds:$H_A : \mu < 210$. The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu = 210 \\ H_A : & \mu < 210 \end{cases} \qquad \text{so,} \qquad \mu_0 = 210$$

2. The test statistic $Z = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma} = \frac{(199-210)\sqrt{83}}{53.1} = -1.8872$.

3. This a a **left tail** test. So, by (9.35), the p-value
$\mathbf{p} = P(Z < z) = normalcdf(-5, z) = normalcdf(-5, -1.8872) = .0296$.

4. Two percent level of significance means $\alpha = .02$. Since,
p-value $\mathbf{p} = .0296 \not< \alpha = .02$,
we ACCEPT the null hypothesis at two percent level of significance.
That means, at two percent level of significance, we DO NOT accept that the weight loss program is effective.

5. Since $\mathbf{p} = .0296$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent; 3 percent ($\alpha = .03$) would be the lowest level at which we would reject the null hypothesis (i.e. accept the alternative).

Exercise 9.1.5. A manufacturer of heating furnace is marketing a new model of energy efficient furnace. The mean gas consumption in January by ordinary furnaces is 153 CCF. A sample of 93 new model furnace had a mean consumption of 142 CCF in January. The standard deviation of consumption in January is known to be $\sigma = 46$ CCF. In terms of mean consumption μ, perform a significance test that the new model is really energy efficient.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 1 percent level of significance will you reject or accept the null hypothesis (or that the new model is energy efficient or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.
Here the population standard deviation $\sigma = 46$,
the sample size $n = 93$,
the sample mean $\overline{X} = 142$,

1. The alternative hypothesis is the claim of the manufacturer: $H_A : \mu < 153$. The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu = 153 \\ H_A : & \mu < 153 \end{cases} \qquad \text{so,} \qquad \mu_0 = 153$$

2. The test statistic $Z = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma} = \frac{(142-153)\sqrt{93}}{46} = -2.3061$.

3. This a a **left tail** test. So, by (9.35), the p-value
$\mathbf{p} = P(Z < z) = normalcdf(-5, z) = normalcdf(-5, -2.3061) = .01055$.

4. One percent level of significance means $\alpha = .01$. Since,
p-value $\mathbf{p} = .01055 \not< \alpha = .01$,
we ACCEPT the null hypothesis at two percent level of significance.
That means, at one percent level of significance, we do not accept that the this model is energy efficient.

5. Since $\mathbf{p} = .01055$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis (i.e. accept the alternative).

Exercise 9.1.6. It is believed that due to favorable weather conditions the mean weight μ of King salmon in Anchor River would be higher than the last year's mean of 33 pounds . The standard deviation of the weight is known to be $\sigma = 16$ pounds. A catch of 53 King had a mean of 39 pounds. In terms of mean weight μ, perform a significance test that the weight would be higher.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic Z.
3. Compute the p-value.
4. At the 2 percent level of significance will you reject or accept the null hypothesis (or that the mean weight has increased or not)?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.
Here the population standard deviation $\sigma = 16$,
the sample size $n = 53$,
the sample mean $\overline{X} = 39$,

1. The alternative hypothesis is that the mean weight has increased: $H_A : \mu > 33$. The null and alternative hypotheses are:
$$\begin{cases} H_0 : & \mu = 33 \\ H_A : & \mu > 33 \end{cases} \qquad \text{so,} \qquad \mu_0 = 33$$
2. The test statistic $Z = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma} = \frac{(39-33)\sqrt{53}}{16} = 2.7300$.
3. This a a **right tail** test. So, by (9.11), the p-value
$\mathbf{p} = P(Z > z) = normalcdf(z, 5) = normalcdf(2.7300, 5) = .0032$.
4. Two percent level of significance means $\alpha = .02$. Since,
p-value $\mathbf{p} = .0032 < \alpha = .02$,
we REJECT the null hypothesis at 2 percent level of significance.
That means, at two percent level of significance, we accept that the mean weight has increased.

5. Since $\mathbf{p} = .0032$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent; .5 percent ($\alpha = .005$) would be the lowest level at which we would reject the null hypothesis (i.e. accept the alternative).

Exercise 9.1.7. The instructor of Math 365 claims that due to his updated method of teaching, the student's learning has improved. The mean percent score of all his Math 365 courses before this semester was 68 percent. This semester in his call of 79 students, the mean percent score is 74 percent. The standard deviation of the percent score is known to be $\sigma = 22$ percent. In terms of mean consumption μ, perform a significance test that the percent score is higher.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic Z.
3. Compute the p-value.
4. At the 2 percent level of significance will you reject or accept the null hypothesis (or that the student's learning has improved or not)?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution.
Here the population standard deviation $\sigma = 22$,
the sample size $n = 79$,
the sample mean $\overline{X} = 74$,

1. The alternative hypothesis is that the mean percent score has increased from 68 percent: $H_A : \mu > 68$.
   The null and alternative hypotheses are:
$$\begin{cases} H_0 : & \mu = 68 \\ H_A : & \mu > 68 \end{cases} \qquad \text{so,} \qquad \mu_0 = 68$$
2. The test statistic $Z = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma} = \frac{(74-68)\sqrt{79}}{22} = 2.4241$.
3. This a a **right tail** test. So, by (9.11), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5) = normalcdf(2.4241, 5) = .0077$.

4. Two percent level of significance means $\alpha = .02$. Since,
   p-value $\mathbf{p} = .0077 < \alpha = .02$,
   we REJECT the null hypothesis at 2 percent level of significance.
   That means, at two percent level of significance, we accept that the mean percent score has increased.

5. Since $\mathbf{p} = .0032$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent; 1 percent (because $.0077 < \alpha = .01$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.1.8. It is believed that the annual mean expenditure, including tuition, for students has increased from the corresponding mean in year 2000. In year 2000, the mean annual expenditure was \$17,000. A sample of 87 students had annual mean expenditure of \$19,500. The standard deviation annual expenditure is known to be $\sigma = \$7,500$. In terms of mean expenditure μ, perform a significance test that the mean annual expenditure μ has increased.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 2 percent level of significance will you reject or accept the null hypothesis (or that the expenditure has or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

9.2 Significance Test for μ: Unknown σ

This section is analogous to section 7.2, on T-intervals. So, the significance tests we develop will be called T-test. As in the case of Z-tests (section 9.1.1), we would do three tests:

Two Tail Test	Left Tail Test	Right Tail Test
$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu < \mu_0 \end{cases}$	$\begin{cases} H_0: & \mu = \mu_0 \\ H_A: & \mu > \mu_0 \end{cases}$

(9.14)

We design the decision rules, as in the case Z-test (section 9.1.1). We refrain from repeating some of the reasoning used above (section 9.1.1) and we only give the decision rules.

1. Similar to section 7.2, we assume that $X \sim N(\mu, \sigma)$, and σ is also unknown.

2. Draw a sample $X_1, X_2, \ldots, X_n$ from the X-population, of size n. Let

$$\begin{cases} \overline{X} = \frac{X_1+X_2+\cdots+X_n}{n} & = \text{the sample mean.} \\ S = \sqrt{\frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n-1}} & = \text{the standard deviation.} \end{cases} \tag{9.15}$$

3. Assume $H_0 : \mu = \mu_0$ is true. Then

$$T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1} \quad \text{has } t \text{ distribution, with degrees of freedom } n-1 \tag{9.16}$$

Expression T above would be called a **test statistic**. We use lower case, to denote

$$t = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s} \quad \text{the observed value of} \quad T. \tag{9.17}$$

For three cases of alternate hypotheses, as in (9.14), we formulate the decision rules:

(a) **Two Tail Test:** When H_0 is true, by (9.16),

$$\begin{cases} P\left(|T| \le t_{n-1,\alpha/2}\right) = P\left(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}\right) = 1-\alpha, & OR \\ P\left(|T| > t_{n-1,\alpha/2}\right) = \alpha \end{cases} \tag{9.18}$$

Before we state the decision rules, we define p-value, for this test, as

$$\mathbf{p} = P\left(|T| > |t|\right) = 1 - tcdf\left(-|t|, |t|, n-1\right) \tag{9.19}$$

Because of equation 9.18, **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if \ |t| > t_{n-1,\alpha/2} \\ \textbf{Accept } H_0 & if \ |t| \le t_{n-1,\alpha/2} \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if \ |t| > t_{n-1,\alpha/2} \\ \textbf{Reject } H_A & if \ |t| \le t_{n-1,\alpha/2} \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \ge \alpha \end{cases} \tag{9.20}$$

(b) **Left Tail Test:** When H_0 is true, by (9.16),

$$P\left(T < -t_{n-1,\alpha}\right) = \alpha \tag{9.21}$$

We define p-value, for this test, as

$$\mathbf{p} = P(T < t) \approx tcdf(-5, t, n-1). \tag{9.22}$$

Because of equation 9.21, **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if \ t < -t_{n-1,\alpha} \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if \ t < -t_{n-1,\alpha} \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \ge \alpha \end{cases} \tag{9.23}$$

(c) **Right Tail Test:** When H_0 is true, by (9.16),

$$P\left(t_{n-1,\alpha} < T\right) = \alpha \tag{9.24}$$

We define p-value, for this test, as

$$\mathbf{p} = P(T > t) \approx tcdf(t, 5, m-1)\ . \tag{9.25}$$

Because of equation 9.24, **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if\ t > t_{n-1,\alpha} \\ \textbf{Accept } H_0 & Otherwise \end{cases} \quad \text{Equivalently,} \begin{cases} \textbf{Accept } H_A & if\ t > t_{n-1,\alpha} \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.26}$$

These three significance tests, for μ, when the value of σ is not known, is also called **T-test**.

Universal Decision rule. As in the Z-test, the **p-value based decision rules** (9.20, 9.23, 9.26) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \quad \text{which is even same as (9.13)} \tag{9.27}$$

However, p-values $\mathbf{p}$ are defined differently (9.19, 9.22, 9.25), are specific to the respective tests. This similarity, regarding p-value based decision rule **will be repeated for all the subsequent sections, as well**.

9.2.1 Problems: T-Test

Exercise 9.2.1. A supplier of lamps claims that the mean lifetime of his lamps is longer than that of the lamps in the market. The mean lifetime of the bulbs on the market is 3456 hours. To test the claim of the supplier, a sample of 26 bulbs were examined. The sample mean was found to be 3720 hours and the sample standard deviation was $s = 552$ hours. In terms of mean lifetime μ, perform a significance test that the supplier's lamps last longer.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic T.
3. Compute the p-value.

4. At the 5 percent level of significance will you reject or accept the null hypothesis (or that the mean lifetime of the lamps is higher or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The population standard deviation is not known. So, the T-test will be used.
We summarize the given data:
The sample size $n = 26$,
The sample mean $\overline{X} = 3720$,
The sample standard deviation $S = 552$,
Degrees of freedom, $df = n - 1 = 26 - 1 = 25$

1. The alternative hypothesis is that mean lifetime μ of the supplier's lamps is higher than 3456 hours: $H_A : \mu > 3456$.
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu = 3456 \\ H_A : & \mu > 3456 \end{cases} \qquad \text{so,} \qquad \mu_0 = 3456$$

2. The test statistic $T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{s} = \frac{(3720-3456)\sqrt{26}}{552} = 2.4387$.

3. This a a **right tail** test. So, by (9.25), the p-value
   $\mathbf{p} = P(T > t) \approx tcdf(t, 5, n-1) \ = cdf(2.4387, 5, 25) = .0111$.

4. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .0111 < \alpha = .05$, we REJECT the null hypothesis at 5 percent level of significance.
   That means, at five percent level of significance, we accept that the supplier's claim.

5. Since p-value $\mathbf{p} = .0111$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.2.2. It is believed that the mean length of babies at birth in the United States is higher than the mean of 16.7 inches in some other nation. A sample of 33 babies in the United States was collected, and the sample mean and standard deviation was found to be $\overline{X} = 19$ inches, $S = 5.5$ inches. Perform a of significance test for this belief as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic T.
3. Compute the p-value.
4. At the 1 percent level of significance will you reject or accept the null hypothesis (or that the birth length is higher or not)?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The population standard deviation is not known. So, the T-test will be used.
Given data summarize, as follows :
The sample size $n = 33$,
The sample mean $\overline{X} = 19$,
The sample standard deviation $S = 5.5$,
Degrees of freedom, $df = n - 1 = 33 - 1 = 32$

1. The alternative hypothesis is that mean birth length μ in US is higher than 16.7 inches: $H_A : \mu > 16.7$.
   The null and alternative hypotheses are:
   $$\begin{cases} H_0 : & \mu = 16.7 \\ H_A : & \mu > 16.7 \end{cases} \qquad \text{so,} \qquad \mu_0 = 16.7$$
2. The test statistic $T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{s} = \frac{(19-16.7)\sqrt{33}}{5.5} = 2.4023$.
3. This a a **right tail** test. So, by (9.25), the p-value
   $\mathbf{p} = P(T > t) \approx tcdf(t, 5, n-1) \ = tcdf(2.4023, 5, 32) = .0111$.
4. One percent level of significance means $\alpha = .01$. Since,
   p-value $\mathbf{p} = .0111 \not< \alpha = .01$, we ACCEPT the null hypothesis at 1 percent level of significance.
   That means, at one percent level of significance, we do NOT accept that the mean birth length μ is longer than 16.7 .
5. Since p-value $\mathbf{p} = .0111$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.2.3. A car manufacturer claims that a new model of car will get more mileage per gallon than the old model. The old model gets a mean mileage of 33 miles per gallon. To test the claim, 19 cars from the new model were tested and the sample mean was found to be $\overline{X} = 35$ miles and standard deviation $S = 3.3$ miles. Perform a significance test for this manufacturer's claim as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic T.

3. Compute the p-value.

4. At the 1 percent level of significance will you reject or accept the null hypothesis (or that the milage is higher or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The population standard deviation is not known. So, the T-test will be used.
Given data summarize, as follows :
The sample size $n = 19$,
The sample mean $\overline{X} = 35$,
The sample standard deviation $S = 3.3$,
Degrees of freedom, $df = n - 1 = 19 - 1 = 18$

1. The alternative hypothesis is that mean milage μ per gallon for the new model is higher than 33 miles: $H_A : \mu > 33$.
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu = 33 \\ H_A : & \mu > 33 \end{cases} \qquad \text{so,} \qquad \mu_0 = 33$$

2. The test statistic $T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{s} = \frac{(35-33)\sqrt{19}}{3.3} = 2.6218$.

3. This a a **right tail** test. So, by (9.25), the p-value
   $\mathbf{p} = P(T > t) \approx tcdf(t, 5, n-1) \ = tcdf(2.6218, 5, 18) = .0086$

4. One percent level of significance means $\alpha = .01$. Since,
   p-value $\mathbf{p} = .0086 < \alpha = .01$, we REJECT the null hypothesis at 1 percent level of significanc That means, at one percent level of significance, we accept that the mean milage μ is higher than 33 miles .

5. Since p-value $\mathbf{p} = .0086$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 1 percent ($\alpha = .01$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.2.4. Consider the following data , on life time X of light bulbs produced in a factory.

5110	4671	6441	3331	5055	5270	5335	4973	1837	5487
7783	4560	6074	4777	4707	5263	4978	5418	5123	5017

It is natural to assumed $X \sim N(\mu, \sigma)$. The mean lifetime for an average light bulb on the market is 4500 hours. The producer claims that the mean lifetime of the bulbs is more than the average bulbs on the market. Perform a significance test for this claim.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic T.

3. Compute the p-value.

4. At the 2 percent level of significance will you reject or accept the null hypothesis (or that the lifetime is higher or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The population standard deviation is not known. So, the T-test will be used.
Use TI-84, as in chapter 2, to summarize the raw data:
Here the sample size $n = 20$,
the sample mean $\overline{X} = 5060.5$,
Sample standard deviation $S = 1143.1106$
The degrees of freedom $df = n - 1 = 20 - 1 = 19$

1. The alternative hypothesis is that the mean lifetime μ of the lamps is higher than 4500 hours: $H_A : \mu > 4500$.

The null and alternative hypotheses are:

$$\begin{cases} H_0: & \mu = 4500 \\ H_A: & \mu > 4500 \end{cases} \qquad \text{so,} \qquad \mu_0 = 4500$$

2. The test statistic $T = \frac{(\overline{X}-\mu_0)\sqrt{n}}{s} = \frac{(5060.5-4500)\sqrt{20}}{1143.1106} = 2.1928$.

3. This a a **right tail** test. So, by (9.25), the p-value
$\mathbf{p} = P(T > t) \approx tcdf(t, 5, n-1) \ = tcdf(2.1928, 5, 19) = .0204$.

4. Two percent level of significance means $\alpha = .02$. Since,
p-value $\mathbf{p} = .0204 \not< \alpha = .02$, we ACCEPT the null hypothesis at 2 percent level of significance. That means, at two percent level of significance, we do NOT accept that the mean lifetime μ is higher than 4500 hours.

5. Since p-value $\mathbf{p} = .0204$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 3 percent ($\alpha = .03$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.2.5. Consider the data, in Ex. 7.2.8, on weight (in pounds) of salmon in a river. It is suspected that, due to pollution, the mean weight has reduced from last year's the mean weight 37 pounds. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic T.

3. Compute the p-value.

4. At the 10 percent level of significance will you reject or accept the null hypothesis (or that the mean weight has reduced or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The population standard deviation is not known. So, the T-test will be used.
Use TI-84, as in chapter 2, to summarize the raw data:
The sample size $n = 19$,
The sample mean $\overline{X} = 34.3737$,
The sample standard deviation $S = 6.7608$,
Degrees of freedom, $df = n - 1 = 19 - 1 = 18$

1. The alternative hypothesis is that the mean weight μ has reduced from 37 pounds: $H_A : \mu < 37$.
   The null and alternative hypotheses are:
$$\begin{cases} H_0 : & \mu = 37 \\ H_A : & \mu < 37 \end{cases} \qquad \text{so,} \qquad \mu_0 = 37$$

2. The test statistic $T = \frac{(\overline{X}-\mu_0)\sqrt{n}}{s} = \frac{(34.3737-37)\sqrt{19}}{6.7608} = -1.6933$.

3. This a a **left tail** test. So, by (9.22), the p-value
   $\mathbf{p} = P(T < t) \approx tcdf(-5, t, n-1) \ = tcdf(-5, -1.6933, 18) = .0538$

4. Ten percent level of significance means $\alpha = .10$. Since,
   p-value $\mathbf{p} = .0538 < \alpha = .10$, we REJECT the null hypothesis at 10 percent level of significance.
   That means, at ten percent level of significance, we accept that the mean weight μ has reduced from 37 pounds.

5. Since p-value $\mathbf{p} = .0538$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 6 percent ($\alpha = .06$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.2.6. It is speculated that the teenage boys in a certain community are under weight. Under normal circumstances the mean weigh of this age group should be 155 pounds. A sample of 27 teenage boys had a mean weight 135 pound and sample standard deviation 32 pounds. Perform a significance test whether the mean weigh μ of this group is below 155 pounds, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic T.

3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis (or that the mean weight is lower or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.2.7. A guess is that the mean time μ needed for a student to arrive at the class from his/her residence would be less than 30 minutes. To test his guess, a sample 37 was collected. The sample mean time needed was 27 minutes and the sample standard deviation was 9 minutes. Perform a significance test whether the mean time μ needed would be below 30 minutes, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic T.
3. Compute the p-value.
4. At the 3 percent level of significance will you reject or accept the null hypothesis (or that the mean time is lower or not)?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

9.3 Significance Test, for Proportion p

This section is analogous to section 7.3, on confidence intervals for p. So, the significance tests we develop will be called One proportion Z-test, for p. As in the case of Z-tests (section 9.1.1), we would do three tests:

Two Tail Test	**Left Tail Test**	**Right Tail Test**
$\begin{cases} H_0: & p = p_0 \\ H_A: & p \neq p_0 \end{cases}$	$\begin{cases} H_0: & p = p_0 \\ H_A: & p < p_0 \end{cases}$	$\begin{cases} H_0: & p = p_0 \\ H_A: & p > p_0 \end{cases}$

(9.28)

We design the decision rules, as in the case Z-test (section 9.1.1). We only give the decision rules.

1. As in section 7.3, a sample of size n is drawn. Let X be the total number of success and $\overline{X} = \frac{X}{m}$ denote the sample proportion of success.
2. Assume $H_0 : p = p_0$ is true. From section 6.3, equation 6.4, we have

$$Z = \frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(\overline{X} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0,1) \quad \text{has standard normal distribution} \tag{9.29}$$

Expression Z above would be called a **test statistic**. We use lower case, to denote

$$z = \frac{(\overline{x} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \quad \text{the observed value of} \quad Z. \tag{9.30}$$

For three cases of alternate hypotheses, as in (9.28), we formulate the decision rules:

(a) **Two Tail Test:** When H_0 is true, by (9.29),

$$\begin{cases} P\left(|Z| \le z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha, & OR \\ P\left(|Z| > z_{\alpha/2}\right) = \alpha \end{cases} \tag{9.31}$$

Define p-value, for this test, as

$$\mathbf{p} = P\left(|Z| > |z|\right) = 1 - normalcdf\left(-|z|, |z|\right) \tag{9.32}$$

Because of (9.31), **at level of significance** α, set the decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \ |z| > z_{\alpha/2} \\ \textbf{Accept } H_0 & if \ \ |z| \le z_{\alpha/2} \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\, |z| > z_{\alpha/2} \\ \textbf{Reject } H_A & if \ \ |z| \le z_{\alpha/2} \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \ge \alpha \end{cases} \tag{9.33}$$

(b) **Left Tail Test:** When H_0 is true, by (9.29),

$$P\left(Z < -z_\alpha\right) = \alpha \tag{9.34}$$

Define p-value, for this test, as

$$\mathbf{p} = P(Z < z) = normalcdf(-5, z). \tag{9.35}$$

Because of (9.31), **at level of significance** α, set the decision rules:

$$\begin{cases} \textbf{Reject } H_0 & if\ z < -z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\ z < -z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \ge \alpha \end{cases} \tag{9.36}$$

(c) **Right Tail Test:** When H_0 is true, by (9.29),

$$P\left(z_\alpha < Z\right) = \alpha \tag{9.37}$$

Define p-value, for this test, as

$$\mathbf{p} = P(Z > z) = normalcdf(z, 5). \tag{9.38}$$

Because (9.31), **at level of significance** α, set the decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ z > z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if z > z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \ge \alpha \end{cases} \tag{9.39}$$

This is known as the **One proportion Z-Test**.

Universal Decision rule. As before (9.13, 9.27), the **p-value based decision rules** (9.33,9.36, 9.39) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \geq \alpha \end{cases} \tag{9.40}$$

However, p-values **p** are defined differently (9.32, 9.35, 9.38), are specific to the respective tests.

9.3.1 Problems: One proportion Z-Test

Exercise 9.3.1. In a sample of 197 apples from a lot, 26 were found to be sour. The lot will be rejected if more than 10 percent is sour. Perform a significance test for the acceptability of this lot.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic Z.
3. Compute the p-value.
4. At the 3 percent level of significance will you reject or accept the null hypothesis (or whether the lot is acceptable or not)?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The proportion of sour apples will be denoted by p.
A one proportion Proportion Z-Test will be performed.
Data is summarized as follows:
The sample size $n = 197$,
The number of success $X = 26$,
The sample proportion of success $\overline{X} = \frac{X}{n} = \frac{26}{197} = .1320$

1. The alternate hypothesis is that p is more that .1 (ten percent).: $H_A : p > .1$
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & p = .1 \\ H_A : & p > .1 \end{cases} \qquad \text{so,} \qquad p_0 = .1$$

2. The test statistic $z = \frac{(\overline{x}-p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} = \frac{(.1320-.1)\sqrt{197}}{\sqrt{.1(1-.1)}} = 1.4971$

3. This a a **right tail** test. So, by (9.38), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5) = normalcdf(1.4971, 5) = .0672$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0672 \not< \alpha = .03$, we accept the null hypothesis at 3 percent level of significance.
   That means, at three percent level of significance, we conclude that the lot is acceptable.

5. Since p-value $\mathbf{p} = .0672$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 7 percent ($\alpha = .07$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.3.2. A new vaccine was tried on 147 randomly selected individuals, and it was determined that 61 of them got the virus. It is known that usually fifty percent of the population get the virus. Perform a significance test to decide if this vaccine is indeed effective or not.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis (or that the vaccine is effective or not)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The proportion of the vaccinated population who benefit from it would be denoted by p.
A one proportion Proportion Z-Test will be performed.
Data is summarized as follows:
The sample size $n = 147$,
The number of success $X = 61$,
The sample proportion of success $\overline{X} = \frac{X}{n} = \frac{61}{147} = .4150$

1. The alternate hypothesis is that p is less that .5 (better than 50 percent).: $H_A : p < .5$
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & p = .5 \\ H_A : & p < .5 \end{cases} \qquad \text{so,} \qquad p_0 = .5$$

2. The test statistic $z = \frac{(\overline{x}-p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} = \frac{(.4150-.5)\sqrt{147}}{\sqrt{.5(1-.5)}} = -2.0611$

3. This a a **left tail** test. So, by (9.35), the p-value
   $\mathbf{p} = P(Z < z) = normalcdf(-5, z) = normalcdf(-5, -2.0611) = .0196$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0196 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
   That means, at three percent level of significance, we conclude that the vaccine in EFFECTIVE.

5. Since p-value $\mathbf{p} = .0196$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,6, 7, 8, 9,10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.3.3. Before an election for a congressional seat, a poll was conducted. Out of 887 randomly selected voters interviewed, 389 said that they would vote for Candidate A. The election strategists have decided that, to win Candidate A needs to get more than 40 percent votes. Perform a significance test whether he/she will get more than 40 percent or not.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis (or whether he/she will get more than 40 percent or not.)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The proportion of the voter population who would vote for the candidate would be denoted by p.
A one proportion Proportion Z-Test will be performed.
Data is summarized as follows:
The sample size $n = 887$,
The number of success $X = 389$,
The sample proportion of success $\overline{X} = \frac{X}{n} = \frac{389}{887} = .4386$

1. The alternate hypothesis is that p is more that .4 (higher than 40 percent).: $H_A : p > .4$.
   The null and alternative hypotheses are:
   $$\begin{cases} H_0 : & p = .4 \\ H_A : & p > .4 \end{cases} \qquad \text{so,} \qquad p_0 = .4$$

2. The test statistic $z = \frac{(\overline{x} - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} = \frac{(.4386-.4)\sqrt{887}}{\sqrt{.4(1-.4)}} = 2.3466$

3. This a a **right tail** test. So, by (9.38), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5) = normalcdf(2.3466, 5) = .0095$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0095 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
   That means, at three percent level of significance, we conclude that the Candidate A will win.

5. Since p-value $\mathbf{p} = .0095$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 1 percent ($\alpha = .01$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.3.4. A pollster was asked to make decision whether the proportion p of the US population who would support Government shutdown due to budget dispute, would be above 55 percent or not? A sample 898 were polled and 522 of them said they would support government shutdown. Perform a significance test whether p would be above 55 percent?

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.3.5. A telephone company wants to know whether the proportion p of calls that are longer than 20 minutes, in a town, would exceed 65 percent. A sample of 1123 class, 761 were longer than 20 minute. Perform a significance test whether p would be above 65 percent?

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.3.6. It is believed that, this year, proportion p of infected oranges will remain below 15 percent. A sample of 1333 class, 175 were were infected. Perform a significance test whether p would remain below 15 percent?

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic Z.

3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

9.4 Testing Hypotheses on Variance σ^2

Let $X \sim N(\mu, \sigma)$ be a normal random variable.

This section is analogous to section 7.4.2, on χ^2-interval for σ^2. So, the significance tests we develop will be called χ^2-Test. We would do two tests:

Left Tail Test	**Right Tail Test**
$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_A: & \sigma^2 < \sigma_0^2 \end{cases}$	$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_A: & \sigma^2 > \sigma_0^2 \end{cases}$

(9.41)

We only give the decision rules.

1. Similar to section 7.4.2, we assume that $X \sim N(\mu, \sigma)$
2. Draw a sample $X_1, X_2, \ldots, X_n$ from the X-population, of size n. As always,

$$\begin{cases} \overline{X} = \frac{X_1+X_2+\cdots+X_n}{n} & = \text{the sample mean.} \\ S^2 = \frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n-1} & = \text{the sample variance.} \end{cases} \tag{9.42}$$

3. Assume $H_0: \sigma^2 = \sigma_0^2$ is true. Then

$$Y = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \quad \text{has } \chi^2 \text{ distribution, with degrees of freedom } n-1 \tag{9.43}$$

Expression Y above would be called a **test statistic**. We use lower case, to denote

$$y = \frac{(n-1)s^2}{\sigma_0^2} \quad \text{the observed value of} \quad Y. \tag{9.44}$$

For two cases of alternate hypotheses, as in (9.41), we formulate the decision rules:

(a) **Left Tail Test:** When H_0 is true, by (9.43),

$$P\left(Y < \chi^2_{n-1,1-\alpha}\right) = \alpha \tag{9.45}$$

We define p-value, for this test, as

$$\mathbf{p} = P(Y < y) \ = \chi^2 cdf(0, y, n-1). \tag{9.46}$$

Because of (9.45), **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if\ y < \chi^2_{n-1,1-\alpha} \\ \textbf{Accept } H_0 & Otherwise \end{cases} \quad \text{Equivalently,} \quad \begin{cases} \textbf{Accept } H_A & if\ y < \chi^2_{n-1,1-\alpha} \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.47}$$

(b) **Right Tail Test:** When H_0 is true, by (9.43),

$$P\left(\chi^2_{n-1,\alpha} < Y\right) = \alpha \tag{9.48}$$

Define p-value, for this test, as

$$\mathbf{p} = P(y < Y) \ = 1 - \chi^2 cdf(0, y, n-1). \tag{9.49}$$

Because of (9.24), **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if\ y > \chi^2_{n-1,\alpha} \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\ y > \chi^2_{n-1,\alpha} \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.50}$$

This is known as the χ^2**-Test**.

Universal Decision rule. As before (9.13, 9.27, 9.40), the ***p*-value based decision rules** (9.47, 9.50) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.51}$$

However, p-values $\mathbf{p}$ are defined differently (9.46, 9.49), are specific to the respective tests.

9.4.1 Problems: on χ^2-Test

Exercise 9.4.1. Suppose that we have collected a sample of size $n = 23$ from a normal population $X \sim N(\mu, \sigma)$. The sample variance was found to be $S^2 = 46.7$. It is believed that the variance σ^2 is higher than 25. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance will you reject or accept the null hypothesis?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. A χ^2-Test will be performed.
Given data is summarized: Sample size $n = 23$
Sample variance $S^2 = 46.7$
Degrees of freedom $df = n - 1 = 23 - 1 = 22$.

1. We are testing the belief, σ^2 is higher than 25.
   The null and alternative hypotheses are:
   $$\begin{cases} H_0: & \sigma^2 = 25 \\ H_A: & \sigma^2 > 25 \end{cases} \qquad \text{so,} \qquad \sigma_0^2 = 25$$

2. The test statistic $Y = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(23-1)46.7}{25} = 41.096$.

3. This is a **right tail** test. So, by (9.49), the p-value
   $\mathbf{p} = P(y < Y) \ = 1 - \chi^2 cdf(0, y, n-1)$
   $= 1 - \chi^2 cdf(0, 41.096, 22) = 1 - .9920 = .008$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .008 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
   That means, we conclude, at 3 percent level of confidence, that the variance σ^2 is higher than 25.

5. Since p-value $\mathbf{p} = .008$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 1 percent ($\alpha = .01$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.4.2. Following is data on the life expectancies of a group of people older than 75.

87 92 81 76 81 87 79 88 88 79 81 89 97 91 82

It is believed that the variance σ^2 life expectancy of this group is higher than 16. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 1 percent level of significance will you reject or accept the null hypothesis?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. A χ^2-Test will be performed.
As in chapter 2, given data is processed:
Sample size $n = 15$
Sample variance $S^2 = (5.8578)^2 = 34.3138$
Degrees of freedom $df = n - 1 = 15 - 1 = 14$.

1. We are testing the belief, that the variance σ^2 is higher than 16.
   The null and alternative hypotheses are:
$$\begin{cases} H_0: & \sigma^2 = 16 \\ H_A: & \sigma^2 > 16 \end{cases} \qquad \text{so,} \qquad \sigma_0^2 = 16$$

2. The test statistic $Y = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(15-1)34.3138}{16} = 30.0246$.

3. This is a **right tail** test. So, by (9.49), the p-value
   $\mathbf{p} = P(y < Y) \ = 1 - \chi^2 cdf(0, y, n-1)$
   $= 1 - \chi^2 cdf(0, 30.0246, 14) = 1 - .9924 = .0076$

4. One percent level of significance means $\alpha = .01$. Since,
   p-value $\mathbf{p} = .0076 < \alpha = .01$, we REJECT the null hypothesis at 1 percent level of significance.
   That means, we conclude, at 1 percent level of confidence, that the variance σ^2 is higher than 16.

5. Since p-value $\mathbf{p} = .0076$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7, 8, 9,10 percent; 1 percent ($\alpha = .01$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.4.3. The following is data on monthly gas consumption (in ccf) by the households in a town during the winter months.

154 222 264 257 127 228 240 393 278 140

It is believed that the variance σ^2 is less than 15600 ccf^2 . Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 5 percent level of significance will you reject or accept the null hypothesis?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. A χ^2-Test will be performed.
As in chapter 2, given data is processed:
Sample size $n = 10$
Sample variance $S^2 = (78.4645)^2 = 6156.6778$
Degrees of freedom $df = n - 1 = 10 - 1 = 9$.

1. We are testing the belief, σ^2 is less than 15600.
   The null and alternative hypotheses are:

$$\begin{cases} H_0: & \sigma^2 = 15600 \\ H_A: & \sigma^2 < 15600 \end{cases} \qquad \text{so,} \qquad \sigma_0^2 = 15600$$

2. The test statistic $Y = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(10-1)6156.6778}{15600} = 3.5519$.

3. This is a **left tail** test. So, by (9.46), the p-value
   $\mathbf{p} = P(Y < y) = \chi^2 cdf(0, y, n-1) = \chi^2 cdf(0, 3.5519, 9) = .0616$

4. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .0616 \not< \alpha = .05$, we ACCEPT the null hypothesis at 5 percent level of significance.
   That means, at the level of significance 5 percent, we do NOT accept that the σ^2 is lower than 15600

5. Since p-value $\mathbf{p} = .0616$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 7 percent ($\alpha = .07$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.4.4. The birth weight of babies has a normal distribution, with variance σ^2. Because of the economic and social diversity of the community, there are concerns about variability of the birth weight. It is believed that the variance σ^2 may be higher than 17 pounds-square. A sample of 26 birth-weight was collected and the sample variance was found to be $S^2 = 26.7$ pounds-square. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance will you reject or accept the null hypothesis?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. A χ^2-Test will be performed.
Given data is summarized: Sample size $n = 26$
Sample variance $S^2 = 26.7$
Degrees of freedom $df = n - 1 = 26 - 1 = 25$.

1. We are testing the belief, σ^2 is higher than 17.
   The null and alternative hypotheses are:
   $$\begin{cases} H_0: & \sigma^2 = 17 \\ H_A: & \sigma^2 > 17 \end{cases} \qquad \text{so,} \qquad \sigma_0^2 = 17$$
2. The test statistic $Y = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(26-1)26.7}{17} = 39.2647$.
3. This is a **right tail** test. So, by (9.49), the p-value
   $\mathbf{p} = P(y < Y) = 1 - \chi^2 cdf(0, y, n-1) =$
   $1 - \chi^2(0, 39.2647, 25) = .0346$
4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0346 \not< \alpha = .03$, we ACCEPT the null hypothesis at 3 percent level of significance..
   That means, , at 3 percent level of confidence, we do not accept that the variance σ^2 is higher than 17.

5. Since p-value $\mathbf{p} = .0346$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 4 percent ($\alpha = .04$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.4.5. It is speculated that the variability of length of babies, in a community, at birth may be small. It is speculated that variance length σ^2 of babies may be lower than 25 square-inches. A sample of size $n = 16$ on birth-lengths was collected. The sample variance was found to be $S^2 = 13$ square-inches. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 5 percent level of significance will you reject or accept the null hypothesis?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. A χ^2-Test will be performed.
Given data is summarized:
Sample size $n = 16$
Sample variance $S^2 = 13$
Degrees of freedom $df = n - 1 = 16 - 1 = 15$.

1. We are testing the speculation, σ^2 is less than 25.
   The null and alternative hypotheses are:
$$\begin{cases} H_0: & \sigma^2 = 25 \\ H_A: & \sigma^2 < 25 \end{cases} \qquad \text{so,} \qquad \sigma_0^2 = 25$$
2. The test statistic $Y = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(16-1)13}{25} = 7.8$.
3. This is a **left tail** test. So, by (9.46), the p-value
   $\mathbf{p} = P(Y < y) = \chi^2 cdf(0, y, n-1) = \chi^2 cdf(0, 7.8, 15) = .0684$
4. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .0684 \not< \alpha = .05$, we ACCEPT the null hypothesis at 5 percent level of

significance.
That means, at the level of significance 5 percent, we do NOT accept that the σ^2 is less than 25.

5. Since p-value $\mathbf{p} = .0684$, from this possibilities of .1, .5, 1,2, 3, 4, 5,4, 7,8, 9,10 percent; 7 percent ($\alpha = .07$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.4.6. The variability of the length of the telephone calls is a concern of the telephone company. It is speculated that the variance σ^2 may be higher that 64 minutes-square. A sample of 14 calls had a sample variance $S^2 = 99$ minutes-square. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 5 percent level of significance will you reject or accept the null hypothesis?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. A χ^2-Test will be performed.
Given data is summarized: Sample size $n = 14$
Sample variance $S^2 = 99$
Degrees of freedom $df = n - 1 = 14 - 1 = 13$.

1. We are testing the speculation , σ^2 is higher than 64.
   The null and alternative hypotheses are:
$$\begin{cases} H_0: & \sigma^2 = 64 \\ H_A: & \sigma^2 > 64 \end{cases} \qquad \text{so,} \qquad \sigma_0^2 = 64$$
2. The test statistic $Y = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(14-1)99}{64} = 20.1094$.
3. This is a **right tail** test. So, by (9.49), the p-value
   $\mathbf{p} = P(y < Y) = 1 - \chi^2 cdf(0, y, n-1)$
   $= 1 - \chi^2 cdf(0, 20.1094, 13) = .0925$

4. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .0925 \not< \alpha = .05$, we ACCEPT the null hypothesis at 5 percent level of significance.
   That means, we conclude, at 5 percent level of confidence, that the variance σ^2 is not higher than 64.

5. Since p-value $\mathbf{p} = .0925$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 10 percent ($\alpha = .10$) would be the lowest level at which we would reject the null hypothesis.

9.5 Two Populations: Known σ_1, σ_2

This section is analogous to section 8.1, on confidence intervals for $\mu_1 - \mu_2$. As in section 8.1, suppose X, Y are two similar random variables. Let mean and st. deviations of X and Y be denoted, as follows:

	Population X	**Population Y**
mean	$E(X) = \mu_1$	$E(Y) = \mu_2$
St. Dev.	σ_1	σ_2

We test equality $\mu_1 = \mu_2$, of means. We deal with the case when σ_1 and σ_2 are known. We would do three tests:

Two Tail Test	**Left Tail Test**	**Right Tail Test**
$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 \neq 0 \end{cases}$	$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 < 0 \end{cases}$	$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 > 0 \end{cases}$

(9.52)

1. A sample $X_1, X_2, \ldots, X_m$, of size m, is drawn from the X-population and a sample $Y_1, Y_2, \ldots, Y_n$, of size n, is drawn from the Ypopulation. Let

$$\begin{cases} \overline{X} = \frac{X_1 + X_2 + \cdots + X_m}{m} \\ \overline{Y} = \frac{Y_1 + Y_2 + \cdots + Y_m}{n} \end{cases} \quad \text{be the sample means}$$

   Assume that the X-samples and Y-samples are drawn independently. As was stated in section 8.1, approximately,

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \sigma) \quad \text{where} \quad \begin{cases} E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \\ \sigma = \sqrt{\frac{\sigma_1^2}{\mathbf{m}} + \frac{\sigma_2^2}{\mathbf{n}}} \end{cases}$$

   If X and Y are normal, then this is exact (not just approximate) distribution.

2. Assume $H_0 : \mu_1 - \mu_2 = 0$ is true. Then,

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1) \quad \text{has standard normal distribution} \tag{9.53}$$

Expression Z above would be called a **test statistic**. We use lower case, to denote

$$z = \frac{\overline{x} - \overline{y}}{\sigma} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \qquad \text{the observed value of} \quad Z. \tag{9.54}$$

For three cases of alternate hypotheses, as in (9.52), we formulate the decision rules:

(a) **Two Tail Test:** When H_0 is true, by (9.53),

$$\begin{cases} P\left(|Z| \le z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha, & OR \\ P\left(|Z| > z_{\alpha/2}\right) = \alpha \end{cases} \tag{9.55}$$

Define p-value, for this test, as

$$\mathbf{p} = P\left(|Z| > |z|\right) = 1 - normalcdf\left(-|z|, |z|\right) \tag{9.56}$$

Because of (9.55), **at level of significance** α, set the decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \ |z| > z_{\alpha/2} \\ \textbf{Accept } H_0 & if \ \ |z| \le z_{\alpha/2} \end{cases} \text{Equivalently,} \begin{cases} \textbf{Accept } H_A & if \ |z| > z_{\alpha/2} \\ \textbf{Reject } H_A & if \ \ |z| \le z_{\alpha/2} \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \ge \alpha \end{cases} \tag{9.57}$$

(b) **Left Tail Test:** When H_0 is true, by (9.53),

$$P\left(Z < -z_\alpha\right) = \alpha \tag{9.58}$$

Define p-value, for this test, as

$$\mathbf{p} = P(Z < z) = normalcdf(-5, z). \tag{9.59}$$

Because of (9.58), **at level of significance** α, set the decision rules:

$$\begin{cases} \textbf{Reject } H_0 & if \ z < -z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{Equivalently,} \begin{cases} \textbf{Accept } H_A & if \ z < -z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \ge \alpha \end{cases} \tag{9.60}$$

(c) **Right Tail Test:** When H_0 is true, by (9.53),

$$P\left(z_\alpha < Z\right) = \alpha \tag{9.61}$$

Define p-value, for this test, as

$$\mathbf{p} = P(Z > z) = normalcdf(z, 5). \tag{9.62}$$

Because of (9.61), **at level of significance** α, set the decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ z > z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \quad \text{Equivalently,} \begin{cases} \textbf{Accept } H_A & if z > z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.63}$$

These are called Two sample Z-Test.

Universal Decision rule. As before (9.13, 9.27, 9.40, 9.51), the **p-value based decision rules** (9.57,9.60, 9.63) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.64}$$

However, p-values $\mathbf{p}$ are defined differently (9.56, 9.59, 9.62), are specific to the respective tests.

9.5.1 Problems: Two sample Z-Test

Exercise 9.5.1. The equality of means μ_1, μ_2 of two populations is to be compared. The standard deviations σ_1, σ_2, respectively, are known to be $\sigma_1 = 8.1$ and $\sigma_2 = 11.3$. A sample of size $m = 64$ was collected from the first population, and the sample mean was found to be $\overline{X} = 3.5$. A sample of size $n = 99$ was collected from the second population, and the sample mean was found to be $\overline{Y} = 7.9$.

It is speculated that these two means are unequal: $\mu_1 \neq \mu_2$? Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.

4. At the 3 percent level of significance will you reject or accept the null hypothesis?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 8.1$	$\sigma_2 = 11.3$
samplemean	$\overline{X} = 3.7$	$\overline{Y} = 7.9$
sample size	$m = 64$	$n = 99$

1. We are testing the speculation, $\mu_1 \neq \mu_2$.
   The null and alternative hypotheses are:
$$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 \neq 0 \end{cases}$$

2. The test statistic
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{3.7 - 7.9}{\sqrt{\frac{8.1^2}{64} + \frac{11.3^2}{99}}} = -2.8919$$

3. This is a **Two tail** test. So, by (9.56), the p-value
   $\mathbf{p} = P(|Z| > |z|) = 1 - normalcdf(-|z|, |z|)$
   $= 1 - normalcdf(-2.8919, 2.8919) = 1 - .9962 = .0038$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0038 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.

5. Since p-value $\mathbf{p} = .0038$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; .5 percent ($\alpha = .005$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.5.2. The birth weight of babies in two hospitals would have to be compared. The birth-weight distributions X and Y of these two hospitals are normal with means μ_1, μ_2 and standard deviations σ_1, σ_2, respectively. It is known the standard deviations $\sigma_1 = 2.3$ pounds and $\sigma_2 = 2.9$ pounds. A sample of size $m = 35$ babies from the first hospitals was

collected, and the sample mean birth weight was found to be $\overline{X} = 8.9$ pounds. A sample of size $n = 48$ babies from the second hospital was collected, and the sample mean birth weight was found to be $\overline{Y} = 7.6$ pounds.

Due to economic disparities between these two neighborhoods, it is speculated that the mean μ_1 is higher than the mean μ_2. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance would you accept this speculation that μ_1 is higher than μ_2?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 2.3$	$\sigma_2 = 2.9$
samplemean	$\overline{X} = 8.9$	$\overline{Y} = 7.6$
sample size	$m = 35$	$n = 48$

1. The alternative hypothesis is that the X-mean μ_1 would be higher than the Y-mean μ_2: $H_A \mu_1 > \mu_2$.
   The null and alternative hypotheses are:
$$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 > 0 \end{cases}$$
2. The test statistic
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{8.9 - 7.6}{\sqrt{\frac{2.3^2}{35} + \frac{2.9^2}{48}}} = 2.2756$$
3. This is a **right tail** test. So, by (9.62), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5)$
   $= P(2.2756 < Z) = normalcdf(2.2756, 5)) = .0114$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0114 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.

5. Since p-value $\mathbf{p} = .0114$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.5.3. Elephants in different parts of the world are different in height, weight, and length of ear and tusk. The mean heights μ_1 and μ_2 of elephants in two different regions would be compared. It is assumed that the height distributions X and Y of the elephants in these two regions are normally distributed. The standard deviation of X and Y are $\sigma_1 = 1.5$ feet and $\sigma_2 = 1.3$ feet, respectively. A sample of size 25 was collected from region-I, and the sample mean height was found to be $\overline{X} = 9.9$ feet. A sample of size 28 was collected from the region-II was collected, and the sample mean height was found to be $\overline{Y} = 9.1$ feet. It is believed that mean height μ_1 is higher than μ_2. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance will you accept μ_1 is higher than μ_2?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 1.5$	$\sigma_2 = 1.3$
samplemean	$\overline{X} = 9.9$	$\overline{Y} = 9.1$
sample size	$m = 25$	$n = 28$

1. We are testing the speculation, μ_1 is higher than μ_2.
   The null and alternative hypotheses are:

$$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 > 0 \end{cases}$$

2. The test statistic
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{9.9 - 9.1}{\sqrt{\frac{1.5^2}{25} + \frac{1.3^2}{28}}} = 2.0631$$

3. This is a **right tail** test. So, by (9.62), the p-value
$\mathbf{p} = P(Z > z) = normalcdf(z, 5)$
$= normalcdf(2.0631, 5)) = .0196$

4. Three percent level of significance means $\alpha = .03$. Since,
p-value $\mathbf{p} = .0196 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance; or we accept the alternate that μ_1 is higher than μ_2.

5. Since p-value $\mathbf{p} = .0038$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.5.4. It is speculated that the mean weight μ_1 of King salmon in Kenai is lower than the mean weight μ_2 of King salmon in Anchor River. The standard deviation weight of the Kings in Kenai is $\sigma_1 = 7.7$ pounds. The standard deviation weight of the Kings in Anchor is $\sigma_2 = 9.1$ pounds. A sample of 51 King from Kenai had a mean $\overline{X} = 30.5$ pounds. A sample of 63 King from Anchor had a mean $\overline{Y} = 33$ pounds. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 5 percent level of significance, do you accept that μ_1 is lower than μ_2?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. The given data is summarized as follows:

	Population I X	**Population II** Y
St. Dev.	$\sigma_1 = 7.7$	$\sigma_2 = 9.1$
samplemean	$\overline{X} = 30.5$	$\overline{Y} = 33$
sample size	$m = 51$	$n = 63$

1. We are testing the speculation, μ_1 is lower than μ_2.
   The null and alternative hypotheses are:
$$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 < 0 \end{cases}$$

2. The test statistic
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{30.5 - 33}{\sqrt{\frac{7.7^2}{51} + \frac{9.1^2}{63}}} = -1.5885$$

3. This is a **left tail** test. So, by (9.59), the p-value
   $\mathbf{p} = P(Z < z) = normalcdf(-5, z)$
   $= normalcdf(-5, -1.5885)) = .0561$

4. Five percent level of significance means $\alpha = .05$. Since,
   p-value $\mathbf{p} = .0561 \not< \alpha = .05$, we ACCEPT the null hypothesis at 5 percent level of significance.
   That means we do not accept that the mean weight of Kings in Kenai is less than that of Anchor, at 3 percent level of significance.

5. Since p-value $\mathbf{p} = .0561$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 6 percent ($\alpha = .06$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.5.5. There is a speculation circulating that the mean percent scores μ_1 in fall semester grades is higher than the mean percent scores μ_2 in spring semester grades. The standard deviation of fall percent scores is $\sigma_1 = 27$ percent and the standard deviation of spring percent scores is $\sigma_2 = 23$ percent. A sample of 87 students in fall had a sample mean score $\overline{X} = 76$ percent. A sample of 77 students in spring had a sample mean score $\overline{Y} = 69$ percent. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test. (This would be a Right Tail.)

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance will you accept μ_1 is higher than μ_2?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.5.6. it is speculated that the mean annual salary μ_1 of the professors in a State University (I) is higher than the mean annual salary μ_2 of the professors in the State University (II). The standard deviation of the annual salary in the University -I $\sigma_1 = \$16,000$ and the standard deviation of the annual salary in the University-II $\sigma_2 = \$11,500$. A sample of 47 professors in University-I had a mean salary $\overline{X} = \$77,000$. A sample of 58 professors in University-II had a mean salary $\overline{Y} = \$71,500$ Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test. (This would be a Right Tail.)
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance will you accept μ_1 is higher than μ_2?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

9.6 Two sample T-Test: Unknown σ_1, σ_2

This section is analogous to section 8.2, on Two sample T-intervals for $\mu_1 - \mu_2$. We develop a test for $H_0 : \mu_1 - \mu_0 = 0$. As in section 8.2, we proceed as follows:

1. As indicated above, $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ have normal distributions, and assume
$$\sigma_1 = \sigma_2 = \sigma$$
2. A sample $X_1, X_2, \ldots, X_m$, of size m, is drawn from the X-population and a sample $Y_1, Y_2, \ldots, Y_n$, of size n, is drawn from the Y population. Use the notations for sample mean and sample standard deviations (variance), as follows:
$$\begin{cases} \overline{X} = \frac{X_1+X_2+\cdots+X_m}{m} \\ \overline{Y} = \frac{Y_1+Y_2+\cdots+Y_m}{n} \end{cases} \qquad \begin{cases} S_1^2 = S_X^2 = \frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_m-\overline{X})^2}{m-1} \\ S_2^2 = S_Y^2 = \frac{(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2+\cdots+(Y_n-\overline{Y})^2}{n-1} \end{cases}$$
We obtain the **pooled estimate**, S_p^2 for σ^2, as follows:
$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

This happens to be the weighted mean of S_X^2 with weight $(m-1)$, and S_Y^2 with weight $(n-1)$. So, **pooled estimate** S_p for σ is

$$S_p = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}$$

We test equality $\mu_1 = \mu_2$, of means. We deal with the case when σ_1 and σ_2 are unknown. We would do three tests:

Two Tail Test	**Left Tail Test**	**Right Tail Test**
$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 \neq 0 \end{cases}$	$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 < 0 \end{cases}$	$\begin{cases} H_0: & \mu_1 - \mu_2 = 0 \\ H_A: & \mu_1 - \mu_2 > 0 \end{cases}$

(9.65)

1. Assume $H_0 : \mu_1 - \mu_2 = 0$ is true. Then,

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2} \qquad \text{has } t\text{−distribution,} \tag{9.66}$$

with degrees of freedom $df = m + n - 2$. Expression T above would be called a **test statistic**. We use lower case, to denote

$$t = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad \text{the observed value of} \quad T. \tag{9.67}$$

As in We formulate the decision rules, analogous to T-test (section 9.2), for three significance test, as in (9.65):

(a) **Two Tail Test:** When H_0 is true, by (9.66),

$$\begin{cases} P\left(|T| \leq t_{n-1,\alpha/2}\right) = P\left(-t_{m+n-2,\alpha/2} \leq T \leq t_{m+n-2,\alpha/2}\right) = 1-\alpha, & OR \\ P\left(|T| > t_{m+n-2,\alpha/2}\right) = \alpha \end{cases} \tag{9.68}$$

We define p-value, for this test, as

$$\mathbf{p} = P\left(|T| > |t|\right) = 1 - tcdf\left(-|t|, |t|, m+n-2\right) \tag{9.69}$$

Because of (9.68), **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if \ |t| > t_{m+n-2,\alpha/2} \\ \textbf{Accept } H_0 & if \ |t| \leq t_{m+n-2,\alpha/2} \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if \ |t| > t_{m+n-2,\alpha/2} \\ \textbf{Reject } H_A & if \ |t| \leq t_{m+n-2,\alpha/2} \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \geq \alpha \end{cases} \tag{9.70}$$

(b) **Left Tail Test:** When H_0 is true, by (9.66),

$$P\left(T < -t_{m+n-2,\alpha}\right) = \alpha \tag{9.71}$$

We define p-value, for this test, as

$$\mathbf{p} = P(T < t) \approx tcdf(-5, t, m+n-2). \tag{9.72}$$

Because of (9.71), **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if\ t < -t_{m+n-2,\alpha} \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\ t < -t_{m+n-2,\alpha} \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.73}$$

(c) **Right Tail Test:** When H_0 is true, by (9.66),

$$P\left(t_{m+n-2,\alpha} < T\right) = \alpha \tag{9.74}$$

We define p-value, for this test, as

$$\mathbf{p} = P(T > t) \approx tcdf(t, 5, m+n-2)\ . \tag{9.75}$$

Because of (9.74), **at level of significance** α, set the decision rule as:

$$\begin{cases} \textbf{Reject } H_0 & if\ t > t_{m+n-2,\alpha} \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\ t > t_{m+n-2,\alpha} \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.76}$$

These are called Two sample T-Test.

Universal Decision rule. As before (9.13, 9.27, 9.40, 9.51, 9.64), the **p-value based decision rules** (9.70,9.73, 9.76) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.77}$$

However, p-values $\mathbf{p}$ are defined differently (9.69, 9.72, 9.75), are specific to the respective tests.

9.6.1 Problems: Two sample T-Test

Exercise 9.6.1. Suppose that two "similar" normal populations have means μ_1, μ_2 respectively and same standard deviations σ. It is believed that μ_1 and μ_2 are not equal. A sample of size $m = 11$ from the first population the sample mean was found to be $\overline{X} = 13.5$ and the sample standard deviation $S_1 = 2.33$. A sample of size $n = 13$ was collected from the second population that had a sample mean $\overline{Y} = 11.5$ and sample standard deviation $S_2 = 2.73$. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance would you accept this speculation that μ_1 is not equal to μ_2?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. Since σ_1, σ_2 are not known, we perform a two sample T-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 11$	$n = 13$
samplemean	$\overline{X} = 13.5$	$\overline{Y} = 11.5$
Sample St. Dev.	$S_1 = 2.33$	$S_2 = 2.73$
Degrees of freedom $= m + n - 2 = 11 + 13 - 2 = 22$		

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}} = \sqrt{\frac{(11-1)2.33^2 + (13-1)2.73^2}{11+13-2}} = 2.5560$$

1. The alternative hypothesis is that the μ_1 is not equal to μ_2: $H_A : \mu_1 \neq \mu_2$.
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu_1 - \mu_2 = 0 \\ H_A : & \mu_1 - \mu_2 \neq 0 \end{cases}$$

2. The test statistic
$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{13.5 - 11.5}{2.5560\sqrt{\frac{1}{11} + \frac{1}{13}}} = 1.9100$$

3. This is a **two tail** test. So, by (9.69), the p-value
$\mathbf{p} = P(|T| > |t|) = 1 - tcdf(-|t|, |t|, m+n-2)$
$= 1 - tcdf(-1.9100, 1.9100, 22) = 1 - .9307 = .0693$

4. Three percent level of significance means $\alpha = .03$. Since,
p-value $\mathbf{p} = .0693 \not< \alpha = .03$, we ACCEPT the null hypothesis at 3 percent level of significance.

5. Since p-value $\mathbf{p} = .0693$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 7 percent ($\alpha = .07$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.6.2. The means μ_1, μ_2 of two normal random variables X and Y would have to be compared. They have equal standard deviation σ. it is believed that μ_1 is lower than μ_2. A sample of size $m = 64$ was collected from the X-population and the sample mean and standard deviation were found to be $\overline{X} = 1.8$, $S_1 = 9.2$. A sample of size $n = 99$ was collected from the Y-population and the sample mean and standard deviation were $\overline{Y} = 4.4$, $S_2 = 8.7$. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you accept that μ_1 is lower than μ_2?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. Since σ_1, σ_2 are not known, we perform a two sample T-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 64$	$n = 99$
samplemean	$\overline{X} = 1.8$	$\overline{Y} = 4.4$
Sample St. Dev.	$S_1 = 9.2$	$S_2 = 8.7$
Degrees of freedom $= m + n - 2 = 64 + 99 - 2 = 161$		

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}} = \sqrt{\frac{(64-1)9.2^2 + (99-1)8.7^2}{64+99-2}} = 8.8990$$

1. The alternative hypothesis is that μ_1 is lower than μ_2: $H_A : \mu_1 < \mu_2$.
   The null and alternative hypotheses are:
   $$\begin{cases} H_0 : & \mu_1 - \mu_2 = 0 \\ H_A : & \mu_1 - \mu_2 < 0 \end{cases}$$

2. The test statistic
   $$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{1.8 - 4.1}{8.8990\sqrt{\frac{1}{64} + \frac{1}{99}}} = -1.6114$$

3. This is a **left tail** test. So, by (9.72), the p-value
   $\mathbf{p} = P(T < t) \approx tcdf(-5, t, m+n-2)$
   $= tcdf(-5, -1.6114, 161) = .0545$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0545 \not< \alpha = .03$, we ACCEPT the null hypothesis at 3 percent level of significance.

5. Since p-value $\mathbf{p} = .0545$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 6 percent ($\alpha = .06$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.6.3. The difference in mean monthly water consumption in two adjacent towns has to be compared. It is speculated that the mean monthly consumption μ_1 in Town-I is lower than the mean monthly consumption μ_2 in Town-II. A sample 37 household in the Town-I had a sample mean 6500 gallons and standard deviation 450 gallons. A sample 49 household in the Town-II had a sample mean 6800 gallons and standard deviation 650 gallons. Assume that the standard deviations are equal. Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you accept would you accept that μ_1 is lower than μ_2?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. Since σ_1, σ_2 are not known, we perform a two sample T-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 37$	$n = 49$
samplemean	$\overline{X} = 6500$	$\overline{Y} = 6800$
Sample St. Dev.	$S_1 = 450$	$S_2 = 650$
Degrees of freedom $= m + n - 2 = 37 + 49 - 2 = 84$		

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}} = \sqrt{\frac{(37-1)450^2 + (49-1)650^2}{11+13-2}} = 572.8990$$

1. The alternative hypothesis is that the μ_1 is lower than μ_2: $H_A : \mu_1 < \mu_2$.
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu_1 - \mu_2 = 0 \\ H_A : & \mu_1 - \mu_2 < 0 \end{cases}$$

2. The test statistic

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{1.8 - 4.1}{572.8990\sqrt{\frac{1}{37} + \frac{1}{49}}} = -2.4043$$

3. This is a **left tail** test. So, by (9.72), the p-value
   $\mathbf{p} = P(T < t) \approx tcdf(-5, t, m+n-2)$
   $= tcdf(-5, -2.4043, 84) = .0092$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = .0092 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
   That means we ACCEPT that the mean monthly consumption in Town-I is lower than that of town-II.

5. Since p-value $\mathbf{p} = .0092$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 1 percent ($\alpha = .01$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.6.4. The birth weight of the babies in developed and developing countries are normally distributed with mean μ_1, μ_2 and equal standard deviation σ. (My data is not real.) It is believed that μ_1 is higher than μ_2. The following data about the birth weight from developed and developing nations were collected.

Developed				
8.8	8.1	6.3	9.7	6.3
7.1	5.3	7.7	9.1	8.1
8.2	7.9	8.3	8.9	9.0
10.1	9.9	8.8	7.8	5.2
7.2				

Developing				
6.3	5.2	8.3	5.9	5.5
7.1	8.1	7.9	6.3	6.9
9.1	8.1	7.0	4.9	5.3
6.3	7.1	6.3	6.1	5.8
5.7	6.8	8.3	7.7	

Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you accept this that μ_1 is higher than μ_2?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. Since σ_1, σ_2 are not known, we perform a two sample T-test.
Use TI-84, as in chapter 2, to summarize the raw data:

	Population I X	**Population II** Y
sample size	$m = 21$	$n = 24$
samplemean	$\overline{X} = 7.9905$	$\overline{Y} = 6.75$
Sample St. Dev.	$S_1 = 1.3758$	$S_2 = 1.1417$
Degrees of freedom $= m + n - 2 = 21 + 24 - 2 = 43$		

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} = \sqrt{\frac{(21-1)1.3758^2 + (24-1)1.1417^2}{21+24-2}} = 1.2560$$

1. The alternative hypothesis is that μ_1 is higher than μ_2: $H_A : \mu_1 > \mu_2$.
   The null and alternative hypotheses are:
   $$\begin{cases} H_0 : & \mu_1 - \mu_2 = 0 \\ H_A : & \mu_1 - \mu_2 > 0 \end{cases}$$

2. The test statistic
   $$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{7.9905 - 6.75}{1.2560\sqrt{\frac{1}{21} + \frac{1}{24}}} = 3.3053$$

3. This is a **right tail** test. So, by (9.75), the p-value
   $\mathbf{p} = P(T > t) \approx tcdf(t, 5, m+n-2)$
   $= tcdf(3.3053, 5, 43) = 9.6030 * 10^{-4}$

4. Three percent level of significance means $\alpha = .03$. Since,
   p-value $\mathbf{p} = 9.6030 * 10^{-4} < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
   That means we ACCEPT that the mean birth weight in Hospital-I is higher than that in Hospital-II.

5. Since p-value $\mathbf{p} = 9.6030 * 10^{-4}$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; .1 percent ($\alpha = .001$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.6.5. Elephants in different parts of the world are different in height, weight, and length of ear and tusk. It is speculated that mean height μ_1 of elephants in a region is higher than the mean height μ_2 in another region. It would be reasonable to assume that

the height distributions X and Y of elephants in these two regions are normally distributed and they have equal standard deviations σ. The following data were collected on the height of the elephants from these two regions:

Region X				
10.9	11.7	9.3	9.9	11.5
8.8	12.9	11.7	9.1	11.1
9.1	8.7	10.5	11.3	12.3
13.1	12.9	9.5	10.7	11.3

Region Y				
8.1	9.3	9.2	10.1	10.3
10.3	10.7	9.9	9.8	10.1
8.9	10.9	10.2	9.8	9.1
9.7	9.8	10.3	11.1	10.9
10.9				

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance would you accept that μ_1 is higher than μ_2?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. Since σ_1, σ_2 are not known, we perform a two sample T-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
sample size	$m = 20$	$n = 21$
samplemean	$\overline{X} = 10.815$	$\overline{Y} = 10.0190$
Sample St. Dev.	$S_1 = 1.4162$	$S_2 = .8072$
Degrees of freedom $= m + n - 2 = 20 + 21 - 2 = 39$		

The pooled estimate of σ is given by

$$S_p = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}} = \sqrt{\frac{(20-1).8072^2 + (21-1)2.73^2}{20+21-2}} = 1.1451$$

1. The alternative hypothesis is that μ_1 is higher than μ_2: $H_A : \mu_1 > \mu_2$.
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & \mu_1 - \mu_2 = 0 \\ H_A : & \mu_1 - \mu_2 > 0 \end{cases}$$

2. The test statistic
$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{0.815 - 10.0190}{1.1451\sqrt{\frac{1}{20} + \frac{1}{21}}} = 2.2249$$

3. This is a **right tail** test. So, by (9.75), the p-value
$\mathbf{p} = P(T > t) \approx tcdf(t, 5, m + n - 2)$
$= tcdf(2.2249, 5, 39) = .0160$

4. Three percent level of significance means $\alpha = .03$. Since,
p-value $\mathbf{p} = .0160 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
That means we ACCEPT that the mean mean height of elephants in Region-I is higher than that in Region-II.

5. Since p-value $\mathbf{p} = .0160$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 2 percent ($\alpha = .02$) would be the lowest level at which we would reject the null hypothesis.

9.7 Two Populations: Two Proportions p_1, p_2

This section is analogous to section 8.3, on confidence intervals for the difference $p_1 - p_2$ of two population proportions. The test will be called the Two proportion Z-interval. Let p_1 and p_2 represent the proportions of the attribute A, in Population I and Population II, respectively. As in section 8.3, we draw two independent samples, and the following date is obtained:

	Population I X	**Population II** Y
sample size	m	n
No. of success	X	Y
Sample proportion of success	$\overline{X} = \frac{X}{m}$	$\overline{Y} = \frac{Y}{n}$

As shown in section 8.3, approximately,

$$Z = \frac{(\overline{X} - \overline{Y}) - (p_1 - p_2)}{\sigma} \sim N(0, 1) \quad \text{has a st. normal distribution, where} \tag{9.78}$$

$$st.\ dev = \sigma = \sqrt{\frac{\mathbf{p_1(1 - p_1)}}{\mathbf{m}} + \frac{\mathbf{p_2(1 - p_2)}}{\mathbf{n}}}$$

We would do three tests:

Two Tail Test	**Left Tail Test**	**Right Tail Test**
$\begin{cases} H_0: & p_1 - p_2 = 0 \\ H_A: & p_1 - p_2 \neq 0 \end{cases}$	$\begin{cases} H_0: & p_1 - p_2 = 0 \\ H_A: & p_1 - p_2 < 0 \end{cases}$	$\begin{cases} H_0: & p_1 - p_2 = 0 \\ H_A: & p_1 - p_2 > 0 \end{cases}$

(9.79)

Assume $H_0 : p_1 - p_2 = 0$ is true. Then,

1. We can combine the estimate $\overline{X} = \frac{X}{m}$ for p_1, and the estimate $\overline{Y} = \frac{Y}{n}$ for p_2, to obtain

$$\overline{p} = \frac{X+Y}{m+n} \qquad \text{as a pooled estimate for} \quad p_1 = p_2$$

We also say that $\overline{p}$ is the grand sample proportion of success. So, σ can be estimated by

$$\sigma = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}} \approx \sqrt{\frac{\overline{p}(1-\overline{p})}{m} + \frac{\overline{p}(1-\overline{p})}{n}} = \sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{m}+\frac{1}{n}\right)} = \mathfrak{s} \text{ (say)}$$

2. By (9.78),

$$Z = \frac{\overline{X}-\overline{Y}}{\sigma} \approx \frac{\overline{X}-\overline{Y}}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim N(0,1) \quad \text{has a st. normal distribution.} \tag{9.80}$$

Expression Z above would be called a **test statistic**. We use lower case, to denote

$$z = \frac{\overline{x}-\overline{y}}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} \qquad \text{the observed value of} \quad Z. \tag{9.81}$$

For three cases of alternate hypotheses, as in (9.79), we formulate the decision rules:

1. **Two Tail Test:** When H_0 is true, by (9.80),

$$\begin{cases} P\left(|Z| \le z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha, & OR \\ P\left(|Z| > z_{\alpha/2}\right) = \alpha \end{cases} \tag{9.82}$$

Define p-value, for this test, as

$$\mathbf{p} = P\left(|Z| > |z|\right) = 1 - normalcdf\left(-|z|, |z|\right) \tag{9.83}$$

Because of (9.82), **at level of significance** α, set the decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \ |z| > z_{\alpha/2} \\ \textbf{Accept } H_0 & if \ \ |z| \le z_{\alpha/2} \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if \ |z| > z_{\alpha/2} \\ \textbf{Reject } H_A & if \ \ |z| \le z_{\alpha/2} \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if \ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if \ \mathbf{p} \ge \alpha \end{cases} \tag{9.84}$$

2. **Left Tail Test:** When H_0 is true, by (9.80),

$$P\left(Z < -z_\alpha\right) = \alpha \tag{9.85}$$

Define p-value, for this test, as

$$\mathbf{p} = P(Z < z) = normalcdf(-5, z). \tag{9.86}$$

Because of (9.85), **at level of significance** α, set the decision rules:

$$\begin{cases} \textbf{Reject } H_0 & if\ z < -z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if\ z < -z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.87}$$

3. **Right Tail Test:** When H_0 is true, by (9.80),

$$P\left(z_\alpha < Z\right) = \alpha \tag{9.88}$$

Define p-value, for this test, as

$$\mathbf{p} = P(Z > z) = normalcdf(z, 5). \tag{9.89}$$

Because of (9.88), **at level of significance** α, set the decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ z > z_\alpha \\ \textbf{Accept } H_0 & Otherwise \end{cases} \text{ Equivalently, } \begin{cases} \textbf{Accept } H_A & if z > z_\alpha \\ \textbf{Reject } H_A & Otherwise \end{cases}$$

This translates to the following p-value based decision rule:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.90}$$

These are called Two proportion Z-Test.

Universal Decision rule. As before (9.13, 9.27, 9.40, 9.51, 9.64, 9.77), the **p-value based decision rules** (9.84,9.87, 9.90) looks the same:

$$\begin{cases} \textbf{Reject } H_0 & if\ \mathbf{p} < \alpha \\ \textbf{Accept } H_0 & if\ \mathbf{p} \geq \alpha \end{cases} \tag{9.91}$$

However, p-values $\mathbf{p}$ are defined differently (9.83, 9.86, 9.89) are specific to the respective tests.

9.7.1 Problems: Two proportion Z-test

Exercise 9.7.1. The proportions p_1, p_2 , respectively, of an attribute A present in two populations would have to be compared. It is believed that p_1 is higher than p_2. A sample of size $m = 117$ was drawn from the first population and $X = 70$ had the attribute A. Similarly, a sample of size $n = 79$ was drawn from the second second population and $Y = 37$ had the attribute A.

Perform a significance test as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance would you accept this speculation that p_1 is higher than p_2?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. We perform a two proportion Z-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
No. of success	$X = 70$	$Y = 37$
sample size	$m = 117$	$n = 79$
sample proportion	$\overline{X} = \frac{70}{117} = .5983$	$\overline{Y} = \frac{37}{79} = .4684$
Grand sample proportion $= \frac{X+Y}{m+n} = \frac{70+37}{117+79} = .5459$		

1. The alternative hypothesis is that the p_1 is higher then p_2: $H_A : p_1 > p_2$.
   The null and alternative hypotheses are:
$$\begin{cases} H_0 : & p_1 - p_2 = 0 \\ H_A : & p_1 - p_2 > 0 \end{cases}$$
2. The test statistic
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{.5983 - .4684}{\sqrt{.5459(1-.5459)\left(\frac{1}{117} + \frac{1}{79}\right)}} = 1.7917$$

3. This is a **right tail** test. So, by (9.89), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5)$
   $= normalcdf(z, 5) = normalcdf(1.7917, 5) = .0366$

4. Three percent level of significance means $\alpha = .03$. Since, p-value
   $\mathbf{p} = .0366 \not< \alpha = .03$, we ACCEPT the null hypothesis at 3 percent level of significance.
   That means, at three percent level of significance, we DO NOT accept that p_1 is higher.

5. Since p-value $\mathbf{p} = .0366$, from this possibilities of .1, .5, 1, 2, 3, 4, 5,4, 7,8, 9,10 percent; 4 percent ($\alpha = .04$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.7.2. To compare the proportions p_1, p_2, of defective lamps produced by new production center and old the production center, respectively, samples were collected. In a sample of 157 lamps from the new center, 26 were found to be defective; and in a sample of 141 lamps from the old center, 32 were defective. Perform a significance test that the new center is performing better, based on proportion of defective lamps, as follows.

Solution. We perform a two proportion Z-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
No. of success	$X = 26$	$Y = 32$
sample size	$m = 157$	$n = 141$
sample proportion	$\overline{X} = \frac{26}{157} = .1656$	$\overline{Y} = \frac{32}{141} = .2270$
Grand sample proportion $= \frac{X+Y}{m+n} = \frac{26+32}{157+141} = .1946$		

1. The alternative hypothesis the new center is performing better: $H_A : p_1 < p_2$.
   The null and alternative hypotheses are:

$$\begin{cases} H_0 : & p_1 - p_2 = 0 \\ H_A : & p_1 - p_2 < 0 \end{cases}$$

2. The test statistic

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{.1656 - .2270}{\sqrt{.1946(1-.1946)\left(\frac{1}{157} + \frac{1}{141}\right)}} = -1.3367$$

3. This is a **left tail** test. So, by (9.86), the p-value
   $\mathbf{p} = P(Z < z) = normalcdf(-5, z)$
   $= normalcdf(z, 5) = normalcdf(-5, -1.3367) = .0907$

4. Three percent level of significance means $\alpha = .03$. Since, p-value
   $\mathbf{p} = .0907 \not< \alpha = .03$, we ACCEPT the null hypothesis at 3 percent level of significance.
   That means, at three percent level of significance, we DO NOT accept that the new center is performing better.

5. Since p-value $\mathbf{p} = .0907$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 10 percent ($\alpha = .10$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.7.3. Data was collected to compare the proportions p_1, p_2 of men and women, respectively, who watch football. In a sample of 199 men, 83 said that they watch football; and in a sample of 161 women, 51 said they watch football. (These are not real data).

Perform a significance test that the proportion of men who watch football is higher, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you accept this speculation that p_1 is higher than p_2?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Solution. We perform a two proportion Z-test.
The given data is summarized as follows:

	Population I X	**Population II** Y
No. of success	$X = 83$	$Y = 51$
sample size	$m = 199$	$n = 161$
sample proportion	$\overline{X} = \frac{83}{199} = .4171$	$\overline{Y} = \frac{51}{161} = .3168$
Grand sample proportion $= \frac{X+Y}{m+n} = \frac{83+51}{199+161} = .3722$		

1. The alternative hypothesis is that the p_1 is higher then p_2: $H_A : p_1 > p_2$.
   The null and alternative hypotheses are:
$$\begin{cases} H_0 : & p_1 - p_2 = 0 \\ H_A : & p_1 - p_2 > 0 \end{cases}$$

2. The test statistic
$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{.4171 - .3168}{\sqrt{.3722(1-.3722)\left(\frac{1}{199} + \frac{1}{161}\right)}} = 1.9574$$

3. This is a **right tail** test. So, by (9.89), the p-value
   $\mathbf{p} = P(Z > z) = normalcdf(z, 5)$
   $= normalcdf(z, 5) = normalcdf(1.9574, 5) = .0251$

4. Three percent level of significance means $\alpha = .03$. Since, p-value
   $\mathbf{p} = .0251 < \alpha = .03$, we REJECT the null hypothesis at 3 percent level of significance.
   That means, at three percent level of significance, we ACCEPT that the proportion of men who watch football is higher than that of women.

5. Since p-value $\mathbf{p} = .0251$, from this possibilities of .1, .5, 1, 2, 3, 4, 5, 6, 7,8, 9,10 percent; 3 percent ($\alpha = .03$) would be the lowest level at which we would reject the null hypothesis.

Exercise 9.7.4. Two varieties of grapes are compared. To compare the proportions p_1, p_2 of acceptable grapes of these two varieties, respectively, samples were drawn. In a sample of 131 grapes from the variety I, 112 were acceptable. In a sample of 143 grapes from the variety II, 113 were acceptable. Perform a significance test that the proportion p1 acceptable grapes of variety-I is higher than that of the variety-II, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you REJECT the null hypothesis (that means *the proportion p_1 of acceptable grapes of variety-I is higher*)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.7.5. To compare the proportions p_1, p_2 of students, respectively, in two state universities who pay more than \$15 K tuition per year, samples were collected. In a sample of 217 students in the university I, 129 paid more than \$15 K. In a sample of 313 students in the university II, 158 paid more than \$15 K. It is speculated that the university I is more expensive than university-II and p_1 is higher than p_2. Perform a significance test that the proportion p_1 of those in university-I that pay more than \$15 K in tuition is higher that the proportion p_2 of those in university-II, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you REJECT the null hypothesis (that means *that the proportion p_1 of acceptable grapes of variety-I is higher*)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.7.6. To compare the proportions p_1, p_2 of college graduates who earn more than \$50 K, in two states, data was collected. In a sample of 444 college graduates in the state I, 354 earn more than \$50 K. In a sample of 546 college graduates in the state II, 414 earn more than \$50 K. It is speculated that higher proportion of graduates in state-I earn more than \$50 K. Perform a significance test that the proportion p_1 higher than p_2, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.

2. Compute the value of the test statistic.

3. Compute the p-value.

4. At the 3 percent level of significance would you REJECT the null hypothesis (that means *that the proportion p_1 is higher*)?

5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

Exercise 9.7.7. It is believed that women are safer drivers than men. Let p_1, p_2 denote the proportions of women and men drivers, respectively, who were involved in an auto accident in a year period. In a sample of a size 770 women drivers 39 were involved in auto accident during this period. During the same period, in a sample of size 1215 men 79 were involved in auto accident in a year. It is speculated that proportion p1 of women drivers who were involved in auto accidents last year is lower than that p2 of men. Perform a significance test that the proportion p_1 lower than p_2, as follows.

1. Formulate the null and alternative hypotheses to perform a significance test.
2. Compute the value of the test statistic.
3. Compute the p-value.
4. At the 3 percent level of significance would you REJECT the null hypothesis (that means *that the proportion p_1 is lower*)?
5. What would be the lowest level of significance, among .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 percent, at which you would REJECT the null hypothesis?

9.8 Paired T-test (read only)

In this section, we deal with a situation, appears similar to Two Sample T-Test (section 9.6), while not the same. However, the problem reduces to a T-Test (section 9.2).

We would have two populations, which we continue to denote the first population random variable by X and the second population random variable by Y.

1. It is also assumed that $X \sim N(\mu_1, \sigma_1)$ and $Y \sim N(\mu_2, \sigma_2)$ have normal distributions.
2. We also assume that X and Y independent.
3. To compare μ_1, and μ_2, this would be situations, when it is natural to collect samples in "pairs" (X, Y) from the two populations and consider the difference $D = X - Y$. So, mean and standard deviation of D is given by
$$\mu_D = E(D) = E(X) - E(Y) = \mu_1 - \mu_2, \qquad \sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2}$$
4. Two compare the X and Y population, we test
$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{Equivalently,} \qquad H_0 : \mu_D = 0 \tag{9.92}$$

5. Samples

   $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ are collected, in pairs, from $(X, Y)-$population.

6. Write
$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \cdots, D_n = X_n - Y_n$$
   which we call the D-sample.

7. Then the mean of the D-sample

   $\overline{D} = \dfrac{D_1 + D_2 + \cdots + D_n}{n} = \overline{X} - \overline{Y}$ is the difference of two sample means.

   Similar to T-test (section 9.2), the test statistic would be

   $T = \dfrac{\overline{D}\sqrt{n}}{S_D}$ where $S_D = \sqrt{\dfrac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}$ is the sample st. deviation.

8. It follows, $T \sim t_{n-1}$ has a t-distribution, with degrees of freedom $n - 1$.

9. Now, decision rules are formulated, as in section 9.2, for three tests:

Two Tail Test	**Left Tail Test**	**Right Tail Test**
$\begin{cases} H_0: & \mu_D = 0 \\ H_A: & \mu_D \neq 0 \end{cases}$	$\begin{cases} H_0: & \mu_D = 0 \\ H_A: & \mu_D < 0 \end{cases}$	$\begin{cases} H_0: & \mu_D = 0 \\ H_A: & \mu_D > 0 \end{cases}$

(9.93)

This is known as Paired T-test.

Example. Suppose two models of cars would have to be comapared to see how fast they accelerate. In this case, to avoid any variation due to individual drivers, a sample of n drivers drive one of each model. So, (X_i, Y_i) are the accelerations of the first and second model driven by i^{th}-driver. Thus, there would be a sample n pairs of observations

$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, D_3 = X_3 - Y_3, \ldots, D_n = X_n - Y_n$$

Now, this can be treated as a T-Test.

Index